AMBIENT LLMS: DATA-EFFICIENT SUPERVISED FINE-TUNING VIA TEMPERATURE MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Collecting large-scale, diverse, and high-quality post-training datasets for autoregressive language models is challenging. Consequently, Supervised Fine-Tuning (SFT) is typically applied to datasets significantly smaller than those used for pre-training, which leads to overfitting and the catastrophic forgetting of the tails of the data distribution. To prevent this collapse, we introduce Ambient SFT, a principled framework for leveraging imperfect predictions to augment the fine-tuning set. In our method, a "frozen" version of the pre-trained model serves as a noisy teacher that the fine-tuned model must match at a coarse level. In particular, we require that the high-temperature versions of the conditional distributions produced by the post-training model should match the high-temperature versions of the pre-training ones. The temperature hyperparameter controls the level of trust in the pre-trained model, with the limit $T \to \infty$ reverting back to regular SFT and $T = 1$ to knowledge distillation. We demonstrate that our approach outperforms standard SFT across domain adaptation, instruction-tuning, and reasoning.

1 INTRODUCTION

The development of state-of-the-art Large Language Models (LLMs) has evolved into a complex, multi-stage pipeline (Liu et al., 2024; Bai et al., 2023; Abdin et al., 2024; Achiam et al., 2023; Olmo et al., 2025; Grattafiori et al., 2024). This process typically progresses from massive-scale pre-training to more specialized post-training phases, including instruction-tuning (Ouyang et al., 2022), alignment (Ziegler et al., 2019; Ouyang et al., 2022), and reasoning (Guo et al., 2025; Lightman et al., 2023). As models advance through this pipeline, data availability decreases by orders of magnitude. While pre-training corpora often encompass trillions of tokens, later stages frequently rely on datasets comprising merely of a few thousand highly curated examples (Zhou et al., 2023a; Guha et al., 2025). This severe bottleneck arises because post-training requires extraordinarily high-quality data, making large-scale collection prohibitively challenging (Zhou et al., 2023a).

The stark contrast in data scale introduces undesirable artifacts. Most notably, post-trained LLMs tend to suffer from catastrophic forgetting at the tails of the pre-training data distribution, overfitting instead to the specific stylistic patterns or narrow viewpoints overrepresented in the limited post-training data (Ouyang et al., 2022; Lin et al., 2024). This phenomenon is known as the *alignment tax* (Ouyang et al., 2022). The prevailing solution to this bottleneck is brute force: allocating vast financial and computational resources to scale up the collection of post-training data. In this work, we ask a different question: are there *algorithmic solutions* that can increase the data efficiency of post-training without inducing severe model collapse or overfitting?

We start our approach to this problem by performing a thought experiment. Imagine having access to an oracle LLM that can generate the high-quality post-training data distribution, but you can only observe a high-temperature, noisy version of its predictions. We show that learning from high-temperature data can match the performance of training on the true, standard-temperature distribution, provided the sample size is large enough. For a practical algorithm, we do not have a perfect high-temperature oracle, but we can effectively approximate it using a high-temperature *weak* model, which is the model prior to post-training. Distributionally, any two models converge as the sampling temperature increases; thus, at a sufficiently high temperature, a readily available weak model serves

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.

as a strong proxy for the unavailable oracle. This motivates our proposed approach: during post-training, we augment our dataset using a weak teacher operating at a high temperature. We enforce that the post-trained model matches the weak teacher's distribution at this high temperature, while remaining unconstrained to refine its conditional distributions at the standard temperature ($T = 1$). The temperature hyperparameter explicitly controls the level of trust placed in the pre-trained model, with the $T \to \infty$ limit gracefully reverting to standard SFT and with $T = 1$ to knowledge distillation.

Our algorithm, termed Ambient SFT, mitigates the overfitting typically observed when applying Supervised Fine-Tuning (SFT) to small datasets. It achieves this by adding to the standard SFT loss a second term, which we call Temperature Matching, that regularizes the post-trained network to match the high-temperature conditional distributions of its pre-trained predecessor. We evaluate Ambient SFT across three standard post-training paradigms: *(i)* domain adaptation, *(ii)* instruction-tuning, and *(iii)* the development of reasoning capabilities. Across all settings, our approach consistently outperforms standard SFT. In summary, our contributions are as follows:

- We introduce *Ambient SFT*, a drop-in replacement for standard SFT that regularizes fine-tuning by matching a high-temperature version of the model's own pre-trained distribution. A single temperature knob interpolates between SFT and knowledge distillation.
- We prove that a target distribution can be recovered from high-temperature conditionals alone, that any two models' conditionals contract as the temperature grows, and that under a log-linear model Ambient SFT strictly beats vanilla SFT and knowledge distillation.
- We show gains in all three post-training regimes: up to 26.7% relative improvement over SFT across five domain-adaptation targets; average relative gains of 8.0% to 20.3% on instruction tuning and wins on $5/6$ reasoning benchmarks.

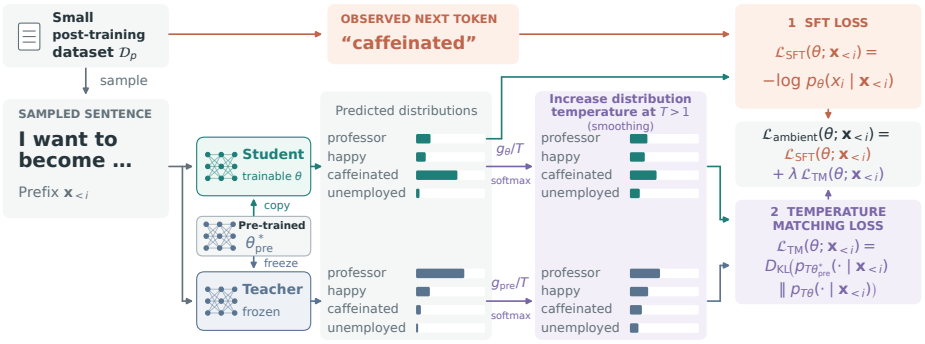


Figure 1: Overview of Ambient SFT for data-efficient post-training. A pre-trained model initializes a trainable student and a frozen teacher. On the small dataset $\mathcal{D}_p$, SFT fits the observed next token, while Temperature Matching aligns the two models' distributions after smoothing. Their weighted sum updates only the student. Temperature controls trust in the teacher: $T \to \infty$ recovers SFT and $T = 1$ knowledge distillation.

2 Background

Notation. Let V be a finite alphabet of tokens, and let $L \subseteq V^l$ be a language associated with a probability measure p. A sample from that language is denoted with $\mathbf{x} = (x_1, \ldots, x_l)$, where $x_i \in V$. We use $\mathbf{x}_{<i} = (x_1, ..., x_{i-1})$ to denote the vector with the first $(i-1)$ co-ordinates.

Training of Auto-Regressive (AR) Language Models. Any distribution, can be factorized as follows: $p(\mathbf{x}) = p(x_1)p(x_2|x_1)\cdot ... \cdot p(x_l|\mathbf{x}_{<l})$. Autoregressive Language Models (LMs) are artificial networks trained to capture these conditional distributions. Specifically, a model $g_\theta(\cdot|\mathbf{x}_{<i})$ outputs logits for each potential token in the vocabulary given a context $\mathbf{x}_{<i}$, inducing conditional distributions:

$$p_\theta(x_i \mid \mathbf{x}_{<i}) = \frac{e^{g_\theta(x_i|\mathbf{x}_{<i})}}{\sum_{y_i \in V} e^{g_\theta(y_i|\mathbf{x}_{<i})}}. \tag{1}$$

The standard training objective targets the minimization of the population negative log-likelihood

$$J(\theta) = \sum_{i=1}^{l} \mathbb{E}_{\mathbf{x}_{\leq i} \sim p_{\leq i}} \left[-\log \left(\frac{e^{g_\theta(x_i|\mathbf{x}_{<i})}}{\sum_{y_i \in V} e^{g_\theta(y_i|\mathbf{x}_{<i})}} \right) \right], \tag{2}$$

where $p_{\leq i}$ denotes the distribution of the first i tokens. Eq. (2) computes the cross-entropy between the model's predicted distribution and the target distributions of each token x_i observed after its prefix $\mathbf{x}_{<i}$. By minimizing this loss, the model approximates the true data conditionals $p(x_i \mid \mathbf{x}_{<i})$.

Adjusting the temperature. At inference time, the sharpness of the learned conditional distributions can be modulated by dividing the network logits by a scalar $T > 0$, known as the *sampling temperature*. We use $p_{T\theta}(x_i|\mathbf{x}_{<i}) \propto p_\theta(x_i|\mathbf{x}_{<i})^{1/T}$ to refer to the temperature-scaled distributions:

$$p_{T\theta}(x_i|\mathbf{x}_{<i}) = \frac{e^{g_\theta(x_i|\mathbf{x}_{<i})/T}}{\sum_{y_i \in V} e^{g_\theta(y_i|\mathbf{x}_{<i})/T}}. \tag{3}$$

The need for post-training. In practice, we do not have oracle access to the true distribution p. Instead, we optimize the objective of Eq. (2) using a finite empirical dataset $\mathcal{D}_p = \{\mathbf{x}^{(k)}\}_{k=1}^{n_1}$ sampled from p. We denote this empirical distribution with $\hat{p}_{n_1}$. Typically, the success of the generative training depends on the size of n_1. However, collecting a sufficiently large dataset directly from p is often prohibitively expensive. For example, p might represent the distribution of expert-level textbooks or highly specialized instruction-following dialogues. While such in-distribution data is scarce, there exists an abundance of generic text on the web, drawn from a broader, different distribution q. This necessitates the use of two distinct datasets: a massive dataset $\mathcal{D}_q = \{\mathbf{y}^{(k)}\}_{k=1}^{n_2}$ sampled from q, and our limited target dataset $\mathcal{D}_p$, where $n_1 \ll n_2$. The prevailing learning paradigm leverages both datasets by first pre-training the model on $\mathcal{D}_q$ and subsequently fine-tuning it on $\mathcal{D}_p$. Hereafter, we use θ^*_{pre} and θ^*_{post} to denote the parameters of the pre-trained and fine-tuned models.

3 Method and Theoretical Results

Problem setting. In our setting, we assume access to a pre-trained model θ^*_{pre} but not necessarily to the pre-training dataset $\mathcal{D}_q$. Our goal is to adapt this model to capture as best as possible the distribution p, from which we only have access to a small dataset $\mathcal{D}_p = \{\mathbf{x}^{(k)}\}_{k=1}^{n_1}$.

3.1 Ambient SFT algorithm

We start by presenting our algorithm rightaway. Ambient SFT fine-tunes a pre-trained model with parameters θ^*_{pre} on the new dataset $\mathcal{D}_p$ with two losses on the same tokens: the usual cross-entropy on the observed next token, and a *temperature matching* loss that pulls the student's conditional distribution at temperature T toward the pre-trained model's conditional distribution at temperature T. With $\mathcal{D}_p = \{\mathbf{x}^{(k)}\}_{k=1}^{n_1}$ the post-training documents, each of length l, and $p_{T\theta}$ the temperature-scaled conditionals of Eq. (3), the objective is

$$\widehat{J}_{\lambda,T}(\theta) = \frac{1}{n_1 l} \sum_{k=1}^{n_1} \sum_{i=1}^{l} \Big[\underbrace{-\log p_\theta(x_i^{(k)} \mid \mathbf{x}_{<i}^{(k)})}_{\text{Clean Cross-entropy}} + \lambda \underbrace{D_{\text{KL}}\big(p_{T\theta^*_{\text{pre}}}(\cdot \mid \mathbf{x}_{<i}^{(k)}) \,\|\, p_{T\theta}(\cdot \mid \mathbf{x}_{<i}^{(k)})\big)}_{\text{Temperature Matching}} \Big] \tag{4}$$

where the student θ is initialized at θ^*_{pre}, the teacher θ^*_{pre} is frozen, and $\lambda > 0$ sets the relative weight of the two terms. The KL term equals the cross-entropy $-\sum_{y \in V} p_{T\theta^*_{\text{pre}}}(y \mid \mathbf{x}_{<i}^{(k)}) \log p_{T\theta}(y \mid \mathbf{x}_{<i}^{(k)})$ up to a constant that does not depend on θ, which is what we optimize.

We refer to the full objective of Eq. (4) as Ambient SFT and to its second term as Temperature Matching. The name connects to Ambient Diffusion (Daras et al., 2023; 2024; 2025; Rodríguez-Muñoz et al., 2026), a line of work that trains diffusion models from corrupted data. Here, we view as corruption the imperfect predictions of the pre-trained teacher, which we only trust at a high T.

Extreme case 1: $T \to \infty$. For any two distributions p, q that have full support, the D_{KL} distance goes to 0 as $T \to \infty$. Hence, the second term disappears, and our method reverts to regular SFT.

Extreme case 2: $T = 1$. On the other hand, setting $T = 1$ gives rise to the classic Knowledge Distillation (KD) loss. Typically, in KD, the student is a weak model, and the teacher is a strong

model. In our case, we have the opposite: the teacher is the pre-trained model, and the student is the model being post-trained that is expected to eventually outperform the teacher.

Clarification and Related Work. Sometimes in the literature the term KD is used to refer to matching any teacher distribution, independent of the smoothing temperature T. In fact, Hinton et al. (2015) experiment with matching at different temperatures. The main difference is that we use the pre-trained model to *supervise itself* during post-training. For more related work, see Appendix E.

Practical Implementation. Pseudocode is given in Algorithm 1, Appendix B. Our implementation optimizes Eq. (4) stochastically. At every step, with probability p_c we use the regular SFT loss (clean cross-entropy only). With probability $1 - pc$, we do a forward pass through the frozen teacher and we minimize the Temperature Matching loss between the tempered teacher and tempered student conditionals at every supervised position. In expectation, this implements (4) with $\lambda = (1 - p_c)/p_c$.

Computational overhead. Since Ambient SFT requires an additional forward pass through the frozen teacher, it is more expensive than the regular SFT. Using the standard approximation that a backward pass costs roughly twice a forward pass, the expected per-step compute relative to plain fine-tuning is $1 + \frac{1-p_c}{3} \cdot \frac{C_{\text{teacher}}}{C_{\text{student}}}$, where $C_{\text{teacher}}/C_{\text{student}}$ is the ratio of teacher to student forward-pass cost. For our Experiments, we typically set $p_c = 0.75$ and we use the same model for the student and the teacher. This makes training $\approx 1.1\times$ slower for the same amount of steps. Hence, Ambient SFT is best suited to settings that are *data-limited* rather than *compute-limited*.

3.2 Theoretical Analysis of the Algorithm

We now analyze Ambient SFT theoretically: we show that pure temperature matching is equivalent to learning from high-temperature samples, characterize how fast temperature contracts distributional differences, and establish the statistical benefits of Ambient SFT under a simple log-linear model. Readers interested mainly in the empirical results can skip to Section 4.

3.2.1 Equivalence to learning from high-temperature data

We start by showing that training with the temperature matching loss of Eq. (4) alone is equivalent to learning from high-temperature data alone. The result holds for matching toward *any* autoregressive model, so we write θ^* for the teacher; in the algorithm $\theta^* = \theta^*_{\text{pre}}$. The proof is in Appendix A.1.

Theorem 3.1 (Learning from noisy data)**.** *Let p_{θ^*} be any AR model over V^l with logits g_{θ^*} as in Eq.* (1)*, let $p_{T\theta^*}(\cdot \mid \mathbf{x}_{<i})$ be its temperature-scaled conditionals of Eq.* (3)*, and let $p_{T\theta^*}(\mathbf{x}) = \prod_{i=1}^{l} p_{T\theta^*}(x_i \mid \mathbf{x}_{<i})$ be the distribution of text sampled autoregressively at temperature T.*

(a) ***Temperature Matching is equivalent to learning from high-temperature samples.*** *Let $R_T(\theta) = \mathbb{E}_{\mathbf{x}\sim p_{T\theta^*}}\big[\sum_{i=1}^{l} D_{\mathrm{KL}}\big(p_{T\theta^*}(\cdot \mid \mathbf{x}_{<i}) \,\|\, p_{T\theta}(\cdot \mid \mathbf{x}_{<i})\big)\big]$ be the (population-level) temperature-matching loss of Eq.* (4) *with teacher θ^*, and let*

$$J_T(\theta) = \mathbb{E}_{\mathbf{x}\sim p_{T\theta^*}}\left[\sum_{i=1}^{l} -\log p_{T\theta}(x_i \mid \mathbf{x}_{<i})\right] \tag{5}$$

be the tempered cross-entropy on text sampled at temperature T. Then J_T, R_T have the same minimizers. Unlike R_T, J_T needs only samples from the tempered model, not its logits.

(b) ***Exact recovery at any finite*** *T.* *If θ minimizes J_T at any $T > 0$, then $p_{T'\theta}(\cdot \mid \mathbf{x}_{<i}) = p_{T'\theta^*}(\cdot \mid \mathbf{x}_{<i})$ for every context $\mathbf{x}_{<i}$ and every $T' > 0$, i.e., matching the teacher at any temperature implies matching the teacher at all temperatures. Moreover, this implies that $g_\theta(\cdot \mid \mathbf{x}_{<i}) = g_{\theta^*}(\cdot \mid \mathbf{x}_{<i}) + c(\mathbf{x}_{<i})$ for a context-dependent constant.*

The theorem shows that the distribution is identifiable from high-temperature data. This is perhaps counterintuitive: at high T, samples look gibberish. Yet they still carry a learning signal, and a model trained with Eq. (5) recovers the distribution, producing reasonable samples as shown in Figure 2a.

3.2.2 Temperature scaling as a contraction mechanism

If we had lots of data from $p_{\theta^*_{\text{post}}}$, we could run temperature matching towards it for any T of our choice. Unfortunately, in our setting we assume we do not have many data from it. What happens

if we run temperature matching to the weak teacher $p_{\theta^*_{\text{pre}}}$ instead? The following result establishes an upper bound on the KL divergence between the two distributions, for any fixed context as the temperature T increases. We show that, for every context $\mathbf{x}_{<i}$ and every $T > 0$:

$$D_{\text{KL}}\Big(p_{T\theta^*_{\text{post}}}(\cdot \mid \mathbf{x}_{<i}) \,\Big\|\, p_{T\theta^*_{\text{pre}}}(\cdot \mid \mathbf{x}_{<i})\Big) \leq \frac{\Delta(\mathbf{x}_{<i})^2}{8T^2}, \tag{6}$$

where $\Delta(\mathbf{x}_{<i})$ is the range over $y \in V$ of the logit gap $g_{\theta^*_{\text{post}}}(y \mid \mathbf{x}_{<i}) - g_{\theta^*_{\text{pre}}}(y \mid \mathbf{x}_{<i})$. Our bound is stated in more detail in Appendix A.2. The bound gives a rate of $O(T^{-2})$ for every context.

3.2.3 Ambient SFT analysis under a log-linear model

We are now ready to put all the pieces together. Theorem 3.1 establishes the equivalence of Temperature Matching to learning from corrupted data from the teacher and shows that for any finite T, the teacher distribution is the minimizer of this term alone. Training with Temperature Matching alone would be problematic since we would not be able to move from the pre-training distribution. Our algorithm trains with both regular SFT and Temperature Matching, with a choice of temperature T controlling the level at which the gradients from the weak teacher can be reliable, as quantified by Eq. (6). The questions left to answer are: *(a)* what is the right choice of temperature in the temperature matching term, and *(b)* for the optimal amount of temperature, do we get any statistical advantage compared to learning from regular SFT alone? In what follows, we answer those questions. To avoid arguing about how LLMs generalize to new contexts not seen in training, we adopt a simple log-linear next-token model with frozen features, and analyze (4) on it exactly.

A log-linear model. Let $\phi : V \to \mathbb{R}^d$ be a fixed token embedding and $\psi : V^* \to \mathbb{R}^m$ a fixed context representation (e.g., the pre-trained network's last hidden state). Assume both functions have bounded ℓ_2 norms. The learnable parameter θ is a single matrix, which we will also denote $A_\theta \in \mathbb{R}^{d\times m}$, and

$$p_\theta(y \mid \mathbf{x}_{<i}) = \frac{\exp\big(\phi(y)^\top A_\theta \psi(\mathbf{x}_{<i})\big)}{\sum_{z\in V}\exp\big(\phi(z)^\top A_\theta \psi(\mathbf{x}_{<i})\big)} \tag{7}$$

We can think of such a model as an architectural choice for g_θ, i.e., we write $g_\theta(y \mid \mathbf{x}_{<i}) = \phi(y)^\top A_\theta \psi(\mathbf{x}_{<i})$. Write $\Lambda(\eta) = \log\sum_{z\in V}\exp(\phi(z)^\top \eta)$ for the log-normalizer of (7), i.e. the logarithm of its denominator, and $\mu(\eta) = \nabla\Lambda(\eta) = \mathbb{E}_{p_\eta}[\phi(y)]$, $\Sigma(\eta) = \nabla^2\Lambda(\eta) = \text{Cov}_{p_\eta}[\phi(y)]$ for the mean embedding and the covariance of the conditional with natural parameter $\eta = A_\theta\psi(\mathbf{x}_{<i})$. Flattening the sum of (4), the n_1 documents supply $n = n_1 l$ token observations $(\mathbf{x}_{<t}, x_t)$, $t = 1, \ldots, n$, and we write $\psi_t = \psi(\mathbf{x}_{<t})$. We will use θ and A_θ interchangebly.

The objective. Instantiating (4) on this model, gives, up to additive constants that do not depend on θ,

$$\widehat{J}_{\lambda,T}(\theta) = \underbrace{\frac{1}{n}\sum_{t=1}^{n}\Big[\Lambda(A_\theta\psi_t) - \phi(x_t)^\top A_\theta\psi_t\Big]}_{\widehat{L}(\theta)\text{ (clean cross-entropy)}} + \lambda\underbrace{\frac{1}{n}\sum_{t=1}^{n}\Big[\Lambda\big(\tfrac{A_\theta\psi_t}{T}\big) - \mu\big(\tfrac{A_{\theta^*_{\text{pre}}}\psi_t}{T}\big)^\top\tfrac{A_\theta\psi_t}{T}\Big]}_{\widehat{R}_T(\theta)\text{ (temperature matching)}}. \tag{8}$$

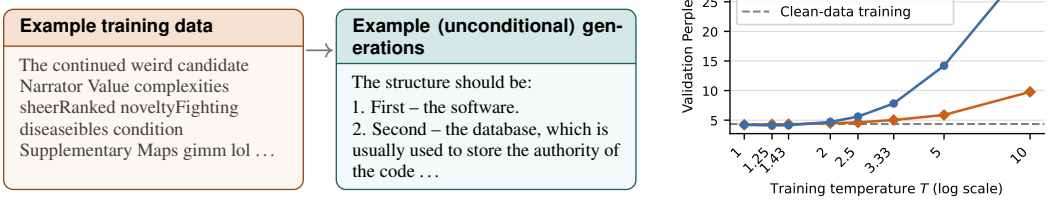
Write $\widehat{\theta}_\lambda$ for the minimizer of (8). Our goal in this section is to bound its distance from θ^*_{post} and study the benefit of including temperature matching term in the objective.

Remark: For simplicity, we assume below that, in the data $(\mathbf{x}_{<t}, x_t)$, $t = 1, \ldots, n$, that we have available for training, the prefixes $\mathbf{x}_{<t}$ are drawn i.i.d. from $p_{\theta^*_{\text{post}}}$ and the tokens x_t are also drawn independently from each other from $p_{\theta^*_{\text{post}}}(\cdot|\mathbf{x}_{<t})$. We denote with $\mathcal{D}_x$ the distribution of prefixes drawn from $p_{\theta^*_{\text{post}}}$. Appendix A.3 discusses how to deal with dependencies of tokens in documents.

Three quantities govern our result, where all expectations below are with respect to prefixes $\mathbf{x}_{<t} \sim \mathcal{D}_x$, $\psi = \psi(\mathbf{x}_{<t})$, and $\otimes$ is the Kronecker product:

$$F = \mathbb{E}\big[\Sigma(A_{\theta^*_{\text{post}}}\psi)\otimes\psi\psi^\top\big], \quad H_T = \frac{1}{T^2}\,\mathbb{E}\big[\Sigma(\tfrac{A_{\theta^*_{\text{post}}}\psi}{T})\otimes\psi\psi^\top\big], \quad B = A_{\theta^*_{\text{post}}} - A_{\theta^*_{\text{pre}}}.$$

We assume throughout that the frozen features are non-degenerate: $G = \mathbb{E}[\psi\psi^\top] \succ 0$, and the embedding differences $\phi(y) - \phi(z)$ span $\mathbb{R}^d$. The latter fails only if some $v \neq 0$ has $\phi(y)^\top v$ constant



(a) Training data (left) and generations (right) of a model that has never seen a proper English sentence.

(b) Clean-data perplexity versus training temperature.

Figure 2: Learning from high-temperature data.

in y, in which case A_θ would not be identifiable; under the two conditions F and H_T are positive definite, so the inverses below are literal and $\|A_{\widehat{\theta}} - A_{\theta^*_{\text{post}}}\|_F$ is a well-defined risk.

F is the Fisher information matrix controlling the curvature of the first term in the loss function. It measures how much the data from the target distribution tell us about each direction of $A_{\theta^*_{\text{post}}}$.

H_T is the curvature of the temperature matching penalty term. As T grows the tempered conditional approaches uniform, so $\Sigma(A_\theta\psi/T) \to \Sigma_0 := \text{Cov}_{\text{unif}(V)}[\phi(y)]$ for every θ and $H_T \approx \Sigma_0 \otimes G/T^2$, where $G = \mathbb{E}[\psi\psi^\top]$, which is the same context geometry as F, but with a token covariance that never degenerates. The teacher thus adds curvature precisely where the data supplies none.

The next theorem quantifies how these quantities affect the estimation error. The expansion concerns the regime in which the teacher's error and the statistical error of maximum likelihood are comparable, both of order $n^{-1/2}$, which is where the tradeoff between them is nontrivial. We therefore state the result under $\|B\| = O(n^{-1/2})$. The proof is in Appendix A.3.

Theorem 3.2 (Risk expansion and generic improvement over $T = 1$)**.** *Suppose* $\|B\| = O(n^{-1/2})$. *Then*

$$\mathbb{E}\big\|A_{\widehat{\theta}_{\lambda,T}} - A_{\theta^*_{\text{post}}}\big\|_F^2 = \rho(\lambda, T) + O(n^{-3/2}),$$
$$\rho(\lambda, T) = \underbrace{\lambda^2 \|(F + \lambda H_T)^{-1} H_T B\|_F^2}_{\textit{bias}^2} + \underbrace{\tfrac{1}{n}\,\text{tr}\big((F + \lambda H_T)^{-1} F (F + \lambda H_T)^{-1}\big)}_{\textit{variance}}, \quad (9)$$

uniformly for λ *and* T *in a fixed range bounded away from* 0 *and* ∞. *Moreover, interior choices of* (λ, T) *do at least as well as all three extremes:*

$$\inf_{\substack{0<\lambda<\infty \\ 0<T<\infty,\ T\neq 1}} \rho(\lambda, T) \;\le\; \min\Big\{ \underbrace{\text{tr}(F^{-1})/n}_{\textit{maximum likelihood}},\ \underbrace{\|B\|_F^2}_{\textit{frozen teacher}},\ \underbrace{\min_\lambda \rho(\lambda, 1)}_{\textit{knowledge distillation}} \Big\}, \quad (10)$$

the three terms inside the minimum being $\rho(0, T)$, $\lim_{\lambda\to\infty} \rho(\lambda, T)$ *and the best* $T = 1$ *run. Under a mild non-degeneracy condition, and for every* B *outside a closed Lebesgue-null cone (see Appendix A.3), the inequality is strict: some finite* (λ, T) *with* $T \neq 1$ *strictly beats all three.*

4 Experiments

4.1 Learning from high-temperature data

We first verify that training on high-temperature samples can still recover the original ($T = 1$) distribution. We train a model on a dataset $\mathcal{D}$, sample from it at temperature T to obtain $\mathcal{D}'_T$, train a new model on $\mathcal{D}'_T$ with Eq. (5), and measure its perplexity on a clean validation set.

For our first toy experiment, we use a GPT-2 architecture, and we train it on the tiny Shakespeare dataset (Karpathy, 2015). Figure 2b shows the validation perplexity of a model trained with the objective of Eq. (5) and a baseline trained with the normal loss function of Eq. (2) for different temperatures. As shown, our algorithm achieves relatively low validation perplexity *with respect to the clean distribution* even for high training corruption strengths. We proceed to repeat the idea at a larger scale. We train a GPT-2 124M model on OpenWebText ($\approx$ 9B tokens), and we use it to

Table 1: Validation losses achieved by different fine-tuning strategies for domain adaptation.

Method	Python Code	CS	BioMed	News	Reviews
Target Only (From Scratch)	1.50	3.04	2.66	2.63	3.37
Pre-trained (No Fine-tuning)	1.46	2.87	2.61	2.59	3.56
Target Only Fine-tuning	0.80	2.56	2.36	2.01	**2.96**
Data Mixture Fine-tuning	0.79	2.55	2.36	1.99	**2.96**
Ambient SFT (ours)	**0.70**	**2.54**	**2.35**	**1.68**	**2.96**

produce synthetic data at different temperatures as before. Figure 2a shows examples of training and generated data. The model is capable of producing clean samples even though it has only been trained with text that appears gibberish. These experiments validate that the high-temperature distributions still have usable learning signal, despite the seemingly uninformative training data. Theorem 3.1 establishes that Temperature Matching is exactly equivalent to learning from corrupted samples.

4.2 Domain Adaptation Experiments

We proceed to post-training experiments, beginning with *domain adaptation*: a model pre-trained on generic text needs to be adapted to a new domain from which we don't have much data. We start with an 1.5B-parameter model pre-trained on OpenWebText. We adapt to five target domains, chosen to span diverse text types: Python code (CodeParrot Team, 2022; Chen et al., 2021)), CS papers (Lo et al., 2020), biomedical text (Stuhlmann et al., 2025), news (Hamborg et al., 2017), and product reviews (Hou et al., 2026), each using ≈225M tokens. We benchmark: *(a)* training from scratch on the domain of interest, *(b)* fine-tuning, *(c)* data-mixing (where we split training updates between the pre-training domain and the target domain), *(d)* our Ambient SFT approach. We always pick the hyperparameters that achieve the lowest validation perplexity in the target domain. Our algorithm has two hyperparameters p_c, which controls how often we apply the Temperature Matching loss and the temperature T. We give a full sweep of these parameters in Appendix C.1.2.

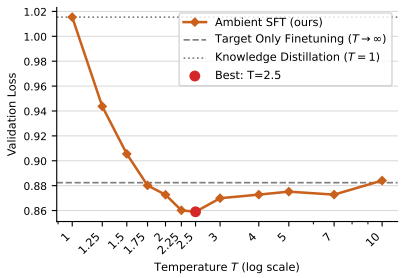
We report results in Table 1. As shown, Ambient SFT matches or outperforms prior baselines across *all* five domains. For News, the relative improvement to SFT in perplexity goes up to 26.7%.

Predicting the success of Ambient SFT. The benefit varies across domains: it is largest on Python Code and News and smaller elsewhere. Because success depends on how much the teacher knows about the target domain, a cheap and reliable predictor of the benefit is the model's perplexity on the target domain before fine-tuning. Across the five domains, this heuristic gives a Spearman rank correlation of $\rho = -0.90$ between zero-shot loss and win margin to the baselines. This is qualitatively consistent with Theorem 3.2, whose bias term depends on the discrepancy $B = A_{\theta^*_{\text{post}}} - A_{\theta^*_{\text{pre}}}$. Full per-domain zero-shot loss and win-margin numbers are reported in Table 14.

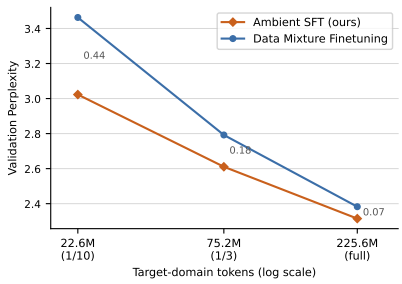
Effect of temperature T. The Temperature Matching term disappears as $T \to \infty$, and it becomes KD for $T = 1$. Figure 3a sweeps T, holding every other setting fixed. Neither extreme is optimal: validation loss falls steadily from $T = 1$ down to a minimum at $T = 2.5$, then rises and plateaus. Beyond this ablation, we clarify that we got the optimal p_c and T values used in Table 1 by optimizing over the *joint* parameter space. So for the optimal $T*$, $T^* \neq 1$.

Effect of the size of the student and teacher. In our experiments so far, the teacher is the frozen model after pre-training and the student is initialized from this teacher. Table 16 isolates the effect of student and teacher capacity, testing four combinations of student size (124M, 1.5B) and teacher size (124M, 1.5B). Two trends are unsurprising: *(a)* a larger student improves results at both teacher sizes, and *(b)* a larger teacher improves results at both student sizes. A third, less obvious finding is that a small (124M) teacher *can* still improve a strictly larger (1.5B) student, a pairing that classical KD would not predict to help. Pairing the 1.5B student with the 124M teacher beats plain fine-tuning of the same 1.5B student with no teacher at all on Python Code and News.

Effect of number of tokens of the target domain. To understand the effect of the size of the fine-tuning dataset, we subsample the Python Code training set, keeping only $1/3$ and $1/10$, and finetune a 124M model on each subsampled set with both Ambient SFT and the Data Mixture baseline. Figure 3b reports the results: our method consistently outperforms the baseline, and the gap widens as the dataset shrinks. This is intuitive: with enough target-domain data, our method should not help,



(a) Effect of temperature T on validation loss.



(b) Effect of dataset size on validation perplexity.

Figure 3: Domain-adaptation ablations on Python Code.

Table 2: Instruction tuning on $\approx$ 21K examples, 1% of Dolci-Instruct-SFT. Top: Qwen3 base models from 1.7B to 14B parameters. Bottom: two other families at the 7–8B scale. Ambient SFT improves on the Vanilla SFT baseline on seven of the eight tasks, consistently across families and sizes; the exception is IFEval, where the gap closes as the model grows. The last column is the average relative improvement.

		MATH	GSM8K	HumanEval+	MMLU	PopQA	TriviaQA	IFEval	BBH	Avg. rel. Δ vs. SFT
Qwen3-1.7B-Base	Vanilla SFT	24.3	61.0	68.0	61.9	12.2	36.3	**58.8**	45.2	—
	Ambient SFT	**40.4**	**72.3**	**83.7**	**64.9**	**14.9**	**43.8**	52.3	**53.3**	+20.3%
Qwen3-4B-Base	Vanilla SFT	36.4	80.2	84.5	72.9	14.8	51.4	**76.3**	63.4	—
	Ambient SFT	**42.2**	**82.9**	**90.1**	**74.3**	**18.1**	**57.5**	71.0	**68.9**	+8.0%
Qwen3-8B-Base	Vanilla SFT	38.7	80.4	82.5	76.2	18.2	65.2	**77.1**	71.4	—
	Ambient SFT	**53.5**	**87.5**	**91.5**	**78.3**	**22.0**	**69.2**	71.7	**72.0**	+10.2%
Qwen3-14B-Base	Vanilla SFT	42.8	86.7	86.9	80.9	21.8	71.6	**78.6**	76.2	—
	Ambient SFT	**58.9**	**89.3**	**92.6**	**81.1**	**24.7**	**75.5**	**78.6**	**78.5**	+8.7%
OLMo-3-7B	Vanilla SFT	22.1	75.2	71.4	65.7	17.5	60.9	**70.8**	61.3	—
	Ambient SFT	**41.5**	**78.6**	**79.5**	**66.4**	**21.0**	**64.8**	64.0	**62.4**	+15.4%
Llama-3.1-8B	Vanilla SFT	9.4	46.0	49.8	58.8	22.1	67.0	**59.0**	50.8	—
	Ambient SFT	**12.4**	**51.3**	**64.9**	**59.7**	**24.5**	**73.7**	54.2	**56.2**	+12.3%

but it is highly effective when data, not compute, is the bottleneck. This matches Theorem 3.2: the bias term is independent of n, while the sampling-variance term is not.

4.3 Instruction-tuning experiments

Having established Ambient SFT as a drop-in replacement for SFT in domain adaptation, we now ask whether it helps elsewhere in the post-training stack, starting with instruction tuning, where a pre-trained LLM is trained via SFT to follow instructions. We start with pre-trained base models, ranging from 1.7B to 14B params, and we fine-tune them on the Dolci-Instruct-SFT (Olmo et al., 2025). We start our evaluation by using only $\approx$ 20K documents ($\approx$ 18M tokens). We validate performance on standard benchmarks that measure instruction following (IFEval (Zhou et al., 2023b)), knowledge (MMLU (Hendrycks et al., 2020), PopQA (Mallen et al., 2023), TriviaQA (Joshi et al., 2017)), math (GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021)), code (HumanEval+ (Chen et al., 2021)), and general reasoning (BBH (Suzgun et al., 2023)). For all our runs, we use $p_c = 0.75$ (i.e., 25% of the updates are spent on the Temperature Matching term) and $T = 2.5$.

Results are presented in Table 2. Ambient SFT leads to consistent average improvements ranging from 8.0% to 20.3%. At every size and model family, Ambient SFT matches or beats vanilla SFT on every benchmark except IFEval, and the IFEval gap closes at the largest size. Interestingly, instruction-following (that IFEval measures) is exactly what we train for. Our method pays a penalty in-distribution, for much better generalization out-of-distribution.

Effect of model size. The Qwen experiments isolate the effect of model size: the gain is largest at the smallest scale but persists as models grow: the average relative improvement is +20.3% at 1.7B and stays above +8.0% from 4B to 14B. On the eight-benchmark average, Ambient SFT at 8B equals vanilla SFT at 14B, and on MATH Ambient SFT at 1.7B already exceeds vanilla SFT at 8B.

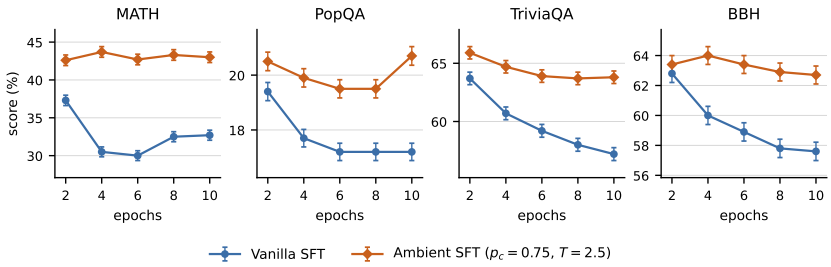


Figure 4: Results versus # of epochs following instruction-tuning on 10% of Dolci-Instruct-SFT.

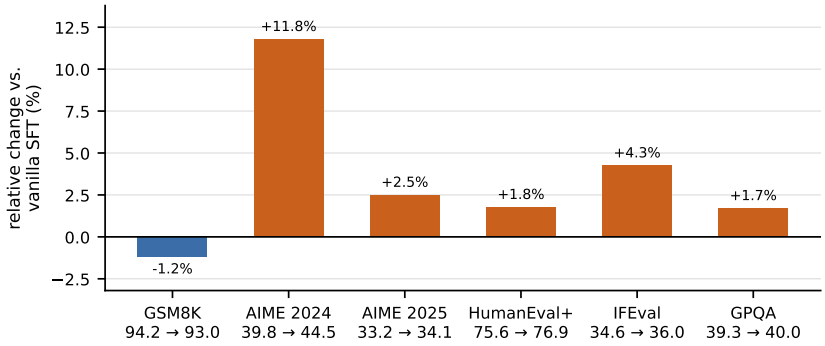


Figure 5: Reasoning results: relative change of Ambient SFT ($p_c = 0.75$, $T = 2.5$) over vanilla SFT.

Effect of dataset size. We now study what happens when we use 1%, 10% and 100% of the dataset. The greatest improvements are observed in the 1% of the dataset setting, as expected, with an average 15.4% relative improvement over vanilla SFT; with 10% of the dataset, the average relative improvement is 3.4%. That said, *even at full data scale*, Ambient SFT leads to an average 8.6% relative improvement to vanilla SFT. Figure 7 shows per-benchmark scores across different scales.

Ambient SFT reduces overfitting. We hypothesize that Ambient SFT reduces overfitting, letting the model benefit from more epochs than vanilla SFT. To test this, we train for 2, 4, 6, 8, and 10 epochs on a 10% subset of Dolci-Instruct-SFT (215k examples). As Figure 4 shows, Ambient SFT reaches its best performance at more epochs and is more robust to overfitting.

4.4 Reasoning Experiments

Finally, we test whether Ambient SFT benefits reasoning. We train OLMo-3-7B on ≈ 22K examples (1%) of Dolci-Think-SFT (Olmo et al., 2025) with the same hyperparameters as before. We evaluate on GSM8K (Cobbe et al., 2021), AIME 2024 and 2025, HumanEval+ (Chen et al., 2021), IFEval (Zhou et al., 2023b), and GPQA (Rein et al., 2023) (details in Appendix C.3.1). As Figure 5 shows, Ambient SFT outperforms vanilla SFT on 5/6 benchmarks and trails by only 1.1 points on GSM8K. While there is a consistent boost in reasoning abilities, its magnitude is smaller compared to the one observed for domain adaptation and instruction-tuning. That said, Table 17 further shows that reasoning models trained with Ambient SFT are also more robust to prompting than pure SFT ones.

5 Conclusion and Future Work

We introduced Ambient SFT, a drop-in replacement for SFT that adds to the standard cross-entropy loss a term matching the model to a high-temperature version of its own pre-trained predictions. Across domain adaptation, instruction tuning, and reasoning, Ambient SFT outperforms vanilla SFT, with the largest gains when fine-tuning data are scarce. A natural next step is to extend this anchoring idea beyond SFT to the reinforcement-learning algorithms used in post-training.

AI USE STATEMENT

In this work, we used generative AI tools for assistance with paper writing, editing figures, developing the mathematical theorems and proofs, implementing parts of the method and conducting the experiments. We have **not** used generative AI tools for conceptualization of the project, results interpretation and developing the structure and the narrative of the paper. We have reviewed all AI-assisted work. In particular, we reviewed the code, the proofs, all the paper edits and parsing of the results of the experiments. The authors fully understand and take ownership of all the materials presented in this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

CodeParrot Team. Codeparrot clean train dataset. `https://huggingface.co/datasets/codeparrot/codeparrot-clean-train`, 2022. Originally developed as part of the code generation models for the book 'Natural Language Processing with Transformers' by Lewis Tunstall, Leandro von Werra, and Thomas Wolf.

Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alex Dimakis, and Adam Klivans. Ambient diffusion: Learning clean distributions from corrupted data. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 288–313. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/012af729c5d14d279581fc8a5db975a1-Paper-Conference.pdf`.

Giannis Daras, Alexandros G. Dimakis, and Constantinos Daskalakis. Consistent diffusion meets tweedie: training exact ambient diffusion models with noisy data. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Giannis Daras, Adrian Rodriguez-Munoz, Adam Klivans, Antonio Torralba, and Constantinos Daskalakis. Ambient diffusion omni: Training good models with bad data. *arXiv preprint arXiv:2506.10038*, 2025.

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

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. OLMES: A standard for language model evaluations. *arXiv preprint arXiv:2406.08446*, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. news-please - a generic news crawler and extractor. In *Intelligence and Security Informatics*, 2017. URL `https://api.semanticscholar.org/CorpusID:5830937`.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

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

Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation: Benchmarking llms as semantic encoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3251–3265, 2026.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Andrej Karpathy. char-rnn. `https://github.com/karpathy/char-rnn`, 2015.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2023.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, et al. Mitigating the alignment tax of rlhf. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 580–606, 2024.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4969–4983, 2020.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Adrián Rodríguez-Muñoz, William Daspit, Adam Klivans, Antonio Torralba, Constantinos Daskalakis, and Giannis Daras. Ambient dataloops: Generative models for dataset refinement. *arXiv preprint arXiv:2601.15417*, 2026.

Linus Stuhlmann, Michael Alexander Saxer, and Jonathan Fürst. Efficient and reproducible biomedical question answering using retrieval augmented generation. In *2025 IEEE Swiss Conference on Data Science (SDS)*, pp. 154–157. IEEE, 2025.

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

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1028–1043, 2024.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3713–3722, 2019.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023a.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023b.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

APPENDIX: CONTENTS

A Theory **14**

A.1 Proof of Theorem 3.1 14

A.2 Proof of Theorem A.1 14

A.3 Proofs for Section 3.2.3 (log-linear model) 16

B Training Algorithm **21**

C Experimental Details **22**

C.1 Domain Adaptation 22

C.1.1 Datasets 22

C.1.2 Full (p, T) sweep and optimal hyperparameters 23

C.2 Instruction Tuning 24

C.2.1 Training hyperparameters 24

C.2.2 Evaluation harness 25

C.3 Reasoning 26

C.3.1 Evaluation harness 26

D Additional Results **27**

D.1 Full temperature sweep on Tiny Shakespeare 27

D.2 Domain Adaptation 27

D.2.1 Predicting the success of Ambient SFT 27

D.2.2 Ablation: effect of a smaller student 28

D.2.3 Ablation: effect of a weaker teacher 28

D.3 Instruction Tuning 29

D.4 Reasoning 29

E Related Work **30**

A THEORY

A.1 PROOF OF THEOREM 3.1

Proof. (a) For distributions q, r on V with r of full support, $D_{\mathrm{KL}}(q\|r) = \sum_y q(y)\log q(y) - \sum_y q(y)\log r(y) = \mathbb{E}_{y\sim q}[-\log r(y)] - H(q)$. For any $\mathbf{x}_{<i}$, we can plug into this identify $q = p_{T\theta^*}(\cdot \mid \mathbf{x}_{<i})$ and $r = p_{T\theta}(\cdot \mid \mathbf{x}_{<i})$, to get the following:

$$D_{\mathrm{KL}}(p_{T\theta^*}(\cdot \mid \mathbf{x}_{<i})\|p_{T\theta}(\cdot \mid \mathbf{x}_{<i})) = \mathbb{E}_{x_i\sim p_{T\theta^*}(\cdot|\mathbf{x}_{<i})}\big[-\log p_{T\theta}(x_i \mid \mathbf{x}_{<i})\big] - H(p_{T\theta^*}(\cdot \mid \mathbf{x}_{<i})). \tag{11}$$

Note that $p_{T\theta^*}(\mathbf{x})$ is by definition the autoregressive product of the tempered conditionals, so under $\mathbf{x} \sim p_{T\theta^*}$ the prefix $\mathbf{x}_{<i}$ has marginal $p_{T\theta^*}(\mathbf{x}_{<i}) = \prod_{j<i} p_{T\theta^*}(x_j \mid \mathbf{x}_{<j})$ and, given the prefix, $x_i \sim p_{T\theta^*}(\cdot \mid \mathbf{x}_{<i})$. By the tower rule, and using (11):

$$\begin{aligned} J_T(\theta) &= \sum_{i=1}^{l} \mathbb{E}_{\mathbf{x}_{<i}\sim p_{T\theta^*}}\, \mathbb{E}_{x_i\sim p_{T\theta^*}(\cdot|\mathbf{x}_{<i})}\big[-\log p_{T\theta}(x_i \mid \mathbf{x}_{<i})\big] \\ &= \sum_{i=1}^{l} \mathbb{E}_{\mathbf{x}_{<i}\sim p_{T\theta^*}}\Big[D_{\mathrm{KL}}\big(p_{T\theta^*}(\cdot \mid \mathbf{x}_{<i}) \,\big\|\, p_{T\theta}(\cdot \mid \mathbf{x}_{<i})\big)\Big] + \text{const}, \end{aligned}$$

i.e., $J_T(\theta) = R_T(\theta) + \text{const}$, where the constant depends on θ^* but not θ.

(b) For all θ, we have that: $J_T(\theta) - J_T(\theta^*) = \sum_i \mathbb{E}_{\mathbf{x}_{<i}\sim p_{T\theta^*}}[D_{\mathrm{KL}}(p_{T\theta^*}(\cdot \mid \mathbf{x}_{<i})\|p_{T\theta}(\cdot \mid \mathbf{x}_{<i}))] \geq 0$. So θ^* is clearly a minimizer. Moreover, θ is a minimizer iff every KL term with positive weight vanishes. Because each tempered conditional is a softmax, $p_{T\theta^*}(\mathbf{x}_{<i}) > 0$ for every $\mathbf{x}_{<i} \in V^{i-1}$, so all contexts receive positive weight, and θ is a minimizer iff $p_{T\theta}(\cdot \mid \mathbf{x}_{<i}) = p_{T\theta^*}(\cdot \mid \mathbf{x}_{<i})$ for every context. Equivalently, this holds iff $g_\theta(y \mid \mathbf{x}_{<i})/T - g_{\theta^*}(y \mid \mathbf{x}_{<i})/T$ is a constant function of $y \in V$ for every context, i.e., iff $g_\theta(\cdot \mid \mathbf{x}_{<i})/T = g_{\theta^*}(\cdot \mid \mathbf{x}_{<i})/T + c(\mathbf{x}_{<i})$ for every context, i.e. iff $g_\theta(\cdot \mid \mathbf{x}_{<i})/T' = g_{\theta^*}(\cdot \mid \mathbf{x}_{<i})/T' + c(\mathbf{x}_{<i})$ for every context and all $T' > 0$, i.e. iff $p_{T'\theta}(\cdot \mid \mathbf{x}_{<i}) = p_{T'\theta^*}(\cdot \mid \mathbf{x}_{<i})$ for every context and every $T' > 0$. □

A.2 PROOF OF THEOREM A.1

Theorem A.1 (High-temperature conditional KL convergence)**.** *Fix a context $x_{<i}$ and let $\Delta(x_{<i})$ be the range of the logit gap between the two models,*

$$\Delta(x_{<i}) = \max_{y\in\mathcal{V}}\big[g_{\theta^*_{\mathrm{post}}}(y \mid x_{<i}) - g_{\theta^*_{\mathrm{pre}}}(y \mid x_{<i})\big] - \min_{y\in\mathcal{V}}\big[g_{\theta^*_{\mathrm{post}}}(y \mid x_{<i}) - g_{\theta^*_{\mathrm{pre}}}(y \mid x_{<i})\big] \tag{12}$$

$$= \max_{y\in\mathcal{V}}\big[g_{\theta^*_{\mathrm{pre}}}(y \mid x_{<i}) - g_{\theta^*_{\mathrm{post}}}(y \mid x_{<i})\big] - \min_{y\in\mathcal{V}}\big[g_{\theta^*_{\mathrm{pre}}}(y \mid x_{<i}) - g_{\theta^*_{\mathrm{post}}}(y \mid x_{<i})\big] \tag{13}$$

For every $T > 0$:

$$D_{\mathrm{KL}}\Big(p_{T\theta^*_{\mathrm{post}}}(\cdot \mid x_{<i}) \,\Big\|\, p_{T\theta^*_{\mathrm{pre}}}(\cdot \mid x_{<i})\Big) \leq \frac{\Delta(x_{<i})^2}{8T^2}.$$

The same bound holds with the two models interchanged.

Proof. The KL divergence of interest is

$$\begin{aligned} &D_{\mathrm{KL}}\Big(p_{T\theta^*_{\mathrm{post}}}(\cdot \mid x_{<i}) \,\Big\|\, p_{T\theta^*_{\mathrm{pre}}}(\cdot \mid x_{<i})\Big) \\ &\quad= \mathbb{E}_{Y\sim p_{T\theta^*_{\mathrm{post}}}(\cdot|x_{<i})}\left[\log \frac{p_{T\theta^*_{\mathrm{post}}}(Y \mid x_{<i})}{p_{T\theta^*_{\mathrm{pre}}}(Y \mid x_{<i})}\right]. \end{aligned}$$

We will write the KL divergence in a form to which Hoeffding's lemma can be applied. Before we do this, let us notice the following easy fact:

$$\begin{aligned} \mathbb{E}_{Y\sim p_{T\theta^*_{\mathrm{post}}}(\cdot|x_{<i})}\left[\exp\left(-\log \frac{p_{T\theta^*_{\mathrm{post}}}(Y \mid x_{<i})}{p_{T\theta^*_{\mathrm{pre}}}(Y \mid x_{<i})}\right)\right] &= \sum_{y\in\mathcal{V}} p_{T\theta^*_{\mathrm{post}}}(y \mid x_{<i}) \frac{p_{T\theta^*_{\mathrm{pre}}}(y \mid x_{<i})}{p_{T\theta^*_{\mathrm{post}}}(y \mid x_{<i})} \\ &= \sum_{y\in\mathcal{V}} p_{T\theta^*_{\mathrm{pre}}}(y \mid x_{<i}) = 1. \end{aligned}$$

We claim that this implies the following:

$$D_{\mathrm{KL}}\Big(p_{T\theta^*_{\mathrm{post}}}(\cdot \mid x_{<i}) \,\Big\|\, p_{T\theta^*_{\mathrm{pre}}}(\cdot \mid x_{<i})\Big) = \log \mathbb{E}\left[\exp\left(-\log \frac{p_{T\theta^*_{\mathrm{post}}}(Y \mid x_{<i})}{p_{T\theta^*_{\mathrm{pre}}}(Y \mid x_{<i})} + \mathbb{E}\left[\log \frac{p_{T\theta^*_{\mathrm{post}}}(Y \mid x_{<i})}{p_{T\theta^*_{\mathrm{pre}}}(Y \mid x_{<i})}\right]\right)\right]. \tag{16}$$

Indeed, the RHS of the above equals:

$$\begin{aligned}
&\log\left(\mathbb{E}\left[\exp\left(-\log \frac{p_{T\theta^*_{\mathrm{post}}}(Y \mid x_{<i})}{p_{T\theta^*_{\mathrm{pre}}}(Y \mid x_{<i})}\right)\right] \cdot \exp\left(\mathbb{E}\left[\log \frac{p_{T\theta^*_{\mathrm{post}}}(Y \mid x_{<i})}{p_{T\theta^*_{\mathrm{pre}}}(Y \mid x_{<i})}\right]\right)\right) \\
&\qquad = \log\left(1 \cdot \exp\left(\mathbb{E}\left[\log \frac{p_{T\theta^*_{\mathrm{post}}}(Y \mid x_{<i})}{p_{T\theta^*_{\mathrm{pre}}}(Y \mid x_{<i})}\right]\right)\right) \\
&\qquad = \mathbb{E}\left[\log \frac{p_{T\theta^*_{\mathrm{post}}}(Y \mid x_{<i})}{p_{T\theta^*_{\mathrm{pre}}}(Y \mid x_{<i})}\right].
\end{aligned}$$

Recall that Hoeffding's lemma implies that, for any random variable Z supported on an interval $[a, b]$,

$$\log \mathbb{E}\left[\exp\left(X - \mathbb{E}[X]\right)\right] \le \frac{(b-a)^2}{8}.$$

We apply Hoeffding's lemma to the random variable

$$Z := -\log \frac{p_{T\theta^*_{\mathrm{post}}}(Y \mid x_{<i})}{p_{T\theta^*_{\mathrm{pre}}}(Y \mid x_{<i})}.$$

This random variable is bounded because the vocabulary is finite and all logits are finite. Its centered exponential moment is exactly the right-hand side of (16). Therefore,

$$D_{\mathrm{KL}}\Big(p_{T\theta^*_{\mathrm{post}}}(\cdot \mid x_{<i}) \,\Big\|\, p_{T\theta^*_{\mathrm{pre}}}(\cdot \mid x_{<i})\Big) \le \frac{1}{8}\left(\max_{y \in \mathcal{V}} \log \frac{p_{T\theta^*_{\mathrm{post}}}(y \mid x_{<i})}{p_{T\theta^*_{\mathrm{pre}}}(y \mid x_{<i})} - \min_{y \in \mathcal{V}} \log \frac{p_{T\theta^*_{\mathrm{post}}}(y \mid x_{<i})}{p_{T\theta^*_{\mathrm{pre}}}(y \mid x_{<i})}\right)^2, \tag{17}$$

where we used the fact that negating a random variable does not change the length of its range.

By the definition of temperature scaling,

$$\log \frac{p_{T\theta^*_{\mathrm{post}}}(y \mid x_{<i})}{p_{T\theta^*_{\mathrm{pre}}}(y \mid x_{<i})} = \frac{g_{\theta^*_{\mathrm{post}}}(y \mid x_{<i}) - g_{\theta^*_{\mathrm{pre}}}(y \mid x_{<i})}{T} + \log \frac{\sum_{z \in \mathcal{V}} \exp\Big(g_{\theta^*_{\mathrm{pre}}}(z \mid x_{<i})/T\Big)}{\sum_{z \in \mathcal{V}} \exp\Big(g_{\theta^*_{\mathrm{post}}}(z \mid x_{<i})/T\Big)}.$$

The second term is independent of y and therefore cancels when taking the difference between the maximum and minimum over tokens. Consequently,

$$\max_{y \in \mathcal{V}} \log \frac{p_{T\theta^*_{\mathrm{post}}}(y \mid x_{<i})}{p_{T\theta^*_{\mathrm{pre}}}(y \mid x_{<i})} - \min_{y \in \mathcal{V}} \log \frac{p_{T\theta^*_{\mathrm{post}}}(y \mid x_{<i})}{p_{T\theta^*_{\mathrm{pre}}}(y \mid x_{<i})} = \frac{\Delta(x_{<i})}{T}.$$

Substituting this identity into (17) gives

$$D_{\mathrm{KL}}\Big(p_{T\theta^*_{\mathrm{post}}}(\cdot \mid x_{<i}) \,\Big\|\, p_{T\theta^*_{\mathrm{pre}}}(\cdot \mid x_{<i})\Big) \le \frac{\Delta(x_{<i})^2}{8T^2}.$$

□

A.3 Proofs for Section 3.2.3 (log-linear model)

Notation. Starred parameters are minimizers of population objectives while hats mark estimators computed from the n observations. $\widehat{\theta}_\lambda$ denotes the minimizer of our obejctive and $\widehat{\theta}_0$ denotes the clean maximum-likelihood estimator ($\lambda = 0$). θ^*_{pre} is the teacher and θ^*_{post} is the target distribution.

Regularity conditions for Theorem 3.2. The minimizer $\widehat{\theta}_\lambda$ of (8) is taken over a fixed ball $\mathcal{B}_R = \{\|A_\theta\|_{\mathrm{op}} \le R\}$ containing θ^*_{pre} and θ^*_{post}, a technical restriction that keeps all logits bounded. The embeddings are bounded, $\|\phi(y)\| \le R_\phi$ for every $y \in V$ and $\|\psi\| \le R_\psi$ almost surely, so all logits on $\mathcal{B}_R$ are bounded by $\varrho = RR_\phi R_\psi$. We assume that the minimizer is in the interior of the ball. We also recall the two non-degeneracy conditions of Section 3.2.3: $G = \mathbb{E}[\psi\psi^\top] \succ 0$, and the embedding differences $\phi(y) - \phi(z)$, $y, z \in V$, span $\mathbb{R}^d$. The second is exactly identifiability of A_θ: if some $v \neq 0$ had $\phi(y)^\top v$ constant in y, then $\Delta = vu^\top$ would leave every conditional unchanged for every $u \in \mathbb{R}^m$. Together they give $\Sigma(\eta) \succ 0$ for every η and hence $F \succ 0$, $H_T \succ 0$: for $\Delta \neq 0$, $\langle \Delta, F\Delta\rangle = \mathbb{E}[(\Delta\psi)^\top \Sigma(A_{\theta^*_{\mathrm{post}}}\psi)(\Delta\psi)] > 0$ because $G \succ 0$.

The non-degeneracy condition of Theorem 3.2. Assume in addition that $B \neq 0$ and that (A2) $D := \partial_T H_T|_{T=1}$ is not a multiple of F. The claim about $T = 1$ additionally excludes a closed Lebesgue-null cone of teacher gaps B, which cannot be dispensed with: Remark A.6 exhibits a model in which $T = 1$ is the unique optimizer within the family. We will show that under these conditions the inequality in (10) is strict: some finite (λ, T) with $T \neq 1$ makes Ambient SFT strictly better than maximum likelihood, than the frozen teacher and than every knowledge-distillation run. We do not claim that the infimum is attained, and Remark A.7 shows that it need not be: the reachable penalty shapes form a curve that is open at its hot end.

Throughout, $\Phi \in \mathbb{R}^{|V|\times d}$ stacks the token embeddings $\phi(z)^\top$, so that $\Lambda(\eta) = \mathrm{lse}(\Phi\eta)$ with $\mathrm{lse}(u) = \log\sum_z e^{u_z}$, $\nabla\Lambda(\eta) = \Phi^\top\sigma(\Phi\eta) = \mu(\eta)$ and $\nabla^2\Lambda(\eta) = \Phi^\top\Sigma(\sigma(\Phi\eta))\Phi = \Sigma(\eta)$, where σ is the softmax and $\Sigma(p) = \mathrm{diag}(p) - pp^\top$. We identify $A_\theta \in \mathbb{R}^{d\times m}$ with $\mathrm{vec}(A_\theta)$, so that for a linear map $\theta \mapsto A_\theta\psi$ the chain rule gives $\nabla_\theta h(A_\theta\psi) = \nabla h(A_\theta\psi)\,\psi^\top$ and $\nabla^2_\theta h(A_\theta\psi) = \nabla^2 h(A_\theta\psi) \otimes \psi\psi^\top$.

Proposition A.2 (Structure of the objective)**.** *For every $\lambda > 0$ and $T > 0$:*

(i) $\widehat{J}_{\lambda,T}$ is convex in θ, and strictly convex along every direction Δ that changes some logit difference on some observed context, i.e., $\phi(y)^\top\Delta\psi_t \neq \phi(z)^\top\Delta\psi_t$ for some t and $y, z \in V$. Hence its minimizer is unique up to directions that leave every conditional $p_\theta(\cdot \mid \mathbf{x}_{<t})$ unchanged.

(ii) The tempreature matching term is a Bregman divergence centered at the pre-trained model: $\widehat{R}_T(\theta) = \frac{1}{n}\sum_t D_T(A_\theta; A_{\theta^*_{\mathrm{pre}}}, \psi_t) + \mathrm{const}$, where

$$D_T(A_\theta; A_{\theta^*_{\mathrm{pre}}}, \psi) = \Lambda\big(\tfrac{A_\theta\psi}{T}\big) - \Lambda\big(\tfrac{A_{\theta^*_{\mathrm{pre}}}\psi}{T}\big) - \mu\big(\tfrac{A_{\theta^*_{\mathrm{pre}}}\psi}{T}\big)^\top \tfrac{(A_\theta - A_{\theta^*_{\mathrm{pre}}})\psi}{T} \;\ge 0,$$

with equality iff $p_{TA_\theta}(\cdot \mid \mathbf{x}_{<i}) = p_{TA_{\theta^*_{\mathrm{pre}}}}(\cdot \mid \mathbf{x}_{<i})$. Thus $\widehat{J}_{\lambda,T}$ is a penalized maximum likelihood with a data-dependent penalty that vanishes exactly at $A_{\theta^*_{\mathrm{pre}}}$.

(iii) (Moment matching.) Its gradient is

$$\nabla_\theta \widehat{J}_{\lambda,T}(\theta) = \frac{1}{n}\sum_{t=1}^{n}\Big[\big(\mu(A_\theta\psi_t) - \phi(x_t)\big) + \frac{\lambda}{T}\big(\mu(\tfrac{A_\theta\psi_t}{T}) - \mu(\tfrac{A_{\theta^*_{\mathrm{pre}}}\psi_t}{T})\big)\Big]\psi_t^\top,$$

so at a minimizer the model's mean embedding matches the data on average, corrected by λ/T times the gap between the student's and teacher's tempered mean embeddings.

(iv) (Curvature.) Its Hessian, as an operator on $\mathbb{R}^{d\times m}$, is

$$\nabla^2_\theta \widehat{J}_{\lambda,T}(\theta) = \frac{1}{n}\sum_{t=1}^{n}\Big[\Sigma(A_\theta\psi_t) + \frac{\lambda}{T^2}\Sigma\big(\tfrac{A_\theta\psi_t}{T}\big)\Big] \otimes \psi_t\psi_t^\top,$$

and as $T \to \infty$, $\Sigma(A_\theta\psi_t/T) \to \Sigma_0 := \mathrm{Cov}_{\mathrm{unif}(V)}[\phi(y)]$ for every θ. The teacher therefore contributes curvature $\frac{\lambda}{T^2}\,\Sigma_0 \otimes \widehat{G} + o(T^{-2})$ with $\widehat{G} = \frac{1}{n}\sum_t \psi_t\psi_t^\top$: a ridge penalty toward

$A_{\theta^*_{\mathrm{pre}}}$ *whose strength is set by* λ/T^2 *and whose geometry is fixed by the embeddings and the context covariance.*

Proof of Proposition A.2. (i) lse is convex, and $u \mapsto \Phi A_\theta \psi_t$ and $u \mapsto \Phi A_\theta \psi_t / T$ are linear in θ, so every summand of (8) is convex in θ; the linear terms do not affect convexity. For strictness, $\nabla^2 \operatorname{lse}(u) = \Sigma(\sigma(u))$ and $v^\top \Sigma(\sigma(u)) v = \operatorname{Var}_{\sigma(u)}(v)$, which is positive unless v is constant on V. Along a direction Δ, the second directional derivative of the t-th clean summand is $\operatorname{Var}_{p_\theta(\cdot|\mathbf{x}_{<t})}\big(\phi(y)^\top \Delta \psi_t\big)$, positive iff $y \mapsto \phi(y)^\top \Delta \psi_t$ is non-constant; the matching summand contributes the same quantity at temperature T times λ/T^2. Summing over t gives strict convexity along every Δ that changes a logit difference on some observed context, and flatness along the others; the latter are exactly the directions that leave every conditional unchanged, since a conditional depends on $A_\theta \psi_t$ only through its logit differences.

(ii) Expanding D_T and comparing with the t-th matching summand, $\Lambda(\frac{A_\theta \psi_t}{T}) - \mu(\frac{A_{\theta^*_{\mathrm{pre}}} \psi_t}{T})^\top \frac{A_\theta \psi_t}{T} = D_T(A_\theta; A_{\theta^*_{\mathrm{pre}}}, \psi_t) + \Lambda(\frac{A_{\theta^*_{\mathrm{pre}}} \psi_t}{T}) - \mu(\frac{A_{\theta^*_{\mathrm{pre}}} \psi_t}{T})^\top \frac{A_{\theta^*_{\mathrm{pre}}} \psi_t}{T}$, and the last two terms do not depend on θ. D_T is the Bregman divergence of the convex function $\eta \mapsto \Lambda(\eta/T)$ between $A_\theta \psi_t$ and $A_{\theta^*_{\mathrm{pre}}} \psi_t$, hence nonnegative, and zero iff $\Lambda(\cdot/T)$ is affine on the segment joining them, i.e., iff $\Phi(A_\theta - A_{\theta^*_{\mathrm{pre}}})\psi_t$ is constant on V by the variance computation in (i), i.e., iff the two tempered conditionals coincide. Equivalently $D_T(A_\theta; A_{\theta^*_{\mathrm{pre}}}, \psi_t) = D_{\mathrm{KL}}\big(p_{TA_{\theta^*_{\mathrm{pre}}}}(\cdot \mid \mathbf{x}_{<t}) \,\|\, p_{T\theta(\cdot|\mathbf{x}_{<t})}\big)$, the standard identity between Bregman divergences of the log-partition and KL divergences in an exponential family.

(iii) and *(iv)* follow from the chain rule stated above applied to $\Lambda(A_\theta \psi_t)$, $\phi(x_t)^\top A_\theta \psi_t$, $\Lambda(A_\theta \psi_t / T)$ and $\mu(\cdot)^\top A_\theta \psi_t / T$; the factor $1/T$ from differentiating $A_\theta \psi_t / T$ appears once in the gradient and twice in the Hessian. Finally $\sigma(\Phi\eta/T) \to \sigma(0) = \mathbf{1}/|V|$ as $T \to \infty$ for every fixed η, so $\Sigma(\eta/T) \to \Phi^\top \Sigma(\mathbf{1}/|V|)\Phi = \operatorname{Cov}_{\mathrm{unif}(V)}[\phi(y)] = \Sigma_0$, uniformly on bounded sets of η. □

Setting for Theorem 3.2. Let $\ell_t(\theta)$ denote the t-th summand of (8) (clean plus λ times matching), so $\widehat{J}_{\lambda,T} = \frac{1}{n}\sum_t \ell_t$ and $J_{\lambda,T} = \mathbb{E}[\ell_t]$, the expectation over $\psi \sim \mathcal{D}_x$ and $x \sim p_{A_{\theta^*_{\mathrm{post}}}}(\cdot \mid \psi)$. We consider the local regime $B = B_0/\sqrt{n}$ with B_0 fixed, so $\|B\| = O(n^{-1/2})$; all $O(\cdot)$ constants below are uniform. Write $H_T^{\mathrm{pre}} = \nabla^2 R_T(\theta^*_{\mathrm{pre}}) = \frac{1}{T^2}\mathbb{E}[\Sigma(A_{\theta^*_{\mathrm{pre}}}\psi/T) \otimes \psi\psi^\top]$ for the curvature of the penalty at its center; $H_T^{\mathrm{pre}} = H_T + O(\|B\|)$ by smoothness of Σ. F and H_T are positive definite by the non-degeneracy conditions above.

Lemma A.3 (Consistency, asymptotic normality, second moments)**.** *For fixed* λ, T *and* B: $\widehat{\theta}_\lambda \to \theta^*_\lambda$ *almost surely and* $\sqrt{n}(\widehat{\theta}_\lambda - \theta^*_\lambda) \Rightarrow \mathcal{N}(0, F_\lambda^{-1} V_\lambda F_\lambda^{-1})$*, with* $F_\lambda = \nabla^2 J_{\lambda,T}(\theta^*_\lambda)$ *and* $V_\lambda = \operatorname{Cov}[\nabla \ell_t(\theta^*_\lambda)]$. *Moreover* $V_\lambda = F + O(\|B\|)$*,* $F_\lambda = F + \lambda H_T + O(\|B\|)$*,* $\mathbb{E}[\widehat{\theta}_\lambda] - \theta^*_\lambda = O(1/n)$ *and* $n\,\mathbb{E}\|\widehat{\theta}_\lambda - \theta^*_\lambda\|^2 = \operatorname{tr}(F_\lambda^{-1} V_\lambda F_\lambda^{-1}) + O(n^{-1/2})$.

Proof. Each ℓ_t is convex and C^∞ in θ with bounded derivatives on bounded sets (the logits are bounded by $\|A_\theta\|\,\|\phi\|\,\|\psi\|$), and $J_{\lambda,T}$ is strictly convex with positive definite Hessian $\nabla^2 J_{\lambda,T}(\theta) = \mathbb{E}[\Sigma(A_\theta\psi) \otimes \psi\psi^\top] + \frac{\lambda}{T^2}\mathbb{E}[\Sigma(\frac{A_\theta\psi}{T}) \otimes \psi\psi^\top]$, so it has a unique minimizer θ^*_λ. Pointwise convergence $\widehat{J}_{\lambda,T} \to J_{\lambda,T}$ (strong law) together with convexity gives uniform convergence on compacts and hence $\widehat{\theta}_\lambda \to \theta^*_\lambda$ almost surely. Asymptotic normality is the classical result for smooth M-estimators: expanding the first-order condition $\sum_t \nabla \ell_t(\widehat{\theta}_\lambda) = 0$ around θ^*_λ, $0 = \frac{1}{\sqrt{n}}\sum_t \nabla \ell_t(\theta^*_\lambda) + \Big[\frac{1}{n}\sum_t \nabla^2 \ell_t(\theta^*_\lambda) + o_P(1)\Big]\sqrt{n}(\widehat{\theta}_\lambda - \theta^*_\lambda)$, where the first term is a sum of i.i.d. mean-zero (by optimality of θ^*_λ) bounded vectors and converges to $\mathcal{N}(0, V_\lambda)$ with $V_\lambda = \operatorname{Cov}[\nabla \ell_t(\theta^*_\lambda)]$, and the bracket converges to $F_\lambda = \nabla^2 J_{\lambda,T}(\theta^*_\lambda)$. Slutsky gives $\sqrt{n}(\widehat{\theta}_\lambda - \theta^*_\lambda) \Rightarrow \mathcal{N}(0, F_\lambda^{-1} V_\lambda F_\lambda^{-1})$. Writing $\nabla \ell_t(\theta) = (\mu(A_\theta\psi_t) - \phi(x_t))\psi_t^\top + \frac{\lambda}{T}(\mu(\frac{A_\theta\psi_t}{T}) - \mu(\frac{A_{\theta^*_{\mathrm{pre}}}\psi_t}{T}))\psi_t^\top$ split it at $\theta = \theta^*_{\mathrm{post}}$ into the clean part $s_{\mathrm{clean}} = (\mu(A_{\theta^*_{\mathrm{post}}}\psi_t) - \phi(x_t))\psi_t^\top$ and the teacher part $r_T(\psi_t) = \frac{\lambda}{T}(\mu(\frac{A_{\theta^*_{\mathrm{post}}}\psi_t}{T}) - \mu(\frac{A_{\theta^*_{\mathrm{pre}}}\psi_t}{T}))\psi_t^\top$, which is a deterministic function of ψ_t. Conditionally on ψ_t the only randomness is $\phi(x_t)$, with covariance $\Sigma(A_{\theta^*_{\mathrm{post}}}\psi_t)$, so s_{clean} has conditional mean zero

and the law of total covariance gives $\mathrm{Cov}[\nabla \ell_t(\theta^*_{\mathrm{post}})] = \mathbb{E}[\Sigma(A_{\theta^*_{\mathrm{post}}}\psi_t) \otimes \psi_t\psi_t^\top] + \mathrm{Cov}(r_T(\psi_t)) = F + O(\|B\|^2)$, the second term being $O(\|B\|^2)$ because $r_T = O(\|B\|)$ uniformly. Moving the evaluation point from θ^*_{post} to θ^*_λ, and since $A_{\theta^*_\lambda} - A_{\theta^*_{\mathrm{post}}} = O(\|B\|)$ by Lemma A.4 and $\nabla \ell_t$ is smooth, $V_\lambda = F + O(\|B\|)$ and $F_\lambda = F + \lambda H_T + O(\|B\|)$.

For the second moments, one cannot simply condition on the prefixes: a finite sample may miss feature directions, and then $\widehat{J}_{\lambda,T}$ is not strongly convex for that realization. Work instead on the event $E_n = \{\widehat{G}_n \succeq \frac{1}{2}G\}$, where $\widehat{G}_n = \frac{1}{n}\sum_t \psi_t\psi_t^\top$. The summands $\psi_t\psi_t^\top$ are i.i.d. and positive semidefinite with $\|\psi_t\psi_t^\top\|_{\mathrm{op}} \le R_\psi^2$ and mean $G \succ 0$, so the matrix Chernoff bound gives $\mathbb{P}(E_n^c) \le m\, e^{-cn}$ with $c = \lambda_{\min}(G)/(8R_\psi^2)$. On E_n the empirical objective is strongly convex on $\mathcal{B}_R$: the set $\{A_\theta\psi : \theta \in \mathcal{B}_R, \|\psi\| \le R_\psi\}$ is compact and $\Sigma(\cdot) \succ 0$ on it by the two non-degeneracy conditions, so $\sigma_0 := \inf \lambda_{\min}(\Sigma(A_\theta\psi)) > 0$ there and, for every Δ, $\langle \Delta, \nabla^2 \widehat{J}_{\lambda,T}(\theta)\Delta\rangle \ge \frac{1}{n}\sum_t (\Delta\psi_t)^\top \Sigma(A_\theta\psi_t)(\Delta\psi_t) \ge \sigma_0 \operatorname{tr}(\Delta \widehat{G}_n \Delta^\top) \ge \frac{1}{2}\sigma_0\lambda_{\min}(G)\|\Delta\|_F^2$, the first inequality discarding the matching term of the Hessian, which is positive semidefinite. With that modulus κ, the two-point inequality at the interior minimizer θ^*_λ of $J_{\lambda,T}$ gives $\kappa\|\widehat{\theta}_\lambda - \theta^*_\lambda\| \le \|\nabla \widehat{J}_{\lambda,T}(\theta^*_\lambda)\|$ on E_n, and $\nabla \widehat{J}_{\lambda,T}(\theta^*_\lambda)$ is an average of i.i.d. bounded terms with mean $\nabla J_{\lambda,T}(\theta^*_\lambda) = 0$, so $\mathbb{E}\|\nabla \widehat{J}_{\lambda,T}(\theta^*_\lambda)\|^2 = O(1/n)$ and $\mathbb{E}[\|\widehat{\theta}_\lambda - \theta^*_\lambda\|^2 \mathbf{1}_{E_n}] \le C/n$. On E_n^c we use only that both iterates lie in $\mathcal{B}_R$, an operator-norm ball, so $\|A_{\widehat{\theta}_\lambda} - A_{\theta^*_\lambda}\|_F \le 2R\sqrt{\min(d,m)}$ and that part contributes at most $4R^2 \min(d,m)\, m\, e^{-cn}$, which is $o(n^{-p})$ for every p; hence $\mathbb{E}\|\widehat{\theta}_\lambda - \theta^*_\lambda\|^2 \le C'/n$ for every n. The same event carries the refinement that follows: all derivatives of ℓ_t are bounded on $\mathcal{B}_R$, so the one-step expansion of the first-order condition can be carried one order further on E_n: $\widehat{\theta}_\lambda - \theta^*_\lambda = -F_\lambda^{-1}\frac{1}{n}\sum_t \nabla \ell_t(\theta^*_\lambda) + \Delta_n$ with $\mathbb{E}[\|\Delta_n\|^2 \mathbf{1}_{E_n}] \le C/n^2$ and the complement again exponentially negligible. That bound is where the fourth moments enter: writing $S_n = \frac{1}{n}\sum_t \nabla \ell_t(\theta^*_\lambda)$ for the empirical score and $Q_n = \frac{1}{n}\sum_t \nabla^2 \ell_t(\theta^*_\lambda) - F_\lambda$ for the Hessian fluctuation, both are averages of bounded i.i.d. mean-zero terms, so $\mathbb{E}\|S_n\|^4 = O(n^{-2})$ and $\mathbb{E}\|Q_n\|^4 = O(n^{-2})$; on E_n the curvature bound gives $\|\widehat{\theta}_\lambda - \theta^*_\lambda\| \le \kappa^{-1}\|S_n\|$, and the second-order Taylor remainder of the first-order condition is bounded by $\|Q_n\|\|\widehat{\theta}_\lambda - \theta^*_\lambda\|$ plus a third-derivative term of order $\|\widehat{\theta}_\lambda - \theta^*_\lambda\|^2$, so $\mathbb{E}[\|\Delta_n\|^2 \mathbf{1}_{E_n}] = O(\mathbb{E}\|Q_n\|^4)^{1/2} O(\mathbb{E}\|S_n\|^4)^{1/2} + O(\mathbb{E}\|S_n\|^4) = O(n^{-2})$ by Cauchy–Schwarz. Hence $\mathbb{E}\|\widehat{\theta}_\lambda - \theta^*_\lambda\|^2 = \frac{1}{n}\operatorname{tr}(F_\lambda^{-1} V_\lambda F_\lambda^{-1}) + O(n^{-3/2})$ by Cauchy–Schwarz on the cross term, and $\mathbb{E}[\widehat{\theta}_\lambda] - \theta^*_\lambda = \mathbb{E}[\Delta_n] = O(1/n)$. □

Lemma A.4 (Population bias)**.** $b_\lambda := A_{\theta^*_\lambda} - A_{\theta^*_{\mathrm{post}}} = -\lambda(F + \lambda H_T^{\mathrm{pre}})^{-1} H_T^{\mathrm{pre}} B + O(\|B\|^2)$ *as* $B \to 0$*, and* $b_0 = 0$*.*

Proof. Let $L(\theta) = \mathbb{E}[\Lambda(A_\theta\psi) - \phi(x)^\top A_\theta\psi]$ and $R_T(\theta) = \mathbb{E}[D_T(A_\theta; A_{\theta^*_{\mathrm{pre}}}, \psi)]$, so $J_{\lambda,T} = L + \lambda R_T + \mathrm{const}$. Since $\mathbb{E}[\phi(x) \mid \psi] = \mu(A_{\theta^*_{\mathrm{post}}}\psi)$, $\nabla L(\theta) = \mathbb{E}[(\mu(A_\theta\psi) - \mu(A_{\theta^*_{\mathrm{post}}}\psi))\psi^\top]$, which vanishes at $A_{\theta^*_{\mathrm{post}}}$ with $\nabla^2 L(A_{\theta^*_{\mathrm{post}}}) = F$; likewise $\nabla R_T(\theta) = \frac{1}{T}\mathbb{E}[(\mu(\frac{A_\theta\psi}{T}) - \mu(\frac{A_{\theta^*_{\mathrm{pre}}}\psi}{T}))\psi^\top]$ vanishes at $A_{\theta^*_{\mathrm{pre}}}$ with $\nabla^2 R_T(A_{\theta^*_{\mathrm{pre}}}) = H_T^{\mathrm{pre}}$. The minimizer solves $\nabla L(\theta^*_\lambda) + \lambda \nabla R_T(\theta^*_\lambda) = 0$. Taylor-expanding ∇L around $A_{\theta^*_{\mathrm{post}}}$ and ∇R_T around $A_{\theta^*_{\mathrm{pre}}}$, with $b_\lambda = A_{\theta^*_\lambda} - A_{\theta^*_{\mathrm{post}}}$ and $A_{\theta^*_\lambda} - A_{\theta^*_{\mathrm{pre}}} = b_\lambda + B$, $F b_\lambda + \lambda H_T^{\mathrm{pre}}(b_\lambda + B) + O(\|b_\lambda\|^2 + \|b_\lambda + B\|^2) = 0$. For $B = 0$ the unique solution is $b_0 = 0$ (then $A_{\theta^*_{\mathrm{post}}} = A_{\theta^*_{\mathrm{pre}}}$ minimizes both terms). For small B, the implicit function theorem applied to the smooth map $(\theta, B) \mapsto \nabla L(\theta) + \lambda \nabla R_T(\theta)$, whose θ-derivative $F + \lambda H_T^{\mathrm{pre}}$ is invertible, gives $b_\lambda = -\lambda(F + \lambda H_T^{\mathrm{pre}})^{-1} H_T^{\mathrm{pre}} B + O(\|B\|^2)$. □

Proof of Theorem 3.2. The expansion. Write $R_n(\lambda, T) = \mathbb{E}\|A_{\widehat{\theta}_{\lambda,T}} - A_{\theta^*_{\mathrm{post}}}\|_F^2$ for the risk. Decompose $A_{\widehat{\theta}_\lambda} - A_{\theta^*_{\mathrm{post}}} = (\widehat{\theta}_\lambda - \theta^*_\lambda) + b_\lambda$, so $R_n(\lambda) = \mathbb{E}\|\widehat{\theta}_\lambda - \theta^*_\lambda\|^2 + \|b_\lambda\|^2 + 2\langle \mathbb{E}[\widehat{\theta}_\lambda] - \theta^*_\lambda, b_\lambda\rangle$. By Lemma A.3 and $\|B\| = O(n^{-1/2})$, the first term is $\frac{1}{n}\operatorname{tr}\left((F + \lambda H_T)^{-1} F (F + \lambda H_T)^{-1}\right) + O(n^{-3/2})$. By Lemma A.4 with $B = B_0/\sqrt{n}$ and $H_T^{\mathrm{pre}} = H_T + O(n^{-1/2})$, the second is $\frac{\lambda^2}{n}\|(F + \lambda H_T)^{-1} H_T B_0\|_F^2 + O(n^{-3/2})$. The cross term is $O(1/n) \cdot O(n^{-1/2})$. Since $B_0 = \sqrt{n}\, B$, this is exactly (9); uniformity in λ on compacts holds because every constant above is continuous in λ. For the rest of the proof ρ denotes the leading term of (9) itself, and $\bar{\rho} = n\rho$ its rescaling, in which

the maximum-likelihood value is $\mathrm{tr}(F^{-1})$ and the teacher gap enters as $B_0 = \sqrt{n}\,B$. The strictness argument below is stated for ρ, the window calculation and the aligned Example for $\bar{\rho}$.

Endpoints. In the rescaled units, $\bar{\rho}(0) = \mathrm{tr}(F^{-1})$ is immediate; unscaled this reads $\rho(0,T) = \mathrm{tr}(F^{-1})/n$. As $\lambda \to \infty$, $(F+\lambda H_T)^{-1}\lambda H_T = I - (F+\lambda H_T)^{-1}F \to I$ because $\|(F+\lambda H_T)^{-1}\|_{\mathrm{op}} \le 1/(\lambda\lambda_{\min}(H_T))$, so the bias term tends to $\|B_0\|_F^2$, i.e. $\lim_{\lambda\to\infty}\rho(\lambda,T) = \|B\|_F^2$, and the variance is at most $\|F\|_{\mathrm{op}}\,\mathrm{tr}\left((F+\lambda H_T)^{-2}\right) \to 0$.

Derivative at zero (Remark A.5). The bias term is $O(\lambda^2)$ and contributes nothing. With $M_\lambda = (F+\lambda H_T)^{-1}$ we have $M_0' = -F^{-1}H_TF^{-1}$, so $\frac{d}{d\lambda}\,\mathrm{tr}(M_\lambda F M_\lambda)\big|_0 = 2\,\mathrm{tr}(M_0' F M_0) = -2\,\mathrm{tr}(F^{-1}H_TF^{-1}FF^{-1}) = -2\,\mathrm{tr}(F^{-2}H_T)$, which is negative because $F^{-2} \succ 0$ and $H_T \succ 0$; equivalently $\partial_\lambda\bar{\rho}(0,T) = -2\,\mathrm{tr}(F^{-2}H_T)$ and $\partial_\lambda\rho(0,T) = -2\,\mathrm{tr}(F^{-2}H_T)/n$.

Strictness. In two steps: the $T=1$ slice already beats the other two extremes, and moving T off 1 beats the $T=1$ slice for generic B.

(i) The $T=1$ slice. Setting $T=1$ in the definition of H_T gives $H_1 = \mathbb{E}[\Sigma(A_{\theta^*_{\mathrm{post}}}\psi)\otimes\psi\psi^\top] = F$. Hence $M = \lambda F$, the shrinkage operator is $(F+M)^{-1}M = \alpha I$ with $\alpha = \lambda/(1+\lambda)$, and (9) becomes, for every F and B,

$$\rho(\lambda,1) = (1-\alpha)^2 V + \alpha^2 W, \qquad V = \frac{\mathrm{tr}(F^{-1})}{n}, \quad W = \|B\|_F^2,$$

a scalar bias–variance tradeoff. It is strictly convex in α and minimized at $\alpha^\circ = V/(V+W) \in (0,1)$, i.e. at the interior point $\lambda_1^* = V/W \in (0,\infty)$, where $\partial_\lambda\rho(\lambda_1^*,1) = 0$. Its value there is the harmonic-type mean $VW/(V+W)$, which is strictly smaller than both V and W whenever $V,W>0$, that is, whenever $B \neq 0$. Since V is the maximum-likelihood risk $\rho(0,T)$ and $W = \lim_{\lambda\to\infty}\rho(\lambda,T)$ is the frozen-teacher risk, the best $T=1$ run already beats those two strictly. No further assumption is needed for this step.

(ii) Off $T=1$. Regard ρ as a function of the penalty $M \succeq 0$ alone. For a symmetric direction Δ, differentiating $(F+M)^{-1}M = I - (F+M)^{-1}F$ and $(F+M)^{-1}F(F+M)^{-1}$ at $M = \lambda F$, where $(F+M)^{-1} = F^{-1}/(1+\lambda)$, gives

$$\left\langle \nabla_M \rho(\lambda F; B), \Delta \right\rangle = \frac{2}{(1+\lambda)^3}\left[\lambda\,\langle B, F^{-1}\Delta B\rangle - \frac{1}{n}\,\mathrm{tr}(F^{-2}\Delta)\right]. \tag{14}$$

Taking $\Delta = F$ returns $2(\lambda W - V)/(1+\lambda)^3$, which vanishes exactly at $\lambda_1^* = V/W$, as it must. Take instead $\Delta = D$. Since $\partial_\lambda M = H_1 = F$ and $\partial_T M|_{T=1} = \lambda_1^* D$, and since $\partial_\lambda\rho(\lambda_1^*,1) = 0$, we get $\partial_T\rho(\lambda_1^*,1) = \lambda_1^*\Pi(B)$ with $\Pi(B) = \langle \nabla_M\rho(\lambda_1^*(B)F; B), D\rangle$. Writing $b = \mathrm{vec}(B)$, $r = b^\top b = \|B\|_F^2$ and $s = \mathrm{tr}(F^{-1})$, so that $\lambda_1^* = s/(nr)$, equation (14) collects into a quadratic form in b:

$$\Pi(B) = \frac{2\,b^\top C b}{n\,r\,(1+\lambda_1^*)^3}, \qquad C = s\cdot\mathrm{sym}(F^{-1}D) - \mathrm{tr}(F^{-2}D)\,I, \tag{15}$$

using $b^\top F^{-1}Db = b^\top\,\mathrm{sym}(F^{-1}D)b$. The matrix C depends on F and D only, and $C=0$ if and only if D is a multiple of F: in an eigenbasis of F with eigenvalues μ_i one has $(\mathrm{sym}(F^{-1}D))_{ij} = \frac{1}{2}D_{ij}(\mu_i^{-1}+\mu_j^{-1})$, so $C=0$ forces $D_{ij}=0$ for $i\neq j$ and $D_{ii} = \mu_i\,\mathrm{tr}(F^{-2}D)/s$, i.e. $D = cF$; conversely $D = cF$ gives $\mathrm{sym}(F^{-1}D) = cI$ and $\mathrm{tr}(F^{-2}D) = cs$, hence $C=0$. Thus (A2) says precisely $C \neq 0$, and then $Z = \{B : b^\top C b = 0\}$ is the zero set of a nonzero quadratic form: closed, of Lebesgue measure zero, and a cone, so membership of Z is unaffected by the $n^{-1/2}$ scaling of B. For $B \notin Z$ we have $\partial_T\rho(\lambda_1^*,1) \neq 0$, so moving T off 1 in the descent direction strictly decreases ρ below $\min_\lambda\rho(\lambda,1)$, and by step (i) below all three extremes. The descent direction is $T>1$ when $b^\top Cb < 0$ and $T<1$ when $b^\top Cb > 0$; the theorem claims improvement, not that it is always on the hot side. Note that this argument needs no commutativity between F, H_T and D, and no analyticity: $b^\top Cb$ is a polynomial.

The weak statement (10). It is pure continuity, and needs neither $B\neq 0$ nor (A2). $\rho(\cdot,T)$ is continuous on $[0,\infty)$ and has a finite limit at ∞, so $\inf_{\lambda\in(0,\infty)}\rho(\lambda,T) \le \min\{\rho(0,T), \lim_{\lambda\to\infty}\rho(\lambda,T)\}$, both extremes being limits of interior values; the same argument in T gives $\inf_{T\neq 1}\rho(\lambda,T) \le \rho(\lambda,1)$ and $\inf_{T<\infty}\rho(\lambda,T) \le \lim_{T\to\infty}\rho(\lambda,T)$. Taking the smallest of the three right-hand sides gives (10); the assumptions are what turn it into a strict inequality.

The window λ_0 (Remark A.5). Write $F+\lambda H_T = F^{1/2}(I+\lambda K)F^{1/2}$ with $K = F^{-1/2}H_T F^{-1/2} \succ 0$ and eigenpairs (k_i, u_i). Then $M_\lambda F M_\lambda = F^{-1/2}(I+\lambda K)^{-2}F^{-1/2}$, so the variance equals $\sum_i (1+\lambda k_i)^{-2}\, u_i^\top F^{-1} u_i$. Since $(1-2x+3x^2)(1+x)^2 - 1 = 4x^3+3x^4 \geq 0$, we have $(1+x)^{-2} \leq 1-2x+3x^2$ for $x \geq 0$, whence the variance is at most $\operatorname{tr}(F^{-1}) - 2\lambda \operatorname{tr}(KF^{-1}) + 3\lambda^2 \operatorname{tr}(K^2F^{-1}) = \operatorname{tr}(F^{-1}) - 2\lambda \operatorname{tr}(H_T F^{-2}) + 3\lambda^2 \operatorname{tr}(H_T F^{-1} H_T F^{-2})$. The bias term is at most $\lambda^2 \|M_\lambda\|_{\mathrm{op}}^2 \|H_T B_0\|_F^2 \leq \lambda^2 \|H_T B_0\|_F^2 / \lambda_{\min}(F)^2$. Hence $\bar\rho(\lambda) - \bar\rho(0) \leq -2\lambda \operatorname{tr}(H_T F^{-2}) + \lambda^2\big(\|H_T B_0\|_F^2/\lambda_{\min}(F)^2 + 3\operatorname{tr}(H_T F^{-1} H_T F^{-2})\big)$, which is negative for $0 < \lambda < \lambda_0$. Finally, for fixed $\lambda \in (0, \lambda_0)$, $n\big(R_n(\lambda) - R_n(0)\big) = \bar\rho(\lambda) - \bar\rho(0) + O(n^{-1/2}) < 0$ once n is large enough. □

Remark A.5 (How much teacher, explicitly). The improvement over maximum likelihood comes from $\partial_\lambda \rho(0,T) = -2\operatorname{tr}(F^{-2}H_T)/n < 0$, the bias entering only at second order in λ; at the other end, $\rho(\lambda, T) \to \|B\|_F^2$ as $\lambda \to \infty$, pure bias. Carrying both terms explicitly, every λ below

$$\lambda_0 = \frac{2\operatorname{tr}(H_T F^{-2})}{n\|H_T B\|_F^2/\lambda_{\min}(F)^2 + 3\operatorname{tr}(H_T F^{-1} H_T F^{-2})},$$

improves on maximum likelihood; this window is of order $\operatorname{tr}(F^{-2}H_T)/(n\|B\|^2)$ when the teacher gap dominates.

Remark A.6 ($T = 1$ can be optimal). The exceptional set Z is not empty. Take $V = \{0,1\}$ with scalar embeddings $\phi(0) = 0$, $\phi(1) = 1$, two equally likely context features e_1, e_2, and target logits 0 and 2, so F and every H_T are diagonal, with $\mu_1 = \frac{1}{8}$ and $\mu_2 \approx 0.0525$. Choose the teacher gap with $b_j^2 = 1/(n\mu_j)$, which is of the order $n^{-1/2}$ the theorem assumes. Each coordinate then contributes $\big((1-\alpha_j)^2 + \alpha_j^2\big)/(n\mu_j)$ to ρ, minimized at $\alpha_j = \frac{1}{2}$; since $H_1 = F$, the single choice $\lambda = 1$ attains that in both coordinates at once, so $\lambda H_1 = F$ minimizes ρ over all *diagonal* penalties, and in particular over the family $\{\lambda H_T\}$, whose every member is diagonal here. Moreover $\kappa_1(T)/\kappa_2(T)$ is strictly decreasing in T, so $\lambda H_T = F$ has the unique solution $(\lambda, T) = (1,1)$: knowledge distillation is not merely tied here but is the unique optimizer. (A2) holds, the diagonal ratios of D to F being -2 and ≈ -0.477; what fails is $B \notin Z$. Indeed $n\mu_j b_j^2$ is constant across directions, which makes $\lambda_1^* = V/W = 1$ and, by (15), $b^\top C b = \frac{1}{n}\big[(\sum_k \mu_k^{-1})(\sum_j d_j \mu_j^{-2}) - (\sum_j \mu_j^{-1})(\sum_k d_k \mu_k^{-2})\big] = 0$, so $\Pi(B) = 0$ and $T = 1$ is stationary. Note that $\Pi(B) = 0$ is weaker than $\nabla_M \rho(\lambda_1^* F) = 0$, which fails here: by (14) the gradient has the off-diagonal entries of $\operatorname{sym}(F^{-1} b b^\top)$, which are nonzero, so F is optimal within the reachable family but not among all positive-definite penalties. The coincidence $b^\top C b = 0$ has measure zero in B, which is what the genericity caveat excludes.

Remark A.7 (The infimum need not be attained). Theorem 3.2 asserts a strict inequality, not the existence of a best (λ, T), and the distinction is real. The risk depends on the pair only through $M = \lambda H_T$, so λ sets the scale of the penalty and T its shape, and the reachable shapes trace the curve $T \mapsto T^2 H_T$, which starts at F at $T = 1$ and tends to $\Sigma_0 \otimes G$ as $T \to \infty$. Convergence by itself decides nothing about attainment: if the target conditionals are uniform then $T^2 H_T = \Sigma_0 \otimes G$ at every temperature and the curve is a single point. But the limiting shape need not be reachable at any finite T, and then the joint infimum can fail to be attained even under the non-degeneracy condition. The following example does that, on an open set of teacher gaps. Keep the model of Remark A.6 but take $b_1^2 = 2/n$ and $b_2^2 = 1/n$. By the Example below the best penalty eigenvalues are $m_j^* = 1/(nb_j^2)$, that is $(\frac{1}{2}, 1)$, whereas the reachable ratio is $m_1/m_2 = \kappa_1(T)/\kappa_2(T) = \cosh^2(1/T)$, which exceeds 1 at every finite T and tends to 1 only as $T \to \infty$. With λ free to set the scale, the reachable set is therefore exactly the open cone $\{m_1 > m_2 > 0\}$, and the unconstrained optimum $(\frac{1}{2}, 1)$ lies outside it. The optimum over the closure is a different point: each coordinate contributes $(m_j^2 b_j^2 + \mu_j/n)/(\mu_j + m_j)^2$, strictly decreasing then strictly increasing with minimum at m_j^*, so from any feasible $m_1 > m_2$ one can move to the diagonal without increasing either contribution, and the constrained minimum sits on the face $m_1 = m_2$. That face is disjoint from the reachable set, so the infimum over finite (λ, T) is approached only as $T \to \infty$ with $\lambda \propto T^2$ and is never attained. Note that this is a different limit from the maximum-likelihood one in (10), which sends $T \to \infty$ at fixed λ and does kill the penalty. Read positively, a missing minimizer says the model would like the hottest teacher available, with λ raised to compensate for the flattening.

Example (aligned case). Suppose F and H_T share eigenvectors, with eigenvalues (μ_j) and (κ_j), and let b_j be the coordinates of B_0 in that basis. Then

$$\bar\rho(\lambda) = \sum_j \Big[(1-\alpha_j)^2 \frac{1}{\mu_j} + \alpha_j^2 b_j^2\Big], \qquad \alpha_j = \frac{\lambda\kappa_j}{\mu_j + \lambda\kappa_j} \in (0,1). \tag{16}$$

Diagonalizing both terms of $\bar{\rho}$ in the common eigenbasis, direction j contributes $\big(\frac{\lambda\kappa_j}{\mu_j+\lambda\kappa_j}\big)^2 b_j^2 + \frac{\mu_j}{(\mu_j+\lambda\kappa_j)^2} = \alpha_j^2 b_j^2 + (1-\alpha_j)^2 \frac{1}{\mu_j}$, using $1-\alpha_j = \mu_j/(\mu_j+\lambda\kappa_j)$. Comparing with the maximum-likelihood value $1/\mu_j$: the shrunk direction is better iff $\alpha_j^2 b_j^2 < \big(1-(1-\alpha_j)^2\big)\frac{1}{\mu_j} = \alpha_j(2-\alpha_j)\frac{1}{\mu_j}$, i.e. iff $b_j^2 < \frac{2-\alpha_j}{\alpha_j}\frac{1}{\mu_j}$. Minimizing $(1-\alpha)^2 v + \alpha^2 b^2$ over α with $v = 1/\mu_j$ gives $\alpha^* = v/(v+b^2) = 1/(1+\mu_j b_j^2)$. In the original units ($b_j = \sqrt{n}\, b'_j$ with b'_j the coordinates of B) these read $b_j'^2 < \frac{2-\alpha_j}{\alpha_j}\frac{1}{n\mu_j}$ and $\alpha_j^* = 1/(1+n\mu_j b_j'^2)$: the amount of teacher to trust in a direction is set by the ratio of the teacher's squared error to the sampling variance $1/(n\mu_j)$ there. Differentiating at $\lambda = 0$ recovers $\bar{\rho}'(0) = -2\sum_j \kappa_j/\mu_j^2$. Finally, since the shrinkage in direction j depends on (λ, T) only through the eigenvalue $m_j = \lambda\kappa_j$ of $M = \lambda H_T$, through $\alpha_j = m_j/(\mu_j+m_j)$, the optimal α_j^* corresponds to $m_j^* = \mu_j\alpha_j^*/(1-\alpha_j^*) = 1/(nb_j'^2)$: the penalty should be strongest in the directions where the teacher's error b'_j is smallest, a shape that depends on the teacher and not on the data, and one that $M = \lambda F$ (i.e. $T = 1$) cannot produce unless $n\mu_j b_j'^2$ is constant across directions (Theorem 3.2).

Lemma A.8 (Curvature of a softmax is at least its smallest probability)**.** *For every $\eta \in \mathbb{R}^d$, $\Sigma(\eta) \succeq e^{-\operatorname{range}(\Phi\eta)}\,\Sigma_0$, where $\operatorname{range}(u) = \max_z u_z - \min_z u_z$ and $\Sigma_0 = \operatorname{Cov}_{\mathrm{unif}(V)}[\phi(y)] = \frac{1}{|V|}\Phi^\top(I - \frac{1}{|V|}\mathbf{1}\mathbf{1}^\top)\Phi$.*

Proof. Let $p = \sigma(\Phi\eta)$ and $u \in \mathbb{R}^{|V|}$. Then $u^\top(\operatorname{diag}(p) - pp^\top)u = \operatorname{Var}_p(u) = \sum_z p_z(u_z - \bar{u}_p)^2 \geq p_{\min}\sum_z(u_z-\bar{u}_p)^2 \geq p_{\min}\sum_z(u_z-\bar{u})^2 = p_{\min}\, u^\top(I - \frac{1}{|V|}\mathbf{1}\mathbf{1}^\top)u$, where $\bar{u}_p$ is the p-weighted mean, $\bar{u}$ the unweighted mean, and the last inequality holds because the unweighted mean minimizes $c \mapsto \sum_z(u_z-c)^2$. Hence $\Sigma(\eta) = \Phi^\top(\operatorname{diag}(p) - pp^\top)\Phi \succeq p_{\min}|V|\,\Sigma_0$, and $p_{\min} = \min_z e^{u_z}/\sum_w e^{u_w} \geq e^{-\operatorname{range}(u)}/|V|$ with $u = \Phi\eta$. □

Dependence within documents. Theorem 3.2 is stated for independent prefixes for clarity, but nothing is lost when the $n = n_1 l$ tokens come from n_1 independent documents of fixed length l, which is the situation of (4). Group the score by document, $Z_k = \frac{1}{l}\sum_{i=1}^{l}\nabla\ell_{k,i}(\theta_\lambda^*)$. The Z_k are i.i.d. across k, so Lemma A.3 applies with $n_1 \to \infty$ at fixed l, its empirical-curvature concentration now run on the document-level averages, and with constants that may depend on l. What has to be rechecked is the covariance. Per document it carries an extra factor of l, so define the token-normalized quantity $V_\lambda^{\mathrm{tok}} := l\operatorname{Cov}(Z_k)$; the claim is $V_\lambda^{\mathrm{tok}} = F + O(\|B\|)$, which is what the sandwich needs because $\frac{1}{n_1}\operatorname{Cov}(Z_k) = \frac{1}{n_1 l}V_\lambda^{\mathrm{tok}} = \frac{1}{n}V_\lambda^{\mathrm{tok}}$. Split the per-token score at θ_{post}^* into a clean part $(\mu(A_{\theta_{\mathrm{post}}^*}\psi_i) - \phi(x_i))\psi_i^\top$ and a teacher part $\frac{\lambda}{T}\big(\mu(\frac{A_{\theta_{\mathrm{post}}^*}\psi_i}{T}) - \mu(\frac{A_{\theta_{\mathrm{pre}}^*}\psi_i}{T})\big)\psi_i^\top$. Only the clean part is a martingale difference for the prefix filtration: conditionally on $\mathbf{x}_{<i}$ the token x_i has mean embedding $\mu(A_{\theta_{\mathrm{post}}^*}\psi_i)$, so clean parts at different positions of the same document are uncorrelated and their covariances add, contributing $\frac{1}{l^2}\sum_i \mathbb{E}[\Sigma(A_{\theta_{\mathrm{post}}^*}\psi_i)\otimes\psi_i\psi_i^\top] = \frac{F}{l}$ to $\operatorname{Cov}(Z_k)$, hence F to V_λ^{tok}, with F now read as an average over positions. The teacher part is *not* a martingale difference: it is a deterministic function of ψ_i whose conditional mean is generally nonzero, and its within-document correlations do not cancel. They are harmless only because the whole term is $O(\|B\|)$, so it adds $O(\|B\|^2)$ to $\operatorname{Cov}(Z_k)$, and the cross term with the clean part adds $O(\|B\|)$ there, only pairs $j > i$ surviving since $r_T(\psi_j)$ for $j \leq i$ is measurable with respect to the prefix at i. In V_λ^{tok} these become $O(l\|B\|^2)$ and $O(l\|B\|)$, both absorbed into $V_\lambda^{\mathrm{tok}} = F + O(\|B\|)$ at fixed l. We do not claim a version with l growing with n_1: there the within-document correlations of the teacher term enter a long-run variance, and a martingale central limit theorem would in any case not supply by itself the second-moment remainder that (9) uses. No repetition of contexts is assumed anywhere.

B Training Algorithm

Algorithm 1 is the stochastic implementation of the objective (4) of Section 3.1. Each step draws one rank-synchronized coin: with probability p_c the batch takes a standard SFT step, and otherwise the frozen teacher is run forward without gradients and the step minimizes the matching loss at every supervised position. In expectation this is (4) with $\lambda = (1-p_c)/p_c$, and $p_c = 1$ recovers SFT exactly.

Algorithm 1 Ambient SFT (drop-in replacement for SFT)

Require: pre-trained weights θ^*_{pre}; post-training documents $\mathcal{D}_p$; clean probability $p_c \in (0, 1]$; temperature $T > 0$; number of steps S; optimizer OPT

$\theta \leftarrow \theta^*_{\text{pre}}$ (student); keep a frozen copy of θ^*_{pre} (teacher)
for $s = 1, \dots, S$ **do**
  sample a batch $\mathcal{B} \subset \mathcal{D}_p$; let $\mathcal{I}$ be its supervised positions (k, i) and $N = |\mathcal{I}|$
  $g_{(k,i)} \leftarrow g_\theta(\cdot \mid \mathbf{x}^{(k)}_{<i})$ for $(k, i) \in \mathcal{I}$ ▷ student logits, one forward pass
  draw $c \sim \text{Bernoulli}(p_c)$ ▷ shared across data-parallel ranks
  if $c = 1$ **then** ▷ clean SFT step
    $\mathcal{L} \leftarrow -\frac{1}{N} \sum_{(k,i)\in\mathcal{I}} \log \text{softmax}(g_{(k,i)}(x^{(k)}_i))$
  else ▷ temperature-matching step
    $\tilde{g}_{(k,i)} \leftarrow g_{\theta^*_{\text{pre}}}(\cdot \mid \mathbf{x}^{(k)}_{<i})$ for $(k, i) \in \mathcal{I}$, without gradients ▷ teacher logits
    $q_{(k,i)} \leftarrow \text{softmax}(\tilde{g}_{(k,i)}/T)$ ▷ tempered teacher
    $\mathcal{L} \leftarrow -\frac{1}{N} \sum_{(k,i)\in\mathcal{I}} \sum_{y\in V} q_{(k,i)}(y) \log \text{softmax}(g_{(k,i)}(y)/T)$ ▷ CE to tempered teacher
  end if
  $\theta \leftarrow \text{OPT}(\theta, \nabla_\theta \mathcal{L})$
end for
return θ

C Experimental Details

C.1 Domain Adaptation

C.1.1 Datasets

All models are pre-trained on OpenWebText (≈9B tokens), a publicly available reproduction of the WebText corpus used to train GPT-2, obtained from the `Skylion007/openwebtext`[1] Hugging Face dataset. For the domain adaptation experiments of Section 4.2, we adapt these pre-trained models to five target domains, chosen to span a diverse range of text types. Each domain corpus is capped at ≈225M target-domain tokens, matching the same order of magnitude across domains so that no single domain benefits from a systematically larger fine-tuning set. The domains, their sources, and the exact Hugging Face datasets used are:

- **Python Code**: an ≈ 225M token subset of the CodeParrot clean-train dataset (CodeParrot Team, 2022), itself built from permissively-licensed GitHub Python repositories in the style of the training corpus used by Chen et al. (2021). `codeparrot/codeparrot-clean-train`[2].
- **CS**: computer science paper abstracts drawn from S2ORC, the Semantic Scholar Open Research Corpus (Lo et al., 2020). `AlgorithmicResearchGroup/s2orc-cs-enriched`[3].
- **BioMed**: biomedical abstracts drawn from PubMed metadata (National Library of Medicine), via the subset released alongside Stuhlmann et al. (2025). `slinusc/PubMedAbstractsSubset`[4].
- **News**: English-language news articles from CC-News, a Common Crawl news dataset extracted with the `news-please` crawler (Hamborg et al., 2017). `vblagoje/cc_news`[5].
- **Reviews**: Amazon product reviews (books category), drawn from the Amazon Reviews 2023 dataset (Hou et al., 2026). `cogsci13/Amazon-Reviews-2023-Books-Review`[6].

[1] https://huggingface.co/datasets/Skylion007/openwebtext
[2] https://huggingface.co/datasets/codeparrot/codeparrot-clean-train
[3] https://huggingface.co/datasets/AlgorithmicResearchGroup/s2orc-cs-enriched
[4] https://huggingface.co/datasets/slinusc/PubMedAbstractsSubset
[5] https://huggingface.co/datasets/vblagoje/cc_news
[6] huggingface.co/datasets/cogsci13/Amazon-Reviews-2023-Books-Review

C.1.2 FULL (p, T) SWEEP AND OPTIMAL HYPERPARAMETERS

Table 8 reports the optimal (p, T) per domain used to produce Table 15 and Table 16's 124M-student rows. The optimal mixing ratio p was 0.5 in every one of the 10 cells tested, so the table reports only T.

For the Small Teacher condition specifically, Tables 3–7 report the full (p, T) grid each domain's optimum was selected from: a uniform 15-point grid ($p \in \{0.5, 0.75, 0.9\} \times T \in \{1.25, 1.5, 1.75, 2.0, 2.5\}$) for CS, BioMed, News, and Reviews, and a wider grid for Python Code.

Table 3: Full (p, T) sweep for the Ambient SFT (Small Teacher) condition, CS, 124M student (15 configurations: $p \in \{0.5, 0.75, 0.9\} \times T \in \{1.25, 1.5, 1.75, 2.0, 2.5\}$). Validation loss, rounded to 2 decimal places; the domain's best cell in **bold** (ties within rounding are all bolded).

$p \backslash T$	1.25	1.5	1.75	2	2.5
0.5	**2.73**	**2.73**	**2.73**	2.75	2.75
0.75	2.75	2.74	2.75	2.76	2.76
0.9	2.77	2.77	2.77	2.78	2.77

Table 4: Full (p, T) sweep for the Ambient SFT (Small Teacher) condition, BioMed, 124M student (15 configurations: $p \in \{0.5, 0.75, 0.9\} \times T \in \{1.25, 1.5, 1.75, 2.0, 2.5\}$). Validation loss, rounded to 2 decimal places; the domain's best cell in **bold** (ties within rounding are all bolded).

$p \backslash T$	1.25	1.5	1.75	2	2.5
0.5	2.50	2.49	2.49	**2.47**	2.49
0.75	2.51	2.50	2.50	2.50	2.49
0.9	2.51	2.50	2.51	2.50	2.50

Table 5: Full (p, T) sweep for the Ambient SFT (Small Teacher) condition, News, 124M student (15 configurations: $p \in \{0.5, 0.75, 0.9\} \times T \in \{1.25, 1.5, 1.75, 2.0, 2.5\}$). Validation loss, rounded to 2 decimal places; the domain's best cell in **bold** (ties within rounding are all bolded).

$p \backslash T$	1.25	1.5	1.75	2	2.5
0.5	2.25	2.24	**2.23**	2.25	2.26
0.75	2.26	2.28	2.29	2.28	2.30
0.9	2.33	2.33	2.33	2.33	2.33

Table 6: Full (p, T) sweep for the Ambient SFT (Small Teacher) condition, Reviews, 124M student (15 configurations: $p \in \{0.5, 0.75, 0.9\} \times T \in \{1.25, 1.5, 1.75, 2.0, 2.5\}$). Validation loss, rounded to 2 decimal places; the domain's best cell in **bold** (ties within rounding are all bolded).

$p \backslash T$	1.25	1.5	1.75	2	2.5
0.5	3.16	**3.15**	**3.15**	**3.15**	3.16
0.75	3.16	3.16	3.16	3.17	3.18
0.9	3.18	3.18	3.18	3.18	3.18

Table 7: Full (p, T) sweep for the Ambient SFT (Small Teacher) condition, Python Code, 124M student. Validation loss, rounded to 2 decimal places; the domain's best cell in **bold** (ties within rounding are all bolded). The high-temperature extension $T \in \{2.25, 2.5, 3\}$ was run only for the two competitive mixing ratios, so $p \in \{0.25, 0.9\}$, which are uniformly worse at every T that was run (1.08 and 0.88 at best), are omitted.

$p \backslash T$	1	1.25	1.5	1.75	2	2.25	2.5	3
0.5	1.02	0.94	0.91	0.88	0.87	**0.86**	**0.86**	0.87
0.75	0.93	0.91	0.89	0.88	0.88	0.87	0.88	0.88

Table 8: Optimal temperature T per domain for the 124M student, both teacher conditions. The optimal mixing ratio was $p = 0.5$ in every one of the 10 (domain $\times$ teacher) cells tested – omitted from the table since it does not vary.

Domain	Small Teacher: best T	Large Teacher: best T
Python Code	2.5	2.5
CS	1.5	1.0
BioMed	2.0	2.0
News	1.75	1.5
Reviews	1.75	1.0

Table 9 reports the analogous optimal (p, T) for the 1.5B student, Ambient SFT's self-teacher condition (Table 1 and Table 16's "1.5B & 1.5B (self)" row) – the only 1.5B-student condition that received a real per-domain (p, T) search in this work. Unlike the 124M student, the optimal p is not uniform across domains, so both p and T are reported. We do not report an analogous table for the 1.5B student paired with the 124M teacher (Table 16's "1.5B & 124M" row): that condition reuses the (p, T) found optimal at the 124M-student scale rather than being independently retuned, so there is no separate search to report.

Table 9: Optimal (p, T) per domain for the 1.5B student, Ambient SFT self-teacher condition (1.5B teacher).

Domain	Best p	Best T
Python Code	0.5	1.5
CS	0.75	2.0
BioMed	0.5	1.0
News	0.5	1.5
Reviews	0.9	2.5

C.2 Instruction Tuning

C.2.1 Training hyperparameters

Every instruction-tuning run in this paper uses the same training recipe, listed in Table 10; vanilla SFT and Ambient SFT differ only in the loss. Runs are trained with the OLMo-core trainer, starting from each model's released base checkpoint. The recipe follows Ai2's instruction-tuning recipe for OLMo 3 (learning rate, schedule, batch size and sequence length) and is kept fixed across base models and model sizes; we did not tune it per model. The only settings that change between experiments are the number of epochs (Table 11) and the tokenizer, which is each base model's own. Every run uses 16 GH200 GPUs; the global batch is fixed in tokens, so the gradient-accumulation factor absorbs the context-parallel degree.

Table 10: Training hyperparameters shared by all instruction-tuning runs.

Setting	Value
Optimization (vanilla SFT and Ambient SFT)	
Optimizer	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$, weight decay 0
Peak learning rate	8×10^{-5}
Schedule	linear warmup over the first 3% of steps, then linear decay to 0
Gradient clipping	global norm 1.0
Global batch	1,048,576 tokens (32 sequences of 32,768 tokens)
Sequence length	32,768 tokens; documents packed, longer documents truncated
Loss tokens	assistant turns only (all other tokens are masked out of the loss)
Precision	bfloat16 parameters, float32 gradient reduction
Parallelism	FSDP, context parallelism of degree 2 or 4, full activation checkpointing
Seeds	data order 34521
Ambient SFT only	
Clean-loss probability p_c	0.75 (the loss is chosen per micro-batch)
Temperature T	2.5
Teacher	the frozen base checkpoint the student starts from
Matching loss	cross-entropy to the teacher's temperature-T distribution over the full vocabulary on the same loss tokens

Table 11: Number of epochs and optimizer steps per experiment. Steps follow from the data size and the fixed global batch.

Experiment	Data	Epochs	Steps
Tables 2 and 2	1% (21,503 examples)	10	170
Figure 4	10% (215,027 examples)	2, 4, 6, 8, 10	340 to 1,700
Figure 7, 1%	1%	10 (both methods)	170
Figure 7, 10%	10%	2 (vanilla SFT), 4 (Ambient SFT)	340, 680
Figure 7, 100%	100% (2,150,273 examples)	2 (both methods)	3,376

C.2.2 Evaluation harness

All instruction-tuning results (Tables 2 and 2, and Figure 4) come from a single, fixed evaluation setup, applied identically to every checkpoint. We use the `olmes` evaluation suite (Gu et al., 2024) (`github.com/allenai/olmes`, commit `5a51f502`, unmodified) with the vLLM backend (vLLM 0.13.0, PyTorch 2.9.0, Transformers 4.57.6, Python 3.11). Trained checkpoints are converted to the Hugging Face format and evaluated with the chat template stored alongside them: the OLMo 3 instruct template (`<|im_start|>`/`<|im_end|>` turns with its default system prompt), which is also the template used to format the training data, for all three base models. Every checkpoint is evaluated once per task, with no selection over seeds or decoding settings.

Table 12 lists the eight task specifications. "Chat" tasks format each prompt (including the few-shot examples, as earlier turns) with the chat template; "plain" tasks feed a raw text prompt. The reported score is the task's primary metric; for MMLU, MATH and BBH it is the unweighted mean over their subjects or subtasks. The relative change reported in the last column of the instruction-tuning tables is the mean over the eight tasks of $(\text{score} - \text{score}_{\text{SFT}})/\text{score}_{\text{SFT}}$.

Table 12: Task specifications used for every instruction-tuning evaluation. All generation is greedy except HumanEval+, which draws 20 samples per problem to estimate pass@10. MATH uses the Minerva chain-of-thought prompt format (Lewkowycz et al., 2022).

Benchmark	`olmes` task	Shots	Prompt	Decoding	Metric
MATH	`minerva_math::tulu`	4, CoT	chat	greedy, $\leq$1024 tokens	exact match, mean of 7 subjects
GSM8K	`gsm8k::olmes`	8, CoT	plain	greedy, $\leq$512 tokens	exact match
HumanEval+	`codex_humanevalplus::tulu`	0	chat	20 samples, T=0.8, top-p 0.95	pass@10
MMLU	`mmlu::olmes`	5	plain	log-likelihood over options	accuracy, mean of 57 subjects
PopQA	`popqa::tulu`	15	chat	greedy, $\leq$15 tokens	exact match
TriviaQA	`triviaqa::olmes`	5	plain	greedy, $\leq$50 tokens	F1 (closed-book)
IFEval	`ifeval::tulu`	0	chat	greedy, $\leq$2048 tokens	prompt-level loose accuracy
BBH	`bbh:cot::olmes`	3, CoT	plain	greedy, $\leq$512 tokens	exact match, mean of 27 tasks

Commands. The environment and the evaluation of one checkpoint `$CKPT` (a Hugging Face model directory) are:

```
git clone https://github.com/allenai/olmes.git && cd olmes
git checkout 5a51f502d463b8cdc4a2dcad7d7096c41ff1197e
pip install -e . && pip install vllm==0.13.0

export VLLM_WORKER_MULTIPROC_METHOD=spawn
MARGS="gpu_memory_utilization=0.85,max_length=32768"
MARGS="$MARGS,trust_remote_code=false,add_bos_token=false"
TASKS="ifeval::tulu mmlu::olmes gsm8k::olmes codex_humanevalplus::tulu"
TASKS="$TASKS minerva_math::tulu popqa::tulu"
TASKS="$TASKS triviaqa::olmes bbh:cot::olmes"
for TASK in $TASKS; do
  olmes --model-type vllm --model "$CKPT" --model-args "$MARGS" \
        --task "$TASK" --output-dir "$OUT/$(echo $TASK | tr ':/' '__')"
done
```

For Llama-3.1-8B we set `add_bos_token=true`: `olmes` encodes prompts without special tokens, and a Llama model trained with a beginning-of-sequence token on every document produces degenerate output without it.

Sharding. To finish each evaluation in minutes rather than hours, we split every task's instances into 16 disjoint shards and run the command above on each shard in parallel, one GPU per shard. Per-subtask scores are combined across shards weighted by instance count, and the subject and subtask means are taken afterwards, which reproduces the aggregation of an unsharded run.

C.3 Reasoning

C.3.1 Evaluation harness

Reasoning ("Think") checkpoints are evaluated with the same software as Appendix C.2.2 (`olmes` commit `5a51f502`, vLLM 0.13.0) but with the `olmo3:adapt` task specifications, the ones behind Ai2's OLMo 3 Think model card. Every task is zero-shot and chat-formatted with the OLMo 3 Think chat template, whose generation prompt ends with `<|im_start|>assistant\n<think>`, so every response starts inside a reasoning block and the model must close it with `</think>` before its final answer. All generation is sampled with temperature 0.6 and top-p 0.95. The model context is capped at 32,768 tokens (prompt plus response); AIME additionally caps the response at 16,384 tokens. Table 13 lists the six tasks.

Table 13: Task specifications used for every reasoning evaluation. All tasks are zero-shot, chat-formatted and sampled at T=0.6, top-p 0.95.

Benchmark	`olmes` task	Samples per problem	Response budget	Metric
GSM8K	`gsm8k::olmo3:adapt`	1	context limit	exact match of the `\boxed{}` answer
AIME 2024	`aime:2024::olmo3:adapt`	32	16,384 tokens	pass@1, mean over the 32 samples
AIME 2025	`aime:2025::olmo3:adapt`	32	16,384 tokens	pass@1, mean over the 32 samples
HumanEval+	`codex_humanevalplus::olmo3:adapt`	10	context limit	pass@1
IFEval	`ifeval::olmo3:adapt`	1	context limit	prompt-level loose accuracy
GPQA	`gpqa::olmo3:adapt`	1	context limit	exact match of the answer letter (448 questions)

Commands. With the environment of Appendix C.2.2, one Think checkpoint `$CKPT` is evaluated with:

```
export VLLM_WORKER_MULTIPROC_METHOD=spawn
MARGS="gpu_memory_utilization=0.85,max_length=32768"
MARGS="$MARGS,trust_remote_code=false,add_bos_token=false"
TASKS="gsm8k::olmo3:adapt aime:2024::olmo3:adapt aime:2025::olmo3:adapt"
TASKS="$TASKS ifeval::olmo3:adapt codex_humanevalplus::olmo3:adapt"
TASKS="$TASKS gpqa::olmo3:adapt"
for TASK in $TASKS; do
  olmes --model-type vllm --model "$CKPT" --model-args "$MARGS" \
        --task "$TASK" --output-dir "$OUT/$(echo $TASK | tr ':/' '__')"
done
```

The same 16-way sharding and instance-weighted aggregation as in Appendix C.2.2 are used. Because generation is sampled, repeated evaluations of the same checkpoint differ slightly; AIME's 32 samples per problem keep the sampling standard error of its score at about one point (0.8 to 1.2 points for the checkpoints in this paper).

D Additional Results

D.1 Full temperature sweep on Tiny Shakespeare

Figure 2b shows the $T \geq 1$ half of the sweep, the regime the method operates in. Figure 6 adds the low-temperature half, which is symmetric: sampling far below $T = 1$ is also a corruption, producing repetitive rather than random text, and the objective of Eq. (5) again tracks clean-data training over a much wider range of T than standard cross-entropy does.

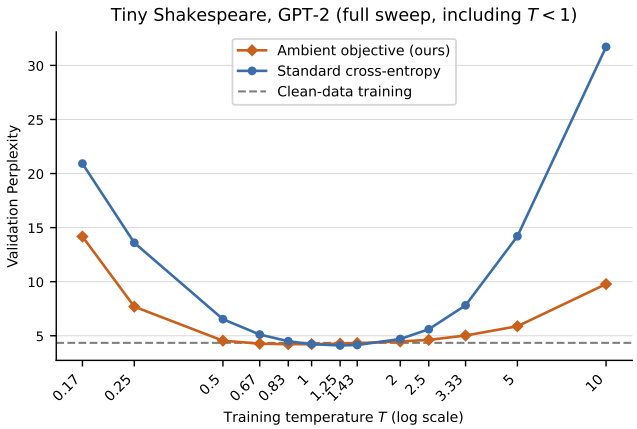


Figure 6: The sweep of Figure 2b over the full range of training temperatures, $T = 1/6$ to $T = 10$. Both losses degrade at both extremes; the objective of Eq. (5) degrades far more slowly.

D.2 Domain Adaptation

D.2.1 Predicting the success of Ambient SFT

Table 14 reports the zero-shot validation loss of the pre-trained (1.5B) model on each target domain, before any fine-tuning, alongside Ambient SFT's win margin over the best baseline from Table 1, as discussed in Section 4.2.

Table 14: Zero-shot validation loss of the pre-trained (1.5B) model on each target domain, before any fine-tuning, alongside Ambient SFT's win margin over the best baseline from Table 1 (best baseline loss minus Ambient SFT loss; higher is a bigger win; negative means Ambient SFT was outperformed). Domains are ordered by zero-shot loss.

	Python Code	News	BioMed	CS	Reviews
Zero-shot validation loss	1.46	2.59	2.61	2.87	3.56
Ambient SFT win margin	0.09	0.31	0.01	0.01	0.00

D.2.2 Ablation: effect of a smaller student

Table 15 repeats the domain-adaptation comparison of Table 1 with a 124M-parameter student in place of the 1.5B one, holding every other condition fixed (same target domains, same OWT-pretrained 1.5B teacher for Ambient SFT). *Ambient SFT (Small Teacher)* additionally reports a self-teacher variant, where the frozen teacher is the same 124M OWT-pretrained checkpoint as the student's own initialization, rather than the larger 1.5B model. Comparing against Table 1 isolates the effect of student capacity: every resume-based method improves substantially with the larger student, and Ambient SFT's margin over the baselines is preserved across the scale change.

Table 15: Best validation loss achieved per domain across fine-tuning strategies, 124M student. *Target Only (From Scratch)*: trained from scratch on target-domain data only. *Target Only Fine-tuning*: finetuned from the OWT-pretrained checkpoint on target-domain data only. *Data Mixture Fine-tuning*: finetuned from the OWT-pretrained checkpoint on a clean mixture of target-domain and OWT data (best mixing ratio p per domain). *Ambient SFT (Small Teacher)*: our method against a frozen self-teacher (the same 124M OWT-pretrained checkpoint), using only target-domain data (best p, T per domain). *Ambient SFT (Large Teacher)*: same mechanism, with a frozen 1.5B-parameter teacher in place of the 124M self-teacher. Values rounded to 2 decimal places, the level of precision supported by measured run-to-run variance across random seeds ($\sigma \approx 0.003$–0.006 across two domains checked); differences below this are not distinguishable from noise. Lower is better; best result(s) per domain in **bold** (ties within rounding are all bolded).

Method	Python Code	CS	BioMed	News	Reviews
Target Only (From Scratch)	0.98	2.99	2.62	2.63	3.34
Target Only Fine-tuning	0.88	2.78	2.51	2.32	3.19
Data Mixture Fine-tuning	0.88	2.69	2.47	2.26	**3.11**
Ambient SFT (Small Teacher)	0.86	2.73	2.47	**2.23**	3.15
Ambient SFT (Large Teacher)	**0.83**	**2.68**	**2.43**	**2.23**	3.12

D.2.3 Ablation: effect of a weaker teacher

Table 16 isolates the effect of teacher strength on Ambient SFT, holding the student fixed and varying only whether the frozen teacher is the small (124M) or large (1.5B) OWT-pretrained model. We report this comparison at both student scales considered in this work. At both scales, using the larger teacher matches or improves over the smaller one on every domain, indicating that our method continues to benefit from a stronger teacher even though, unlike classical KD, the teacher is never the stronger model at the target task – it is only used as a coarse, high-temperature regularizer (Section 3.2.2).

Table 16: Effect of teacher strength on Ambient SFT's best validation loss, at both student scales. The 124M teacher is a self-teacher when the student is also 124M; the 1.5B teacher is the same OWT-pretrained checkpoint used as the teacher in Table 1. Values rounded to 2 decimal places; best teacher choice per domain (within each student-size block) in **bold** (ties within rounding are all bolded).

Student	Teacher	Python Code	CS	BioMed	News	Reviews
124M	124M (self)	0.86	2.73	2.47	**2.23**	3.15
124M	1.5B	**0.83**	**2.68**	**2.43**	**2.23**	**3.12**
1.5B	124M	0.79	2.61	2.39	1.96	3.01
1.5B	1.5B (self)	**0.70**	**2.54**	**2.35**	**1.68**	**2.96**

D.3 Instruction Tuning

Figure 7 reports the full data-availability sweep discussed in Section 4.3, and Figure 8 completes Figure 4 with the four tasks left out of the main paper.

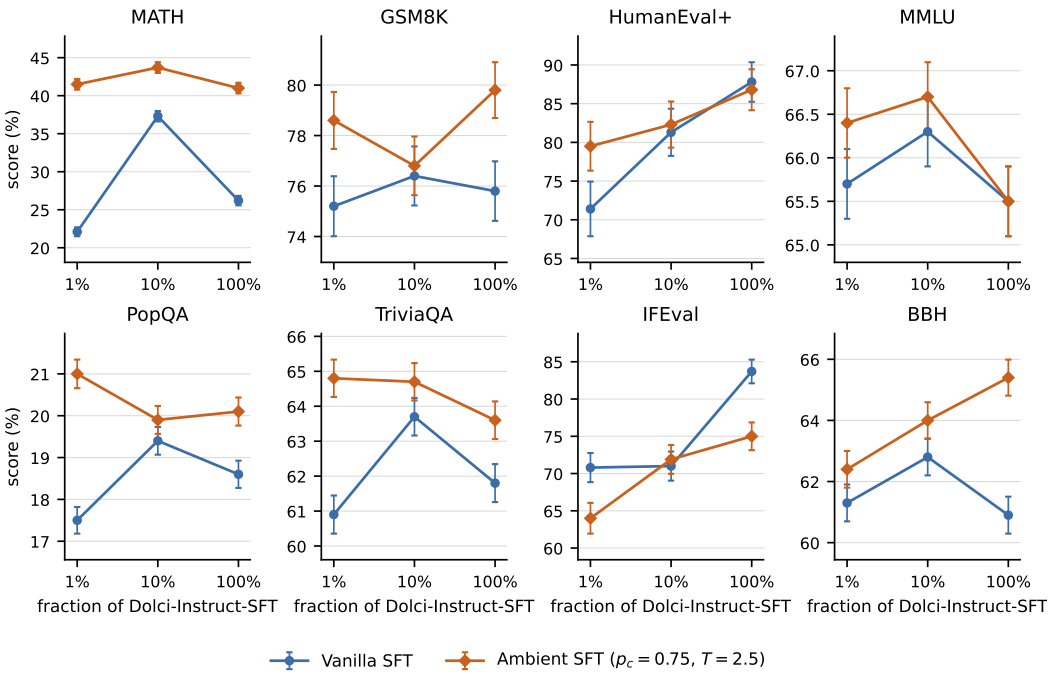


Figure 7: Benchmark scores versus the fraction of Dolci-Instruct-SFT data used for training. For all these results, we start from an OLMo-3-7B model. Error bars are ± 1 binomial standard error over each task's items. For the majority of the benchmarks, Ambient SFT improves upon regular SFT across all data scales. The highest improvements are observed in the 1% of the dataset setting.

D.4 Reasoning

Table 17 evaluates the two reasoning checkpoints of Figure 5 on knowledge benchmarks with the few-shot, non-chat specifications of the instruction-tuning harness (Table 12) on the full test sets. Ambient SFT is ahead on all four benchmarks.

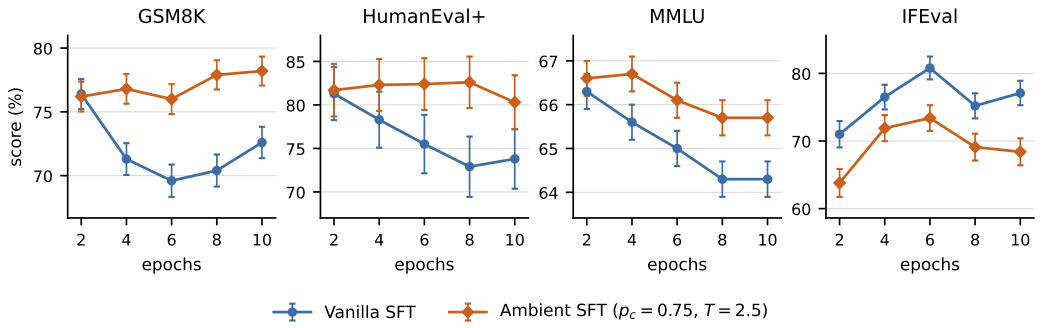


Figure 8: The remaining four tasks of Figure 4: benchmark scores versus the number of epochs on a 10% subset of Dolci-Instruct-SFT, OLMo-3-7B, error bars ± 1 binomial standard error.

Table 17: Knowledge benchmarks for the reasoning checkpoints of Figure 5 (OLMo-3-7B, 1% of Dolci-Think-SFT, ten epochs; Ambient SFT with $p_c = 0.75$, $T = 2.5$), with few-shot, non-chat prompting on the full test sets.

Benchmark	Vanilla SFT	Ambient SFT
MMLU	64.6	**66.0**
GSM8K	77.6	**83.7**
TriviaQA	57.0	**64.3**
BBH	57.0	**66.2**

E Related Work

Our work builds upon the Knowledge Distillation (KD) literature pioneered by Buciluǎ et al. (2006) and Hinton et al. (2015). In the context of LLMs, KD has been successfully applied in the works of (Kim & Rush, 2016; Gu et al., 2023; Zhang et al., 2019). However, whereas classical KD employs a teacher that is stronger or more aligned than the student, our framework inverts this dynamic: the teacher model is strictly *weaker* at the target task. Because the student (the fine-tuned model) is expected to outperform the teacher (the frozen pre-trained model), the teacher's role is not to act as an oracle. Instead, it serves to regularize the fine-tuning process by providing coarse, high-temperature supervision. We also note that applying KD on smoothed distributions has been explored in the paper by (Hinton et al., 2015) and an explicit scaling term $1/T^2$ has been proposed. In our experiments, we do not find this scaling to be optimal across all of our settings, and we apply Knowledge Distillation to a previous version of the same model instead of an external more powerful teacher. The concept of self-distillation has also been explored in (Yang et al., 2024). The main difference between this work and ours is that we match towards the full high-temperature distribution of the teacher, while the authors of (Yang et al., 2024) run SFT over datasets produced by the model.

Closer to our setting is the continual-learning literature on avoiding catastrophic forgetting without access to the original training data. Li & Hoiem (2017) propose recording the pre-update network's outputs on the new task's own inputs and adds a temperature-softened distillation term to the new task's loss, anchoring the update to the model's previous behavior without replaying old data. Our method shares that structure: the frozen pre-trained model is queried on the fine-tuning data itself, and the match is enforced after raising the temperature.

The concept of anchoring a model to a prior state also frequently appears in reinforcement learning (RL) alignment pipelines, where a KL-divergence penalty against a reference model is used to prevent reward hacking (Ouyang et al., 2022). Our framework diverges from this paradigm in two critical ways. First, our method targets the Supervised Fine-Tuning (SFT) stage rather than the RL phase. Second, we exclusively match the high-temperature distributions of the reference model. Enforcing a strict match at $T = 1$ would actively degrade performance by forcing the student to adopt the unaligned base model's lack of task awareness, a manifestation of the phenomenon known as alignment tax (Ouyang et al., 2022). By matching the reference model only at a coarse level, our method successfully avoids this tax while maintaining structural regularization.

Table 18: Comparison of our technique (Ambient SFT) with existing techniques in the literature. Our method targets the SFT part of the post-training process, and it does **not** rely on model rollouts. It is using a noisy teacher (model before fine-tuning) during fine-tuning. To account for the imperfections of the teacher, we only use it at high-temperatures.

Method	**Input Context (x) Source**	**Target (y) Source**	**Loss Type**
Standard SFT	Fine-tuning Dataset	Dataset / Human	Hard labels
SFT Teacher Data	Teacher Rollouts	Teacher Final Output	Hard labels
Knowledge Distillation	Fine-tuning Dataset	Teacher Logits	Soft labels
On-Policy Distillation	Student Rollouts	Teacher Rewards	RL
Off-Policy Distillation	Teacher Rollouts	Teacher Rewards	RL
Ambient SFT (Ours)	Fine-tuning Dataset	**Pre-Trained Model (Same Size)**	**Soft at high-temperature**