LOSSLESS ANTI-DISTILLATION SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Frontier commercial generative models face a growing threat from distillation, whereby a distiller harvests generated responses and trains a competing model at drastically lower cost. Existing defenses either modify the generation to degrade distillation performance, sacrificing response quality, or rely on behavioral detection mechanisms that can be readily bypassed through multi-account querying. In this work, we propose Lossless Anti-Distillation Sampling (LADS), which leaves the generation itself unchanged while substantially reducing the effectiveness of distillation. Concretely, LADS controls the latent randomness underlying inference through a coupling mechanism that preserves within-account generation independence while inducing cross-account dependence. By construction, each benign user, who typically holds only a single account, receives the same experience under LADS as they would without any defense, thereby enjoying a lossless experience. However, for a multi-account task-specific distiller, semantically similar queries submitted across different accounts are assigned coupled randomness, inducing dependence in the harvested data and thereby degrading the generalization performance of the distilled model. Using uniform convergence theory, we show that LADS provably degrades the distiller's generalization gap relative to standard i.i.d. sampling. Experiments on image generation, mathematical reasoning, and code generation confirm that LADS substantially degrades the performance of distilled students while preserving exact statistical fidelity for individual users.

1 INTRODUCTION

Commercial AI evolves rapidly. Once a leading company releases a new frontier model, its competitors come under strong pressure to catch up. A well-known low-risk approach is distillation, in which adversaries harvest large volumes of generated responses by querying the API and then use them to train their own models via supervised fine-tuning or reinforcement fine-tuning. Such a strategy is undesirable for service providers, as it undermines the technical advantage of original innovators and weakens incentives for continued investment in fundamental research.

Given the intense competition among commercial LLM providers, an adversary may need to collect sufficient data and complete fine-tuning within a short period to achieve competitive performance. If the data are collected through a single account, this requires a high query rate, making account-level behavioral detection and restriction relatively straightforward. In practice, a more challenging threat is multi-account distillation, where the adversary controls hundreds of accounts, each designed to behave like a benign user. Existing defenses either degrade benign-user response quality through output-side perturbations (Orekondy et al., 2019; Ma et al., 2021; Savani et al., 2025), attribute distillation only post hoc through watermarking (Zhao et al., 2023; Sander et al., 2024; Xu et al., 2026), or rely on accurate detection models that are readily circumvented under the multi-account distillation setting (Juuti et al., 2019; Chen et al., 2025; Mei et al., 2025). This highlights the need for a defense that is lossless for benign users while remaining effective against multi-account adversaries.

In this paper, we introduce Lossless Anti-Distillation Sampling (LADS). At a high level, LADS views the accesses of different accounts as separate sampling processes and couples the randomness across these processes. The coupling mechanism is carefully designed so that each individual process behaves exactly as it would without the defense, thereby preserving a lossless experience for every individual account, while the samples collected across multiple accounts become statistically dependent, thereby hindering distillation. Concretely, for each generation request q, LADS derives the random seed from the semantic content of the query and the user's visit count within the corresponding

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.

semantic bucket, and generation is then performed using the resulting seed. We show that, for any individual account, the response distribution at each access is identical to that of the original model, while responses across accesses remain independent. Consequently, benign users experience no distortion. At the same time, semantically similar queries submitted across different accounts are coupled through their shared semantics and visit counts. As a result, the samples collected by a multi-account distiller become dependent rather than independent, which reduces sample diversity and consequently degrades the generalization performance of the student model. We support this claim through both a uniform convergence analysis and empirical experiments, as summarized below.

Theoretical Guarantee of LADS. We formalize the multi-account threat as a distiller controlling K accounts, each issuing T queries, and minimizing the standard empirical risk defined over the collected data. We consider two main settings: a warm-up unconditional generation setting (LADS-SIMPLE), where the query q is omitted, and a conditional generation setting (LADS), which is more representative of real-world distillation scenarios. Through a uniform convergence analysis, we show that the generalization gap decreases as $\mathcal{O}(1/\sqrt{KT})$ under standard i.i.d. sampling, but degrades to $\mathcal{O}(1/\sqrt{T})$ under LADS-SIMPLE (Proposition 3.1); under LADS, the generalization gap behaves similarly, depending on the size of the task-associated query space (Theorem 3.2). In other words, by introducing cross-account coupled noise, additional dummy accounts no longer provide sufficiently informative training signals. As a result, the student model's test error cannot be reduced at the same rate by simply scaling up the number of accounts.

Empirical Validation Across Language and Image Scenarios. For image generation, we deploy LADS on diffusion-based EDM2 models. Specifically, we use EDM2-Large (777M) and EDM2-XS (125M) as teachers and EDM2-XXS (31M) as the student. To model the multi-account setting, we distribute ImageNet class queries across 50 accounts. Under this setup, we introduce a noise-mixing mechanism that interpolates between standard i.i.d. sampling and the full LADS scheme, while preserving the exact per-request noise distribution $\mathcal{N}(0, I)$. For language generation, we use Qwen3.5-397B-A17B-FP8 as the teacher, and Llama-3.1-8B, DeepSeek-Math-7B-Base, and Qwen2.5-7B as students. We evaluate distillation quality on MATH, GSM8K and HumanEval. Across both scenarios, we observe that for the student models, FID, validation loss, and task accuracy degrade under LADS, while samples observed by benign users remain of high quality — confirming that LADS achieves the multi-account distillation defense predicted by our theory under a strict losslessness guarantee.

2 Related Work

Distillation in LLMs. Existing work broadly follows two lines. task-specific distillation transfers a targeted capability, such as mathematical reasoning (Tong et al., 2024; Pei et al., 2025), code generation (Luo et al., 2023; Wei et al., 2023), or step-by-step reasoning (Hsieh et al., 2023; Magister et al., 2023), with substantial gains reported in recent years. Task-agnostic distillation, in contrast, aims to replicate the teacher's general instruction-following capability (Liu et al., 2026; Wang et al., 2023; Mukherjee et al., 2023; Agarwal et al., 2024). Its query distribution therefore needs to span a broad range of domains, typically requiring a substantially larger data-collection budget. In practice, task-specific distillation represents a particularly relevant threat for API-based model extraction, as a distiller can concentrate its query budget on a high-value capability. This threat model is also supported by recent industry evidence: Anthropic[1] reported large-scale multi-account distillation campaigns targeting a small number of high-value capabilities, including reasoning, tool use, and coding. This real-world evidence motivates our focus on the task-specific distillation regime.

Anti-distillation by perturbing distribution. A natural defense strategy is to perturb the model's outputs so as to prevent the distiller from reliably recovering the targeted capability from the collected data. Orekondy et al. (2019) perturbed the returned posterior probabilities to maximally distort attacker's gradient signal. Kariyappa & Qureshi (2020) proposed to return incorrect predictions for out-of-distribution queries while preserving correct outputs for in-distribution inputs. Ma et al. (2021) fine-tuned a teacher model that encourages the distribution over incorrect labels to be uniform. More recently, Savani et al. (2025) modified the token-level distribution, which steers sampling toward tokens that increase the student's loss on downstream tasks. However, these approaches inherently introduced bias, which can degrade the output quality perceived by benign users.

[1] https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks

Anti-distillation by adding watermark. Another line of work embeds identifiable signals into the model's outputs so that distillation can be detected post hoc. Zhao et al. (2023) proposed GINSEW, which injects a secret sinusoidal signal into the sampling process. Sander et al. (2024) showed that text watermarks are "radioactive": when a student model is fine-tuned on watermarked text, residuals of the watermark signal persist in the student and can be statistically detected even when only a small fraction of the training data is watermarked. Xu et al. (2026) proposed Antidistillation Fingerprinting, which injects a fingerprint penalty into the generation process. While watermarking provides a mechanism for attribution and legal recourse, it does not prevent the distillation itself, and most schemes also alter the output distribution.

Anti-distillation with session-dynamic defense. Rather than perturbing every response, session-dynamic methods monitor user query patterns and activate defenses only when suspicious behavior is detected. Juuti et al. (2019) proposed PRADA, a detection-based defense that monitors the distribution of consecutive API queries and raises an alarm when this distribution deviates from benign user behavior. Chen et al. (2025) proposed QUEEN, which continuously monitors the sensitivity of incoming queries and begins perturbing output logits once the query pattern exceeds a predefined threshold. Mei et al. (2025) introduced D-ADD, which identifies malicious queries by analyzing per-account distribution discrepancy. However, these approaches can be circumvented by a multi-account adversary who distributes queries across many accounts, each individually appearing benign.

3 METHOD

In this section, we formalize the multi-account distillation threat, present a warm-up algorithm LADS-SIMPLE for unconditional generation, and extend it to the full conditional setting LADS via a semantic-bucket noise coupling mechanism.

3.1 PROBLEM DEFINITION, NOTATIONS AND MOTIVATION

We denote by $\mathcal{X}$ the data space (e.g., image, text). Given any query $q \in \mathcal{Q}$ (e.g., a text prompt), the service provider possesses a private generative model that can produce samples from the conditional distribution $P_{\mathcal{X}}(\cdot \mid q)$. We regard this conditional distribution as the teacher model. Note that sampling from $P_{\mathcal{X}}(\cdot \mid q)$ can often be reparameterized as first drawing a noise variable ϵ (or a collection of independent noise variables) from a simple distribution, and then producing the output via a deterministic mapping $G(q, \epsilon)$. Appendix A provides detailed constructions for diffusion-based image generation and autoregressive text generation. For convenience and ease of presentation, throughout the remainder of the paper, we represent the sampling process by $G(q, \epsilon)$ instead of $P_{\mathcal{X}}(\cdot \mid q)$. Accordingly, we write $x = G(q, \epsilon)$ to denote a sample generated by the model G, regardless of whether x is an image or a sequence of text, and regardless of whether ϵ is a single noise variable, as in typical image generation, or a collection of noise variables, as in language generation.

Definition of the distillation adversary. A single-account distiller can be readily detected through behavioral patterns and subsequently restricted if it needs to collect a large amount of data within a short period for fine-tuning. We therefore focus on a distributed distiller who controls K accounts, indexed by $k \in \{1, \ldots, K\}$, to distill from the teacher model G. Given a limited collection period, we assume that each account issues T queries at most. Let $q_{t,k}$ denote the t-th query submitted by account k, with $x_{t,k}$ denoting the corresponding response induced by noise $\epsilon_{t,k}$. We denote by D_{distill} the collection of query-response pairs obtained from these accounts. The distiller then fine-tunes a student model, parameterized by $\theta \in \Theta$, using an empirical risk of the general form

$$L_{\text{standard}}(\theta) = \frac{1}{|D_{\text{distill}}|} \sum_{(q,x) \in D_{\text{distill}}} \ell(\theta; q, x) = \frac{1}{KT} \sum_{k=1}^{K} \sum_{t=1}^{T} \ell(\theta; q_{t,k}, x_{t,k}), \tag{1}$$

where $\ell(\theta; q, x)$ denotes the per-instance training objective. In supervised fine-tuning, ℓ may take the form of a negative log-likelihood loss for language models or a score-matching loss for diffusion models. In more general reinforcement- or preference-based fine-tuning pipelines, ℓ may represent any surrogate objective computed from the observed query-response pairs.

Definition of lossless benign-user experience. Ideally, a benign user accesses the model through a single account to obtain responses. We call the experience lossless if a user receives the same experience under the defense mechanism as they would without any defense. To formalize this concept, we decompose losslessness into two key properties: distribution preservation and independence.

- The distribution-preservation property holds if, for any individual account—-regardless of whether it is benign or controlled by an adversary—-each response follows the original teacher distribution conditioned on the submitted query, thereby preserving per-query response fidelity.
- The independence property holds if the responses observed across accesses from the same account remain independent, thereby preserving the original repeated-sampling behavior within each account.

Existing defenses primarily modify G to introduce bias into the collected data, thereby breaking the distribution-preservation property. Rather than modifying G, we control the realization of the noise ϵ so as to preserve within-account independence while introducing cross-account dependence.

3.2 A Warm-up Case: Anti-distillation Sampling in Unconditional Generation

For sake of clarity, we begin with the simplest setting in which the user query q is omitted. This setting is known as unconditional generation and has been widely studied in the generative modeling literature (Goodfellow et al., 2014; Ho et al., 2020; Radford et al., 2018). In this case, the generation mapping $G(q, \epsilon)$ and distillation loss $\ell(\theta; q, x)$ reduce to $G(\epsilon)$ and $\ell(\theta; x)$, respectively.

To achieve the goal, we leverage the coupling method, a classical tool in stochastic-process theory (Andersson & Britton, 2012; Lindvall, 2002). Our algorithm LADS-SIMPLE associates the noise ϵ with the access count of each account. Specifically, let an injective mapping $\mathrm{SG} : \mathbb{N} \to \mathbb{N}$ denote a private seed-generation function maintained by the service provider, which assigns the seed $s = \mathrm{SG}(t)$ to the t-th access of any account. A standard seeded pseudorandom sampling routine is then used to draw ϵ from its target distribution. In modern sampling libraries, this can be implemented as $\epsilon = \mathsf{Expand}(F_\kappa(s))$, where F_κ is a keyed pseudorandom generator with a secret key κ sampled once at deployment, and Expand maps the resulting pseudorandom bit sequence to the target noise distribution (e.g., a Gaussian distribution). The output $x = G(\epsilon)$ is then returned to the account.

We first establish that LADS-SIMPLE is lossless for each individual account. Under the standard pseudorandomness assumption, distinct seeds produce independent pseudorandom realizations from the same target distribution. Since G remains unchanged and ϵ follows its original target distribution, the response x is distributed exactly as a sample from the teacher model, and thus the distribution-preservation property holds. Moreover, because SG is injective, distinct access indices $t \neq t'$ are mapped to distinct seeds, so the corresponding noise variables within any fixed account are independent across accesses. Therefore, the independence property also holds. The result follows immediately by combining the two properties.

LADS-SIMPLE can be viewed as a common-random-number coupling (Thorisson, 2000) of the generative distributions across accounts. We next show how this strategy hinders distillation. While each individual account receives independent responses across accesses through distinct seeds, accesses from different accounts at the same access index are assigned the same seed. Thus, across the K accounts controlled by the distiller, the t-th access uses the same noise, i.e., $\epsilon_{t,1} = \cdots = \epsilon_{t,K}$, and consequently produces the same output, $x_{t,1} = \cdots = x_{t,K}$. If the distiller uses these data for fine-tuning, the objective (1) reduces to

$$L_{\mathrm{LADS}}(\theta) = \frac{1}{KT} \sum_{k=1}^{K} \sum_{t=1}^{T} \ell(\theta; x_{t,k}) = \frac{1}{T} \sum_{t=1}^{T} \ell(\theta; x_{t,1}), \tag{2}$$

To theoretically formalize how this mechanism hinders the distiller, we analyze the learning process through the lens of uniform convergence. Let $\mathcal{F} := \{x \mapsto \ell(\theta; x) : \theta \in \Theta\}$ be the induced loss class, and assume that $\ell(\theta; x) \in [0, B]$ for all $\theta \in \Theta$ and x. We define the population risk as $\mathcal{E}(\theta) := \mathbb{E}_{x \sim P_X}[\ell(\theta; x)]$. The following proposition gives a uniform convergence bound between the empirical risk and the population risk.

Proposition 3.1 (Effective Sample Size Collapse). *Let $\mathfrak{R}_n(\mathcal{F})$ denote the Rademacher complexity of $\mathcal{F}$ with sample size n. For any $\delta > 0$, with probability at least $1 - \delta$, the following hold:*

1. For standard i.i.d. sampled data in (1), we have

$$\sup_{\theta \in \Theta} |L_{\mathrm{standard}}(\theta) - \mathcal{E}(\theta)| \leq 2\mathfrak{R}_{KT}(\mathcal{F}) + B\sqrt{\frac{\log(1/\delta)}{2KT}}. \tag{3}$$

2. For LADS-SIMPLE *sampled data in (2), we have*

$$\sup_{\theta \in \Theta} |L_{\mathrm{LADS}}(\theta) - \mathcal{E}(\theta)| \leq 2\mathfrak{R}_{T}(\mathcal{F}) + B\sqrt{\frac{\log(1/\delta)}{2T}}. \tag{4}$$

Algorithm 1 LADS-SIMPLE

Require: Generative model G, private seed generator $\mathrm{SG} : \mathbb{N} \to \mathbb{N}$, keyed pseudorandom generator F_κ with secret key κ, per-account access counters $\{t_u\}$ initialized to 0

function SERVEREQUEST(u)
  $t_u \leftarrow t_u + 1$ ▷ Update the access count of account u
  Seed $s \leftarrow \mathrm{SG}(t_u)$
  Draw noise $\epsilon \leftarrow \mathsf{Expand}(F_\kappa(s))$
  Generate output $x \leftarrow G(\epsilon)$
  return x
end function

The above result follows directly from the standard uniform convergence theorem, and the detailed proof is provided in Appendix B. Although it states only upper bounds, classical results show that Rademacher-based bounds of this form are tight up to constants (Bartlett & Mendelson, 2002; Koltchinskii, 2006; Wainwright, 2019), so the gap between the two rates reflects a genuine statistical separation. To interpret the bound, we adopt the canonical scaling $\mathfrak{R}_n(\mathcal{F}) \le C/\sqrt{n}$, which holds for many hypothesis classes, including linear predictors, multilayer perceptrons, and Transformers under suitable norm constraints (Neyshabur et al., 2015; Golowich et al., 2018; Trauger & Tewari, 2024). Under this scaling, the generalization gap under LADS-SIMPLE is worse than that under standard i.i.d. sampling by a factor of $\sqrt{K}$, demonstrating a substantial reduction in learning effectiveness. Under LADS-SIMPLE, the generalization error depends only on the per-account query budget T. The distiller can improve learning efficiency only by substantially increasing T. This either raises the risk of detection and restriction due to a higher query rate per account, or requires a much longer distillation period, thereby reducing the benefit of rapid model catch-up.

A drawback of LADS-SIMPLE is its limited diversity, which arises from the use of extreme coupling. A natural refinement is to use noise decomposition. For example, when the required noise is Gaussian, we can decompose it into a cross-account-coupled component and an independent component (see Experiment 4.1). In this way, within-account independence is preserved, cross-account dependency is maintained, and overall output diversity is improved. The corresponding uniform convergence analysis (Yu, 1994; Mohri & Rostamizadeh, 2008) has to be carried out case by case, depending on the specific randomization and the dependency structure. However, the general conclusion remains the same: introducing dependency degrades learning efficiency and makes generalization more difficult.

3.3 ANTI-DISTILLATION SAMPLING FOR CONDITIONAL GENERATION

LADS-SIMPLE introduces dependence across accounts through noise coupling. We now extend this idea to the more realistic conditional-generation setting, where users may submit different queries and receive query-conditioned responses. In this setting, directly applying LADS-SIMPLE may induce only weak cross-account dependence: Although two accounts may share the same noise at a given access index, their outputs are given by $G(q, \epsilon)$ and $G(q', \epsilon)$; when q and q' differ substantially, the effect of the shared noise may be dominated by the difference between the queries.

To overcome this limitation, we extend LADS-SIMPLE to incorporate query semantics into the coupling process, leading to our conditional-anti-distillation method, LADS. We assume that the whole query space $\mathcal{Q}$ can be partitioned into p semantic buckets, as specified by a private mapping $h : \mathcal{Q} \to [p] := \{1, 2, \ldots, p\}$. For each user, we maintain a bucket-count array $a \in \mathbb{N}^p$, where $a[i]$ records the number of previous queries from this user that have fallen into bucket i. Initially, all entries in a are set to zero. When a user submits a query q, the provider first identifies the semantic bucket $i = h(q)$ and updates the corresponding counter for this specific user:

$$a[i] \leftarrow a[i] + 1.$$

The random seed s is then determined by

$$s = \mathrm{SG}(i, a[i]),$$

where the seed generator $\mathrm{SG} : [p] \times \mathbb{N} \to \mathbb{N}$ is a private injective mapping. Similar to LADS-SIMPLE, this seed s is used to draw ϵ. The final output is then generated as $x = G(q, \epsilon)$. The full procedure is summarized in Algorithm 2. LADS ties the generation noise to both the semantics of the query and

Algorithm 2 LADS

Require: Generative model G, private clustering function $h : \mathcal{Q} \to [p]$, private seed generator $\mathrm{SG} : [p] \times \mathbb{N} \to \mathbb{N}$, keyed pseudorandom generator F_κ with secret key κ, per-account bucket-count arrays $\{a_u \in \mathbb{N}^p\}$ initialized to 0

function SERVEREQUEST(u, q)
  $i \leftarrow h(q)$ ▷ Identify the semantic bucket of query q
  $a_u[i] \leftarrow a_u[i] + 1$ ▷ Update the bucket count of account u
  Seed $s \leftarrow \mathrm{SG}(i, a_u[i])$
  Draw noise $\epsilon \leftarrow \mathsf{Expand}(F_\kappa(s))$
  Generate output $x \leftarrow G(q, \epsilon)$
  return x
end function

the user's access count within the corresponding semantic bucket. It is straightforward to verify that both the distribution-preservation and independence properties continue to hold. Therefore, LADS remains lossless for each individual account.

As discussed in the related work, task-specific distillation represents an important practical threat. We next show that, under reasonable conditions, LADS can effectively hinder distillation in this setting. Let $\mathcal{Q}_{\text{task}} \subseteq \mathcal{Q}$ denote the query space associated with the task, let $P_{\mathcal{Q}_{\text{task}}}(\cdot)$ be the corresponding query distribution, and let $d_{\mathcal{Q}}(q, q')$ be a semantic metric on $\mathcal{Q}$, which in practice can be instantiated by measuring the distance in the embedding space. Because the distiller targets a specific downstream capability, its queries are often semantically concentrated, in the sense that they occupy only a few localized subregions of the query space rather than being broadly distributed. Denote $H(\mathcal{Q}_{\text{task}}) = \{h(q) \mid q \in \mathcal{Q}_{\text{task}}\}$ for the set of semantic buckets triggered by the task. We then assume that $|H(\mathcal{Q}_{\text{task}})| = N \ll p$.

We assume that each semantic bucket admits a center under $d_{\mathcal{Q}}$, and that the private mapping h assigns each query to its nearest center. Without loss of generality, let $c_1, \ldots, c_N$ denote the centers of the N triggered buckets, and let R denote the maximum distance between any query $q \in \mathcal{Q}_{\text{task}}$ and its assigned center. Thus, R measures how tightly the task queries are concentrated within their corresponding semantic buckets. We define the natural conditional population risk as:

$$\mathcal{E}(\theta) := \mathbb{E}_{q \sim P_{\mathcal{Q}_{\text{task}}}} \mathbb{E}_{x \sim P_{\mathcal{X}}(\cdot|q)}[\ell(\theta; q, x)]. \tag{5}$$

The distiller's empirical risk is $L_{\text{LADS}}(\theta) = \frac{1}{KT} \sum_{k=1}^{K} \sum_{t=1}^{T} \ell(\theta; q_{t,k}, x_{t,k})$, following previous notations. Correspondingly, we define the loss function class for conditional generation as $\mathcal{F} := \{(q, x) \mapsto \ell(\theta; q, x) : \theta \in \Theta\}$ and establish the following uniform convergence bound.

Theorem 3.2 (Effective Sample Size Collapse for LADS)**.** *Assume $G(q, \epsilon)$ is L_G-Lipschitz in q under the metric $d_{\mathcal{Q}}$, and the loss ℓ is L_ℓ-Lipschitz with respect to the input pair (q, x) under their joint metric, and bounded by B. For any $\delta \in (0, 1)$, with probability at least $1 - 2\delta$, we have*

$$\sup_{\theta \in \Theta} |L_{LADS}(\theta) - \mathcal{E}(\theta)| \leq \underbrace{2\mathfrak{R}_{NT}(\mathcal{F}) + B\sqrt{\tfrac{\log(1/\delta)}{2NT}}}_{\textit{Estimation Error w.r.t. Surrogate Target}} + \underbrace{2L_\ell(1 + L_G)R + \Delta_{\text{align}}}_{\textit{Systematic Bias}}, \tag{6}$$

where $\mathfrak{R}_{NT}(\mathcal{F}) = \mathcal{O}(1/\sqrt{NT})$ is the corresponding Rademacher complexity, and $\Delta_{\text{align}} := BN\sqrt{\frac{\log(2N/\delta)}{2KT}}$ captures the statistical fluctuation of cluster weights.

The proof is provided in the appendix C. Theorem 3.2 relies on the Lipschitz regularity of G and ℓ, an assumption commonly used to make theoretical analyses of neural networks tractable (Edelman et al., 2022; Li et al., 2023; Castin et al., 2024). Based on the theorem above, we can see that for LADS, the generalization gap shrinks at rate $\mathcal{O}(1/\sqrt{NT})$ rather than the i.i.d. rate $\mathcal{O}(1/\sqrt{KT})$, so a substantially larger gap persists whenever the number of affected clusters is much smaller than the number of controlled accounts, namely $N \ll K$. Moreover, both the number and quality of the semantic clusters affect the performance of the algorithm. When p is large and the clustering is fine-grained, queries within each bucket are more semantically concentrated, leading to a smaller R and potentially smaller values of L_G and L_ℓ, at the cost of increased storage overhead. In contrast,

when p is small or the clustering quality is poor, queries within the same bucket are less semantically coherent, resulting in a larger R and potentially larger values of L_G and L_ℓ. This leads to a looser generalization bound and weaker control over the student model's generalization performance.

Practical implementation for large-scale system. Industrial-scale deployment of LADS requires scalable construction of the bucket assignments h and maintenance of per-user count arrays, both of which can leverage well-established infrastructure with manageable overhead. First, the bucket assignments h can be instantiated by embedding each query into a dense vector and assigning it to a bucket via locality-sensitive hashing or, alternatively, through a nearest-center lookup in the embedding space. Mature and highly optimized systems are readily available, including FAISS (Johnson et al., 2021; Douze et al., 2026) and ScaNN (Guo et al., 2020), which support billion-scale dense retrieval with millisecond-level query latency. Second, although we write the bucket counter as a dense array $a \in \mathbb{N}^p$ for notational convenience, it can be implemented sparsely. Since each account typically touches only a small subset of the p buckets, the provider only needs to maintain counters for those visited buckets. This requires only lightweight state access and keeps the memory overhead manageable even when the global bucket space is large.

Potential adversarial strategy and task-agnostic distillation. Our current analysis primarily targets task-specific distillation, where the adversary's queries concentrate in a relatively small number of semantic regions. LADS can also be applied to task-agnostic distillation, but its effectiveness may decrease as the number N of activated semantic buckets grows, thereby reducing the amount of cross-account coupling. This reveals a natural trade-off in semantic granularity: finer clustering yields tighter semantic coherence but activates more buckets, whereas coarser clustering increases coupling coverage at the cost of weaker dependence among outputs generated from queries within the same bucket. LADS may also face an adaptive distiller that strategically modifies its behavior to circumvent the defense. For example, the distiller may attempt to reverse-engineer the bucket assignment and then either concentrate semantically similar queries within a single account or adversarially rewrite them so that they are mapped to different buckets. Both strategies can increase the independence of the collected samples. Although such reverse engineering is highly nontrivial, its difficulty alone does not constitute a formal security guarantee. This motivates an empirical evaluation of LADS under imperfect or noisy semantic mappings, which we present in the experimental section.

4 Experiment

In this section, we evaluate LADS across two scenarios and two representative generative paradigms: diffusion-based image generative models and autoregressive language models. Both experimental results support the effectiveness of our proposed algorithm.

4.1 Image Generation

We conduct experiments on class-conditional image generation using the EDM2 models (Karras et al., 2024). EDM2 is a family of pre-trained diffusion models on the ImageNet dataset. For all experiments, the checkpoints with the lowest validation loss are reported.

Simulating multi-account distillation. In our experiment, we employ two pretrained teacher models, EDM2-Large (777M parameters) and EDM2-XS (125M parameters), and we use EDM2-XXS (31M parameters) as the student model. They share the same UNet backbone and differ only in the base channel width. For both teacher-side data generation and student-side evaluation, we use a 32-step second-order Heun ODE sampler. To emulate the threat model, we set the total query budget to be 1M, and set the number of distiller's accounts $K = 50$. Therefore, each of the accounts issues $T = 20k$ queries. The student model is then trained on the collected dataset, with all training/inference hyperparameters following the original EDM2 settings.

Simulating noisy semantic mapping. In this task, we represent each query q as a class label. To simulate a more realistic setting, we construct non-overlapping buckets, each containing four randomly selected class labels. This mimics practical scenarios where semantically distinct queries may be assigned to the same bucket because of clustering errors. Furthermore, for each submitted query, we assign it to the true bucket with probability β, and to a randomly selected bucket with probability $1 - \beta$, where β controls the noise level.

Teacher	α	β	Student FID		Student Loss		Benign User FID
			To real	To teacher	Train	Valid	
EDM2-Large	0.0	0.0	8.50	4.53	0.301	0.340	2.06
	0.7	0.7	9.20	5.20	0.305	0.346	
	0.9	0.9	21.09	16.32	0.317	0.366	
	1.0	1.0	29.8	26.95	0.130	0.848	
EDM2-XS	0.0	0.0	9.39	2.80	0.309	0.345	3.53
	0.7	0.7	9.75	3.13	0.312	0.349	
	0.9	0.9	22.67	13.61	0.322	0.370	
	1.0	1.0	25.94	29.76	0.119	0.823	

Table 1: Evaluation results of student models. When $\alpha = 0$ (highlighted in blue), corresponding to the standard i.i.d. setting, the student model performs well in terms of both FID and loss. As α and β both increase toward 1 (highlighted in red), distillation performance becomes progressively worse, indicating that the proposed algorithm is effective.

LADS implementation and distillation training. In EDM2 models, ϵ follows a standard Gaussian distribution. To quantify the strength of cross-account noise coupling and its impact on distillation performance, we decompose ϵ into a *bucket-count-shared* component and a *per-request independent* component. Concretely, let $\mathrm{GRNG}(s)$ denote a Gaussian Random Number Generator that outputs a sample from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ when initialized with seed s, and let $\mathrm{GRNG}()$ denote the same generator with a randomly chosen seed. SG is a random number list and fixed during our experiment. For any user submitting query that falls into the i-th bucket under the noisy mapping, the provider computes the seed $s = \mathrm{SG}(i, a[i])$ and constructs the noise ϵ according to:

$$\epsilon = \sqrt{\alpha} \cdot \mathrm{GRNG}(s) + \sqrt{1 - \alpha} \cdot \mathrm{GRNG}(), \tag{7}$$

where $\alpha \in [0, 1]$ is the mixing coefficient. It is straightforward to verify that ϵ still follows $\mathcal{N}(\mathbf{0}, \mathbf{I})$, thereby ensuring that generation remains lossless for benign users. The parameter α directly controls the strength of cross-account coupling: when $\alpha = 0$, all noise variables are independent, and the resulting sampling procedure reduces to standard i.i.d. sampling regardless of the value of β; $\alpha = 1$ recovers the extreme coupling regime of LADS.

Evaluation metrics. We evaluate the distilled student model under different configurations using two classes of metrics. The first class is the Fréchet Inception Distance (FID) (Heusel et al., 2017). We consider two comparisons: FID (to real), which calculates the FID score between the student-generated images and the real ImageNet images, and FID (to teacher), which measures how faithfully the student recovers the teacher's output distribution. The second class is the loss value, which more directly reflects the approximation and generalization of the learned models. To evaluate the quality experienced by a benign user, we fix a user account, generate samples independently, and compute the FID between the generated images and the real ones.

Results. Detailed results are presented in Table 1. We observe that the coupling strength α affects the performance of the student model in terms of both FID and validation loss. For the EDM2-Large teacher, FID (to real) rises from 8.50 at $\alpha = 0$ to 29.8 at $\alpha = 1.0$, and FID (to teacher) grows from 4.53 to 26.95, with a similar trend for the EDM2-XS teacher. Moreover, even under an imperfect semantic mapping ($\beta < 1$), the student model still exhibits degraded performance, indicating that LADS may remain effective in realistic settings. The validation loss also increases monotonically with α, indicating a growing failure to generalize to the true data distribution. As for the training loss, note that checkpoints are selected based on validation loss, so the total number of training steps differs across experiments. At $\alpha = 0.0$, the number of training steps is significantly larger than in the other settings, suggesting that the model has not yet overfit. In contrast, at $\alpha = 1.0$, the severely reduced data diversity causes the training loss to decrease much more rapidly. For benign users, the FID remains similar to Karras et al. (2024), indicating that they receive high-quality samples.

4.2 Language Model

Mathematical reasoning and code generation are among the most representative and widely studied distillation targets, as demonstrated by recent works such as DART-Math (Tong et al., 2024), MathFusion (Pei et al., 2025), WizardCoder (Luo et al., 2023), and Magicoder (Wei et al., 2023). Following

Benchmark	Model	No SFT (the pre-trained student)	Standard i.i.d Sampling (SFT student)	LADS Sampling (SFT student)
GSM8K	Llama-3.1-8B	43.59%	73.39% (+29.8)	62.72% (+19.1)
	DeepSeek-Math-7B	23.70%	73.16% (+49.5)	60.13% (+36.4)
MATH	Llama-3.1-8B	13.54%	28.18% (+14.6)	22.47% (+8.9)
	DeepSeek-Math-7B	6.20%	32.66% (+26.5)	27.57% (+21.4)
HumanEval	Qwen2.5-7B	56.70%	71.95% (+15.3)	65.20% (+8.5)

Table 2: Evaluation results of student models fine-tuned on samples collected under different sampling strategies of the teacher model. When the teacher uses standard i.i.d. sampling, student performance (highlighted in blue) improves significantly after fine-tuning on these samples. In contrast, the performance gains are substantially reduced under LADS sampling (highlighted in red).

this line of work, we adopt a similar distillation pipeline: a frontier-scale teacher generates responses for a curated set of prompts, and a smaller student is then fine-tuned on these responses. We use Qwen3.5-397B-A17B-FP8 (Qwen Team, 2026) as the teacher model. For mathematical reasoning, we consider Llama-3.1-8B (Grattafiori et al., 2024) and DeepSeek-Math-7B-Base (Shao et al., 2024) as student models; for code generation, we use Qwen2.5-7B (Qwen et al., 2025) as the student model.

LADS implementation and distillation training. We use training prompts from MATH (Hendrycks et al., 2021) and GSM8K (Cobbe et al., 2021) for mathematical reasoning, and Code Alpaca (Chaudhary, 2023) for code generation. For each base prompt, the teacher generates 9 semantically equivalent rephrasings. In the standard i.i.d. setting, the teacher independently rolls out all 10 variants. In the LADS setting, we set the numbers of semantic buckets to 7.5k, 7.5k, and 20k for the three tasks, respectively. Each query is then mapped to a semantic bucket through nearest-neighbor search in the embedding space. Same-bucket prompts are assigned the same decoding seed with probability 0.7 and different seeds otherwise, simulating a noisy semantic mapping. For teacher rollout, we sample with a maximum token length of 65,536, temperature 0.7, top-p of 0.8, top-k of 20, and a presence penalty of 1.5. The thinking mode of the teacher is disabled during rollout. All student models are fine-tuned for 2 epochs with a learning rate of 2×10^{-5}, cosine learning rate schedule with a warmup ratio of 0.01, weight decay of 0.1. For evaluation, we report the accuracy on GSM8K and MATH test set, and pass@1 on HumanEval (Chen et al., 2021) using the official unit tests.

Results. Slightly different from prior works, we adopt a zero-shot evaluation protocol for all models to ensure a fair comparison across different sampling settings, which results in lower baseline numbers than those reported in the original works. As shown in Table 2, the standard i.i.d. sampling yields substantial gains over the pre-trained student across all models and benchmarks (e.g., +29.8 and +49.5 points on GSM8K). In contrast, under LADS sampling, the improvement is significantly reduced. Specifically, on GSM8K, the distillation gain drops from +29.8 to +11.2 for Llama-3.1-8B and from +49.5 to +35.1 for DeepSeek-Math-7B, corresponding to absolute reductions of 18.6 and 14.4 points, respectively, and a similar trend is observed on MATH and HumanEval. These results confirm that LADS effectively limits the statistical benefit of multi-account distillation.

5 Conclusion and Future Work

We proposed Lossless Anti-Distillation Sampling (LADS), a principled mechanism for mitigating multi-account distillation without compromising benign-user experience. Unlike prior defenses that rely on output perturbation or detection, LADS operates at the level of randomness, preserving exact sampling fidelity for individual users while introducing structured dependence across accounts. Our theoretical analysis shows that this coupling mechanism fundamentally limits the statistical efficiency of distillation. Extensive experiments across both image and language domains demonstrate that LADS weakens distillation performance while leaving benign-user outputs unaffected. Looking forward, an important direction is to design more adaptive and fine-grained semantic grouping strategies that further tighten cross-account coupling under realistic workloads. It is also of interest to study the interaction between LADS and more advanced adversarial strategies. In addition, extending the framework to broader generative paradigms and integrating it with complementary system-level protections may enable a more comprehensive defense against large-scale distillation.

Ethics Statement

This work studies a server-side defense that a model provider can deploy against multi-account distillation of its own model. It does not introduce a new attack, and it does not require collecting, storing, or analyzing any additional user content beyond the per-user, per-bucket access counters described in Section 3.3. By construction the mechanism is lossless: the response distribution seen by any individual account is identical to that of the undefended model, so no user experiences degraded output quality. The main foreseeable side effect is that legitimate parties who intentionally query the service from many accounts with semantically equivalent prompts (for example, large-scale academic evaluation efforts) would observe reduced sample diversity; providers deploying LADS should therefore offer sanctioned, non-coupled access paths for such use cases. All datasets and pre-trained models used in our experiments are publicly available and used in accordance with their licenses.

Reproducibility Statement

All components needed to reproduce our results are specified in the paper. The full mechanism is given as pseudocode in Algorithm 1 (unconditional) and Algorithm 2 (conditional), and the noise-coupling construction used in the image experiments is stated explicitly in Equation (7). Section 4.1 lists the teacher/student checkpoints, sampler settings, query budget, and account count for the EDM2/ImageNet experiments, and Section 4.2 lists the prompt pools, decoding parameters, and fine-tuning hyperparameters for the language experiments. Assumptions and complete proofs for Proposition 3.1 and Theorem 3.2 are given in Appendices B and C.

AI Use Statement

Generative AI tools were used in this work in the following ways. First, they are part of the object of study: the teacher language model was used to construct the prompt pool, rephrasing each seed prompt into nine semantics-preserving variants, and to generate the responses that constitute the distillation training data for all student models (Section 4.2). These generated data are experimental material rather than content of the paper, and their construction is described in full in the corresponding sections. Second, a large language model was used as a general-purpose writing aid for grammar, phrasing, and wording, and for routine coding assistance in the experimental scripts.

We did not use generative AI tools to develop the theoretical model or conceptual framework, to formulate the mathematical claims, to write any of the proofs, to propose or refine the research hypotheses, to design the research methodology, or to interpret the experimental results. All AI-assisted output was reviewed by the authors: generated text was checked for accuracy and originality, and generated code was verified for correctness against the reported experimental setup. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024.

Hakan Andersson and Tom Britton. *Stochastic epidemic models and their statistical analysis*. Springer Science & Business Media, 2012.

Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Communications of the ACM*, 51(1):117–122, 2008.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5817–5840. PMLR, 2024.

Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 380–388, 2002.

Sahil Chaudhary. Code alpaca: An instruction-following llama model for code generation. https://github.com/sahil280114/codealpaca, 2023.

Huajie Chen, Tianqing Zhu, Lefeng Zhang, Bo Liu, Derui Wang, Wanlei Zhou, and Minhui Xue. Queen: Query unlearning against model extraction. *IEEE Transactions on Information Forensics and Security*, 2025.

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.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *IEEE Transactions on Big Data*, 12(2):346–361, 2026. doi: 10.1109/TBDATA.2025.3618474.

Benjamin L. Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5793–5831. PMLR, 2022.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Proceedings of the 31st Conference on Learning Theory (COLT)*, pp. 297–299, 2018.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

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.

Emil Julius Gumbel. *Statistical theory of extreme values and some practical applications: a series of lectures*, volume 33. US Government Printing Office, 1954.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating large-scale inference with anisotropic vector quantization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3887–3896. PMLR, 2020.

Tamir Hazan and Tommi Jaakkola. On the partition function and random maximum a-posteriori perturbations. *arXiv preprint arXiv:1206.6410*, 2012.

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.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2023.

Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 604–613, 1998.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021.

Mika Juuti, Sebastian Szyller, Samuel Marchal, and N Asokan. Prada: protecting against dnn model stealing attacks. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 512–527. IEEE, 2019.

Sanjay Kariyappa and Moinuddin K Qureshi. Defending against model stealing attacks with adaptive misinformation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 770–778, 2020.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. URL https://arxiv.org/abs/2206.00364.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24174–24184, 2024.

Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19565–19594. PMLR, 2023.

Torgny Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002.

Xin Liu, Simin Ma, Shujian Liu, Song Wang, Sathish Reddy Indurthi, Haoyun Deng, Lu Wang, and Kaiqiang Song. Understanding knowledge distillation in post-training: When it helps and when it fails. *arXiv preprint arXiv:2606.22942*, 2026.

R Duncan Luce et al. *Individual choice behavior*, volume 4. Wiley New York, 1959.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*, 2023.

Haoyu Ma, Tianlong Chen, Ting-Kuei Hu, Chenyu You, Xiaohui Xie, and Zhangyang Wang. Undistillable: Making a nasty teacher that cannot teach students. *arXiv preprint arXiv:2105.07381*, 2021.

Chris J Maddison, Daniel Tarlow, and Tom Minka. A* sampling. *Advances in neural information processing systems*, 27, 2014.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Jian-Ping Mei, Weibin Zhang, Jie Chen, Xuyun Zhang, and Tiantian Zhu. Defense against model stealing based on account-aware distribution discrepancy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 604–611, 2025.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. *Advances in neural information processing systems*, 21, 2008.

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

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, pp. 1376–1401, 2015.

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Prediction poisoning: Towards defenses against dnn model stealing attacks. *arXiv preprint arXiv:1906.10908*, 2019.

George Papandreou and Alan L Yuille. Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *2011 international conference on computer vision*, pp. 193–200. IEEE, 2011.

Qizhi Pei, Lijun Wu, Zhuoshi Pan, Yu Li, Honglin Lin, Chenlin Ming, Xin Gao, Conghui He, and Rui Yan. Mathfusion: Enhancing mathematical problem-solving of llm through instruction fusion. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7400–7420, 2025.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze, and Teddy Furon. Watermarking makes language models radioactive. *Advances in Neural Information Processing Systems*, 37: 21079–21113, 2024.

Yash Savani, Asher Trockman, Zhili Feng, Yixuan Even Xu, Avi Schwarzschild, Alexander Robey, Marc Finzi, and J Zico Kolter. Antidistillation sampling. *arXiv preprint arXiv:2504.13146*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

Hermann Thorisson. *Coupling, Stationarity, and Regeneration*. Springer, 2000.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. *Advances in Neural Information Processing Systems*, 37:7821–7846, 2024.

Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1405–1413, 2024.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 13484–13508, 2023.

Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. Magicoder: Empowering code generation with oss-instruct. *arXiv preprint arXiv:2312.02120*, 2023.

Yixuan Even Xu, John Kirchenbauer, Yash Savani, Asher Trockman, Alexander Robey, Tom Goldstein, Fei Fang, and J Zico Kolter. Antidistillation fingerprinting. *arXiv preprint arXiv:2602.03812*, 2026.

John I Yellott Jr. The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2): 109–144, 1977.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pp. 94–116, 1994.

Xuandong Zhao, Yu-Xiang Wang, and Lei Li. Protecting language generation models via invisible watermarking. In *International Conference on Machine Learning*, pp. 42187–42199. PMLR, 2023.

APPENDIX CONTENTS

A Unified Representation of Generative Models **16**

B Proof of Proposition 3.1 **16**

C Proof of Theorem 3.2 **18**

C.1 Task-Specific Conditional Distillation Setup 18

C.2 Task-Specific Covering and LSH-Induced Bucket Compression 19

C.3 Center-Depth Representatives 20

C.4 Local Smoothness and Empirical Replacement Error 21

C.5 Weighted Empirical Process Bound 22

C.6 Alignment with the Natural Conditional Population Risk 24

C.7 Conditional LADS Bound 25

C.8 Derivation of Theorem 3.2 26

C.9 Reduction to the Unconditional Case 27

A Unified Representation of Generative Models

We denote by $\mathcal{X}$ the data space (e.g., image, text). Given any query $q \in \mathcal{Q}$ (e.g., a text prompt), the service provider possesses a private generative model that can produce samples from the conditional distribution $P_{\mathcal{X}}(\cdot \mid q)$. We regard this conditional distribution as the teacher model. Note that sampling from $P_{\mathcal{X}}(\cdot \mid q)$ can often be reparameterized as first drawing a noise variable ϵ (or a collection of independent noise variables) from a simple distribution, and then producing the output via a deterministic mapping $G(q, \epsilon)$.

In diffusion-based image generative models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020; Karras et al., 2022), ϵ is typically drawn from a standard multivariate Gaussian distribution, while the mapping function G is given by the corresponding ODE solver (Song et al., 2021). In language models, output tokens are generated autoregressively, with each token sampled from the categorical distribution over the vocabulary V induced by the softmax of the logits. Equivalently, this sampling procedure can be realized by first drawing a noise vector $\epsilon \in \mathbb{R}^{|V|}$, whose entries are independently sampled from the standard Gumbel distribution (Gumbel, 1954), and then selecting the token index $\arg\max_{i \in V} (z_i + \epsilon_i)$ according to the Gumbel-max trick (Luce et al., 1959; Yellott Jr, 1977; Papandreou & Yuille, 2011; Hazan & Jaakkola, 2012; Maddison et al., 2014), where $z \in \mathbb{R}^{|V|}$ denotes the logit vector and z_i is its i-th element.

B Proof of Proposition 3.1

We prove Proposition 3.1 in a self-contained manner. Recall the unconditional-generation setting in Section 3.2, where the query q is omitted, the generation rule is written as $x = G(\epsilon)$, and the distillation loss reduces to $\ell(\theta; x)$. The induced loss class is

$$\mathcal{F} := \{x \mapsto \ell(\theta; x) : \theta \in \Theta\}, \qquad 0 \le \ell(\theta; x) \le B \quad \text{for all } (\theta, x),$$

and the population risk is

$$\mathcal{E}(\theta) := \mathbb{E}_{x \sim P_{\mathcal{X}}}[\ell(\theta; x)].$$

The proof consists of two steps. We first establish a generic uniform convergence bound for an arbitrary i.i.d. sample of size n. We then apply this bound with $n = KT$ for standard i.i.d. sampling and with $n = T$ for LADS-SIMPLE.

Step 1: A self-contained uniform deviation bound. Let $Z_1, \dots, Z_n \overset{\text{i.i.d.}}{\sim} P_{\mathcal{X}}$, and define

$$\Psi(Z_1, \dots, Z_n) := \sup_{\theta \in \Theta} \left| \tfrac{1}{n} \textstyle\sum_{i=1}^n \ell(\theta; Z_i) - \mathcal{E}(\theta) \right|.$$

We claim that for every $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\Psi(Z_1, \dots, Z_n) \le 2\mathfrak{R}_n(\mathcal{F}) + B\sqrt{\tfrac{\log(1/\delta)}{2n}}. \tag{8}$$

We prove this claim in two parts.

Step 1a: Concentration around the mean. Consider Ψ as a function of the sample $Z_1, \dots, Z_n$. Fix an index $j \in \{1, \dots, n\}$, and let

$$Z^{(j)} := (Z_1, \dots, Z_{j-1}, Z'_j, Z_{j+1}, \dots, Z_n)$$

be the sample obtained by replacing Z_j with another value Z'_j. For every fixed $\theta \in \Theta$,

$$\begin{aligned} &\left| \left(\tfrac{1}{n} \textstyle\sum_{i=1}^n \ell(\theta; Z_i) - \mathcal{E}(\theta) \right) - \left(\tfrac{1}{n} \textstyle\sum_{i \ne j} \ell(\theta; Z_i) + \tfrac{1}{n} \ell(\theta; Z'_j) - \mathcal{E}(\theta) \right) \right| \\ &\quad = \tfrac{1}{n} \left| \ell(\theta; Z_j) - \ell(\theta; Z'_j) \right| \le \tfrac{B}{n}, \end{aligned}$$

because $\ell(\theta; \cdot) \in [0, B]$. Moreover, for any two collections $\{a_\theta\}_{\theta \in \Theta}$ and $\{b_\theta\}_{\theta \in \Theta}$,

$$|\sup_{\theta \in \Theta} a_\theta - \sup_{\theta \in \Theta} b_\theta| \le \sup_{\theta \in \Theta} |a_\theta - b_\theta|.$$

Applying this inequality to the absolute-deviation functions inside Ψ, we obtain

$$\left| \Psi(Z_1, \dots, Z_n) - \Psi(Z^{(j)}) \right| \le \tfrac{B}{n}.$$

Thus changing one coordinate of the sample can change Ψ by at most B/n. By McDiarmid's inequality, for every $t > 0$,

$$\Pr(\Psi(Z_1, \dots, Z_n) - \mathbb{E}[\Psi(Z_1, \dots, Z_n)] \geq t) \leq \exp\Big(-\tfrac{2t^2}{\sum_{i=1}^n (B/n)^2}\Big) = \exp\Big(-\tfrac{2nt^2}{B^2}\Big).$$

Setting

$$t = B\sqrt{\tfrac{\log(1/\delta)}{2n}}$$

yields, with probability at least $1 - \delta$,

$$\Psi(Z_1, \dots, Z_n) \leq \mathbb{E}[\Psi(Z_1, \dots, Z_n)] + B\sqrt{\tfrac{\log(1/\delta)}{2n}}. \tag{9}$$

Step 1b: Bounding the mean by Rademacher complexity. It remains to bound $\mathbb{E}\Psi$. Introduce an independent ghost sample $Z'_1, \dots, Z'_n \overset{\text{i.i.d.}}{\sim} P_{\mathcal{X}}$. Since

$$\mathcal{E}(\theta) = \mathbb{E}_{Z'}\left[\tfrac{1}{n}\sum_{i=1}^n \ell(\theta; Z'_i)\right],$$

Jensen's inequality gives

$$\begin{aligned}\mathbb{E}\Psi &= \mathbb{E}_Z\left[\sup_{\theta\in\Theta}\left|\tfrac{1}{n}\sum_{i=1}^n \ell(\theta; Z_i) - \mathbb{E}_{Z'}\left[\tfrac{1}{n}\sum_{i=1}^n \ell(\theta; Z'_i)\right]\right|\right] \\ &\leq \mathbb{E}_{Z,Z'}\left[\sup_{\theta\in\Theta}\left|\tfrac{1}{n}\sum_{i=1}^n \big(\ell(\theta; Z_i) - \ell(\theta; Z'_i)\big)\right|\right].\end{aligned}$$

Now let $\sigma_1, \dots, \sigma_n$ be i.i.d. Rademacher random variables, independent of Z and Z'. Because (Z_i, Z'_i) and (Z'_i, Z_i) have the same joint law, the vector

$$(\ell(\theta; Z_i) - \ell(\theta; Z'_i))_{i=1}^n$$

has the same distribution as

$$\big(\sigma_i\big(\ell(\theta; Z_i) - \ell(\theta; Z'_i)\big)\big)_{i=1}^n.$$

Hence

$$\begin{aligned}\mathbb{E}\Psi &\leq \mathbb{E}_{Z,Z',\sigma}\left[\sup_{\theta\in\Theta}\left|\tfrac{1}{n}\sum_{i=1}^n \sigma_i\big(\ell(\theta; Z_i) - \ell(\theta; Z'_i)\big)\right|\right] \\ &\leq \mathbb{E}_{Z,Z',\sigma}\left[\sup_{\theta\in\Theta}\left|\tfrac{1}{n}\sum_{i=1}^n \sigma_i\ell(\theta; Z_i)\right|\right] + \mathbb{E}_{Z,Z',\sigma}\left[\sup_{\theta\in\Theta}\left|\tfrac{1}{n}\sum_{i=1}^n \sigma_i\ell(\theta; Z'_i)\right|\right],\end{aligned}$$

where the second inequality follows from the triangle inequality. The two terms are identical by symmetry, so

$$\mathbb{E}\Psi \leq 2\mathbb{E}_{Z,\sigma}\left[\sup_{\theta\in\Theta}\left|\tfrac{1}{n}\sum_{i=1}^n \sigma_i\ell(\theta; Z_i)\right|\right].$$

By the definition of the Rademacher complexity used in the main text, the last display is exactly $2\mathfrak{R}_n(\mathcal{F})$. Therefore,

$$\mathbb{E}\Psi \leq 2\mathfrak{R}_n(\mathcal{F}). \tag{10}$$

Combining (9) and (10) proves (8).

Step 2: Applying the generic bound to the two sampling schemes.

Standard i.i.d. sampling. Under the standard protocol, the KT queried outputs

$$x_{t,k}, \qquad t = 1, \dots, T, \quad k = 1, \dots, K,$$

are i.i.d. samples from $P_{\mathcal{X}}$. Relabel them as $Z_1, \dots, Z_{KT}$. Then

$$L_{\text{standard}}(\theta) = \tfrac{1}{KT}\sum_{k=1}^K\sum_{t=1}^T \ell(\theta; x_{t,k}) = \tfrac{1}{KT}\sum_{i=1}^{KT} \ell(\theta; Z_i).$$

Applying (8) with $n = KT$, we obtain that with probability at least $1 - \delta$,

$$\sup_{\theta\in\Theta}|L_{\text{standard}}(\theta) - \mathcal{E}(\theta)| \leq 2\mathfrak{R}_{KT}(\mathcal{F}) + B\sqrt{\tfrac{\log(1/\delta)}{2KT}}.$$

LADS-SIMPLE *sampling.* Under LADS-SIMPLE, the t-th access of every controlled account uses the same seed $\mathrm{SG}(t)$. Since generation is deterministic once the latent noise is fixed, all controlled accounts receive the same output at the same access index:

$$x_{t,1} = x_{t,2} = \cdots = x_{t,K}, \qquad t = 1, \dots, T.$$

Therefore,

$$\begin{aligned} L_{\mathrm{LADS}}(\theta) &= \tfrac{1}{KT}\sum_{k=1}^{K}\sum_{t=1}^{T}\ell(\theta; x_{t,k}) \\ &= \tfrac{1}{KT}\sum_{t=1}^{T}\sum_{k=1}^{K}\ell(\theta; x_{t,1}) \\ &= \tfrac{1}{T}\sum_{t=1}^{T}\ell(\theta; x_{t,1}). \end{aligned}$$

Thus, although the distiller issues KT total queries, the empirical risk depends only on the T distinct outputs

$$x_{1,1}, \ldots, x_{T,1}.$$

Because SG is injective, distinct access indices are mapped to distinct seeds. Under the pseudorandomness assumption stated in Section 3.2, the resulting $\epsilon_{1,1}, \ldots, \epsilon_{T,1}$ are independent draws from the target noise distribution. Since $x = G(\epsilon)$, the corresponding outputs are therefore i.i.d. draws from $P_{\mathcal{X}}$. Consequently,

$$x_{1,1}, \ldots, x_{T,1} \overset{\text{i.i.d.}}{\sim} P_{\mathcal{X}}.$$

Relabel $Y_t := x_{t,1}$, $t = 1, \ldots, T$. Then

$$L_{\mathrm{LADS}}(\theta) = \tfrac{1}{T}\sum_{t=1}^{T}\ell(\theta; Y_t), \qquad Y_1, \ldots, Y_T \overset{\text{i.i.d.}}{\sim} P_{\mathcal{X}}.$$

Applying (8) with $n = T$, we obtain that with probability at least $1 - \delta$,

$$\sup_{\theta\in\Theta} |L_{\mathrm{LADS}}(\theta) - \mathcal{E}(\theta)| \le 2\mathfrak{R}_T(\mathcal{F}) + B\sqrt{\tfrac{\log(1/\delta)}{2T}}.$$

Combining the two cases proves Proposition 3.1.

C Proof of Theorem 3.2

This appendix provides the proof of Theorem 3.2. We introduce the task-specific covering, bucket-level compression, center-depth multiplicities, and population-alignment quantities needed for the analysis. Throughout, N denotes the number of activated task-specific semantic buckets, consistently with Section 3.3.

C.1 Task-Specific Conditional Distillation Setup

We use the threat model from Section 3.1: over the limited collection period, the distiller controls K accounts and each account issues at most T queries. For account $k \in [K]$ and within-account round $t \in [T]$, the distiller submits a query $q_{t,k}$ and receives

$$x_{t,k} = G(q_{t,k}, \epsilon_{t,k}).$$

When convenient, we flatten the collected query-response pairs and write

$$\{(q_m, x_m)\}_{m=1}^{KT}, \qquad x_m = G(q_m, \epsilon_m).$$

We focus on task-specific distillation. Let

$$\mathcal{Q}_{\mathrm{task}} \subseteq \mathcal{Q}$$

be the corresponding family of task-related prompts, and let $P_{\mathcal{Q}_{\mathrm{task}}}$ be the task-specific query distribution. The collected queries are task-local:

$$q_m \in \mathcal{Q}_{\mathrm{task}}, \qquad 1 \le m \le KT.$$

The empirical distillation risk is

$$L_{\mathrm{LADS}}(\theta) := \tfrac{1}{KT}\sum_{m=1}^{KT}\ell(\theta; q_m, x_m).$$

The natural conditional population risk is

$$\mathcal{E}(\theta) := \mathbb{E}_{q\sim P_{\mathcal{Q}_{\mathrm{task}}}}\mathbb{E}_{x\sim P_{\mathcal{X}}(\cdot|q)}[\ell(\theta; q, x)].$$

The service provider fixes a private semantic hash

$$h : \mathcal{Q} \to [p] := \{1, 2, \ldots, p\}.$$

For each account k, the provider maintains a bucket-count array $a_k \in \mathbb{N}^p$. When account k submits query q, the provider computes $i = h(q)$, updates $a_k[i] \leftarrow a_k[i] + 1$, and sets the random seed to

$$s = \mathrm{SG}(h(q), a_k[h(q)]).$$

Thus, in the conditional setting, coupling is induced by the pair

$$\text{semantic bucket} + \text{bucket-local depth}.$$

Requests from different accounts that fall into the same bucket and have the same bucket-local depth share the same latent noise. Requests in different bucket-depth cells use independent latent noises.

C.2 Task-Specific Covering and LSH-Induced Bucket Compression

The purpose of the task-specific covering assumption is to formalize the idea that a multi-account distiller may submit many surface-level variants of prompts, but the semantic complexity of a fixed task remains much smaller.

For consistency with the simplified notation in Section 3.3, we use N to denote the nominal task-specific semantic complexity. In the general construction below, the number of occupied semantic buckets may exceed N due to queries that fail the center-and-hash assignment. Proposition C.2 quantifies this deviation explicitly. The simplified regime used in Theorem 3.2 corresponds to the clean assignment case in which these additional bad-mass contributions vanish.

As introduced in Section 3.3, the centers $c_j \in \mathcal{Q}$ are representative queries in the query space. The theory holds for any valid task-relevant semantic metric $d_{\mathcal{Q}}$ over $\mathcal{Q}$, which in practical implementations may be realized through distances between dense query embeddings.

Assumption C.1 (task-specific covering)**.** There exist centers $c_1, \ldots, c_N \in \mathcal{Q}$ and a radius $R > 0$ such that

$$\Pr_{q \sim P_{\mathcal{Q}_{\mathrm{task}}}} \left(q \in \bigcup_{j=1}^{N} B(c_j, R) \right) \geq 1 - \eta,$$

where

$$B(c_j, R) := \{q \in \mathcal{Q} : d_{\mathcal{Q}}(q, c_j) \leq R\},$$

and $\eta \in (0, 1)$ is small.

This assumption says that most task-specific queries lie near one of N semantic centers under the task-relevant metric $d_{\mathcal{Q}}$.

We next connect this covering structure to bucket-level compression through the lens of locality-sensitive hashing (Indyk & Motwani, 1998; Charikar, 2002; Andoni & Indyk, 2008). Following the standard definition of Indyk & Motwani (1998), let $\mathcal{H}$ be an (R, cR, π_1, π_2)-sensitive hash family on $(\mathcal{Q}, d_{\mathcal{Q}})$: for any $q, q' \in \mathcal{Q}$,

$$d_{\mathcal{Q}}(q, q') \leq R \implies \Pr_{h \sim \mathcal{H}}[h(q) = h(q')] \geq \pi_1,$$

and

$$d_{\mathcal{Q}}(q, q') \geq cR \implies \Pr_{h \sim \mathcal{H}}[h(q) = h(q')] \leq \pi_2,$$

where $c > 1$ and $\pi_1 > \pi_2$. Such families are known to exist for the metrics relevant here, including Hamming and ℓ_p distances (Indyk & Motwani, 1998; Andoni & Indyk, 2008) and angular or cosine distance (Charikar, 2002); the last is the natural choice when $d_{\mathcal{Q}}$ is instantiated through distances between normalized dense query embeddings, as in Section 3.3. The provider samples a private hash $h \sim \mathcal{H}$ at deployment time and keeps it fixed over the collection period.

Let

$$B_{\mathrm{task}} := \{h(q_m) : 1 \leq m \leq KT\}$$

be the set of occupied buckets among the collected task-specific queries.

Proposition C.2 (LSH-induced bucket compression)**.** *Consider the KT collected query-response pairs. Suppose Assumption C.1 holds, and suppose the private hash $h \sim \mathcal{H}$ is sampled from an (R, cR, π_1, π_2)-sensitive hash family. Then there exists a nonnegative random variable M_{bad} such that*

$$|B_{\text{task}}| \le N + M_{\text{bad}},$$

and

$$\mathbb{E}[M_{\text{bad}}] \le KT(\eta + 1 - \pi_1).$$

If the collected queries $q_1, \dots, q_{KT}$ are independently sampled from $P_{\mathcal{Q}_{\text{task}}}$, then for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$|B_{\text{task}}| \le N + KT(\eta + 1 - \pi_1) + \sqrt{\tfrac{KT}{2} \log \tfrac{1}{\delta}}.$$

Proof. For each query q_m, define the bad event E_m as follows: either q_m does not lie in the covering union $\cup_{j=1}^{N} B(c_j, R)$, or it lies in some covering ball but fails to hash to the same bucket as its assigned center. More precisely, fix an assignment $J(q_m) \in [N]$ whenever $q_m \in \cup_{j=1}^{N} B(c_j, R)$ and $d_{\mathcal{Q}}(q_m, c_{J(q_m)}) \le R$. Then E_m occurs if either the assignment does not exist or

$$h(q_m) \ne h(c_{J(q_m)}).$$

By Assumption C.1 and the near-neighbor collision property of LSH,

$$\Pr(E_m) \le \eta + (1 - \pi_1).$$

Let

$$M_{\text{bad}} := \textstyle\sum_{m=1}^{KT} \mathbf{1}_{E_m}.$$

Then

$$\mathbb{E}[M_{\text{bad}}] \le KT(\eta + 1 - \pi_1).$$

If q_m is not bad, then it hashes to one of the center buckets $\{h(c_1), \dots, h(c_N)\}$. Therefore all occupied buckets are contained in the N center buckets plus the buckets contributed by bad queries, giving

$$|B_{\text{task}}| \le N + M_{\text{bad}}.$$

When the collected queries are independent, M_{bad} is a sum of Bernoulli variables with mean at most $KT(\eta + 1 - \pi_1)$. Hoeffding's inequality yields the high-probability bound. □

C.3 Center-Depth Representatives

On the high-probability bucket-compression event, we call a query q_m good if both of the following hold:

$$q_m \in B(c_{J(q_m)}, R), \qquad h(q_m) = h(c_{J(q_m)}).$$

Let $I_{\text{good}} \subseteq [KT]$ be the set of good indices and $I_{\text{bad}} = [KT] \setminus I_{\text{good}}$. Define the empirical bad query fraction

$$\rho_\delta := \tfrac{|I_{\text{bad}}|}{KT}.$$

For each good query q_m, let $j(m) \in [N]$ be its assigned center index, so that

$$d_{\mathcal{Q}}(q_m, c_{j(m)}) \le R, \qquad h(q_m) = h(c_{j(m)}).$$

Let $s(m) \in [T]$ denote the bucket-local depth of query q_m in the bucket $h(c_{j(m)})$. Since each account issues at most T queries over the collection period, $s(m) \le T$.

For each center-depth pair $(j, s) \in [N] \times [T]$, define

$$\epsilon_{j,s}$$

as the latent noise generated by seed

$$\text{SG}(h(c_j), s),$$

and define the corresponding center output

$$\widetilde{x}_{j,s} := G(c_j, \epsilon_{j,s}).$$

For every good query q_m, the true output and the center output share the same latent noise:

$$x_m = G(q_m, \epsilon_{j(m),s(m)}), \qquad \widetilde{x}_{j(m),s(m)} = G(c_{j(m)}, \epsilon_{j(m),s(m)}).$$

Define the multiplicity of each center-depth cell by

$$n_{j,s} := \big|\{m \in I_{\text{good}} : (j(m), s(m)) = (j,s)\}\big|,$$

and define the empirical weight

$$w_{j,s} := \tfrac{n_{j,s}}{KT}.$$

Then

$$\textstyle\sum_{j=1}^{N}\sum_{s=1}^{T} w_{j,s} = \frac{|I_{\text{good}}|}{KT} = 1 - \rho_\delta.$$

The compressed weighted empirical risk is

$$\widetilde{L}(\theta) := \textstyle\sum_{j=1}^{N}\sum_{s=1}^{T} w_{j,s}\, \ell(\theta; c_j, \widetilde{x}_{j,s}).$$

The corresponding weighted surrogate population target is

$$\widetilde{\mathcal{E}}(\theta) := \textstyle\sum_{j=1}^{N}\sum_{s=1}^{T} w_{j,s}\, \mathbb{E}_{x\sim P_{\mathcal{X}}(\cdot|c_j)}[\ell(\theta; c_j, x)].$$

This target is not the original conditional population risk. It is the surrogate population target naturally induced by the center-depth compressed empirical process.

C.4 Local Smoothness and Empirical Replacement Error

We use the following local regularity assumptions.

Assumption C.3 (task-specific conditional smoothness)**.** For the fixed task-specific query space $\mathcal{Q}_{\text{task}}$, there exists $L_G > 0$ such that for all $q, q' \in \mathcal{Q}_{\text{task}}$ and all latent noises ϵ,

$$d_{\mathcal{X}}(G(q,\epsilon), G(q',\epsilon)) \le L_G\, d_{\mathcal{Q}}(q,q').$$

Assumption C.4 (Lipschitz and bounded distillation loss)**.** There exist constants $L_\ell > 0$ and $B > 0$ such that for all $\theta \in \Theta$ and all $(q,x), (q',x') \in \mathcal{Q}_{\text{task}} \times \mathcal{X}$,

$$|\ell(\theta; q, x) - \ell(\theta; q', x')| \le L_\ell\big(d_{\mathcal{Q}}(q,q') + d_{\mathcal{X}}(x,x')\big),$$

and

$$0 \le \ell(\theta; q, x) \le B.$$

Lemma C.5 (Uniform empirical replacement by centers)**.** *On the bucket-compression event, if the empirical bad-query fraction is at most ρ_δ, then*

$$\sup_{\theta\in\Theta} |L_{\text{LADS}}(\theta) - \widetilde{L}(\theta)| \le L_\ell(1 + L_G)R + B\rho_\delta.$$

Proof. Decompose the empirical risk into good and bad queries:

$$L_{\text{LADS}}(\theta) = \tfrac{1}{K}T\textstyle\sum_{m\in I_{\text{good}}} \ell(\theta; q_m, x_m) + \tfrac{1}{K}T\sum_{m\in I_{\text{bad}}} \ell(\theta; q_m, x_m).$$

By the definition of the weights,

$$\widetilde{L}(\theta) = \tfrac{1}{K}T\textstyle\sum_{m\in I_{\text{good}}} \ell(\theta; c_{j(m)}, \widetilde{x}_{j(m),s(m)}).$$

Therefore, by the triangle inequality,

$$\begin{aligned}\Big|L_{\text{LADS}}(\theta) - \widetilde{L}(\theta)\Big| &= \Big|\tfrac{1}{K}T\textstyle\sum_{m\in I_{\text{good}}}\Big[\ell(\theta; q_m, x_m) - \ell(\theta; c_{j(m)}, \widetilde{x}_{j(m),s(m)})\Big] + \tfrac{1}{K}T\sum_{m\in I_{\text{bad}}} \ell(\theta; q_m, x_m)\Big| \\ &\le \tfrac{1}{KT}\textstyle\sum_{m\in I_{\text{good}}}\big|\ell(\theta; q_m, x_m) - \ell(\theta; c_{j(m)}, \tilde{x}_{j(m),s(m)})\big| + \tfrac{1}{KT}\sum_{m\in I_{\text{bad}}} |\ell(\theta; q_m, x_m)|.\end{aligned}$$

For a good query q_m, the true output and representative output share the same latent noise:

$$x_m = G(q_m, \epsilon_{j(m),s(m)}), \qquad \widetilde{x}_{j(m),s(m)} = G(c_{j(m)}, \epsilon_{j(m),s(m)}).$$

Using Assumptions C.3 and C.4,

$$\begin{aligned}&\big|\ell(\theta; q_m, x_m) - \ell(\theta; c_{j(m)}, \widetilde{x}_{j(m),s(m)})\big| \\ &\quad\le L_\ell\Big(d_{\mathcal{Q}}(q_m, c_{j(m)}) + d_{\mathcal{X}}(G(q_m, \epsilon_{j(m),s(m)}), G(c_{j(m)}, \epsilon_{j(m),s(m)}))\Big) \\ &\quad\le L_\ell(1 + L_G)R.\end{aligned}$$

For bad queries, boundedness gives

$$0 \le \ell(\theta; q_m, x_m) \le B.$$

Averaging the good-query replacement error and the bad-query contribution gives

$$\sup_{\theta\in\Theta} |L_{\text{LADS}}(\theta) - \widetilde{L}(\theta)| \le L_\ell(1 + L_G)R + B\rho_\delta.$$

□

C.5 Weighted Empirical Process Bound

The compressed empirical process is indexed by center-depth cells $(j, s) \in [N] \times [T]$. Its multiplicities need not be uniform. We therefore use a weighted empirical-process bound.

Define the weighted effective sample size associated with the unnormalized good-query weights by

$$n_{\text{eff}} := \frac{1}{\sum_{j=1}^{N} \sum_{s=1}^{T} w_{j,s}^2}.$$

Recall that these weights satisfy

$$\sum_{j=1}^{N} \sum_{s=1}^{T} w_{j,s} = 1 - \rho_\delta,$$

rather than summing to one in general. Since at most NT weights are nonzero, Cauchy's inequality gives

$$\sum_{j=1}^{N} \sum_{s=1}^{T} w_{j,s}^2 \geq \frac{(1-\rho_\delta)^2}{NT},$$

and therefore

$$n_{\text{eff}} \leq \frac{NT}{(1-\rho_\delta)^2}.$$

In the clean-bucket case $\rho_\delta = 0$, this reduces to

$$n_{\text{eff}} \leq NT.$$

Moreover, when the transcript is balanced over the activated center-depth cells, so that

$$w_{j,s} \approx \frac{1-\rho_\delta}{NT},$$

we have

$$n_{\text{eff}} \approx \frac{NT}{(1-\rho_\delta)^2}.$$

In the clean balanced special case, $\rho_\delta = 0$ and hence

$$n_{\text{eff}} \approx NT.$$

Lemma C.6 (Weighted uniform convergence)**.** *Let $\mathcal{F}$ be a class of measurable functions taking values in $[0, B]$. Let $Z_1, \ldots, Z_m$ be mutually independent, and let $w_1, \ldots, w_m \geq 0$ be fixed weights. Define*

$$L_w(f) := \sum_{i=1}^{m} w_i f(Z_i), \qquad E_w(f) := \sum_{i=1}^{m} w_i \mathbb{E}[f(Z_i)].$$

Let $\mathbf{w} := (w_1, \ldots, w_m)$. Also define the weighted Rademacher complexity

$$\mathfrak{R}_{\mathbf{w}}(\mathcal{F}) := \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\sum_{i=1}^{m} \sigma_i w_i f(Z_i)\right|\right],$$

where $\sigma_1, \ldots, \sigma_m$ are independent Rademacher random variables. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F}} |L_w(f) - E_w(f)| \leq 2\mathfrak{R}_{\mathbf{w}}(\mathcal{F}) + B\sqrt{\frac{\log(1/\delta)}{2} \sum_{i=1}^{m} w_i^2}.$$

Equivalently,

$$\sup_{f \in \mathcal{F}} |L_w(f) - E_w(f)| \leq 2\mathfrak{R}_{\mathbf{w}}(\mathcal{F}) + B\sqrt{\frac{\log(1/\delta)}{2n_{\text{eff}}}}.$$

Proof. We give the argument for completeness, since this is the weighted analogue of the uniform convergence bound used in Appendix B. Define

$$\Phi(Z_1, \ldots, Z_m) := \sup_{f \in \mathcal{F}} \left|\sum_{i=1}^{m} w_i \big(f(Z_i) - \mathbb{E}f(Z_i)\big)\right|.$$

First, we bound its expectation. Let $Z_1', \ldots, Z_m'$ be an independent ghost sample, where Z_i' has the same distribution as Z_i. By Jensen's inequality,

$$\mathbb{E}\Phi \leq \mathbb{E}_{Z,Z'} \sup_{f \in \mathcal{F}} \left|\sum_{i=1}^{m} w_i \big(f(Z_i) - f(Z_i')\big)\right|.$$

Introducing independent Rademacher variables $\sigma_1, \ldots, \sigma_m$, and using the symmetry of (Z_i, Z_i'), we obtain

$$\mathbb{E}\Phi \leq \mathbb{E}_{Z,Z',\sigma} \sup_{f \in \mathcal{F}} \left|\sum_{i=1}^{m} \sigma_i w_i \big(f(Z_i) - f(Z_i')\big)\right|.$$

By the triangle inequality and the fact that the two ghost-sample terms have the same distribution,

$$\mathbb{E}\Phi \le 2\mathbb{E}_{Z,\sigma} \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^m \sigma_i w_i f(Z_i)\right|.$$

Equivalently, applying the same definition of weighted Rademacher complexity to the symmetrized class $\mathcal{F} \cup (-\mathcal{F})$, whose complexity is the same up to the displayed absolute value, gives

$$\mathbb{E}\Phi \le 2\mathfrak{R}_{\mathbf{w}}(\mathcal{F}).$$

It remains to concentrate Φ around its mean. If only the i-th coordinate is replaced by an arbitrary Z_i'', then for every $f \in \mathcal{F}$,

$$|w_i(f(Z_i) - \mathbb{E}f(Z_i)) - w_i(f(Z_i'') - \mathbb{E}f(Z_i))| \le Bw_i,$$

because $f \in [0, B]$. Taking the supremum over f preserves this bounded-difference constant. Hence McDiarmid's inequality gives, for every $t > 0$,

$$\Pr\{\Phi - \mathbb{E}\Phi \ge t\} \le \exp\left(-\tfrac{2t^2}{B^2\sum_{i=1}^m w_i^2}\right).$$

Setting

$$t = B\sqrt{\tfrac{\log(1/\delta)}{2}\sum_{i=1}^m w_i^2}$$

yields, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}} |L_w(f) - E_w(f)| \le 2\mathfrak{R}_{\mathbf{w}}(\mathcal{F}) + B\sqrt{\tfrac{\log(1/\delta)}{2}\sum_{i=1}^m w_i^2}.$$

Since $n_{\text{eff}} = (\sum_i w_i^2)^{-1}$, the equivalent form follows. □

Let

$$\mathcal{F} := \{(q, x) \mapsto \ell(\theta; q, x) : \theta \in \Theta\}.$$

Theorem C.7 (task-specific weighted compression bound)**.** *Suppose Assumptions C.1, C.3, and C.4 hold. For any $\delta \in (0, 1)$, suppose further that the bucket-compression event described in Proposition C.2 holds with probability at least $1-\delta$. Conditional on the submitted queries, their bucket assignments, and the resulting weights $\{w_{j,s}\}$, distinct center-depth cells receive distinct seeds $\mathrm{SG}(h(c_j), s)$ and hence $\{(c_j, \tilde{x}_{j,s})\}$ are mutually independent with*

$$\tilde{x}_{j,s} \sim P_{\mathcal{X}}(\cdot \mid c_j).$$

Then, with probability at least $1-2\delta$,

$$\sup_{\theta\in\Theta} |L_{\mathrm{LADS}}(\theta) - \widetilde{\mathcal{E}}(\theta)| \le L_\ell(1+L_G)R + B\rho_\delta + 2\mathfrak{R}_{\mathbf{w}}(\mathcal{F}) + B\sqrt{\tfrac{\log(1/\delta)}{2n_{\text{eff}}}}.$$

Proof. Let $\mathcal{A}_{\text{comp}}$ denote the bucket-compression event described in Proposition C.2. By assumption, $\mathbb{P}(\mathcal{A}_{\text{comp}}) \ge 1-\delta$. On $\mathcal{A}_{\text{comp}}$, Lemma C.5 gives

$$\sup_{\theta\in\Theta} |L_{\mathrm{LADS}}(\theta) - \widetilde{L}(\theta)| \le L_\ell(1+L_G)R + B\rho_\delta,$$

where ρ_δ is the empirical bad-query fraction introduced in Section C.3.

Condition on the submitted queries, their bucket assignments, and hence on the weights $\{w_{j,s}\}$. Under this conditioning, the weights are fixed, and the compressed risk is a weighted empirical process over independent center-depth cells:

$$\widetilde{L}(\theta) = \sum_{j=1}^N \sum_{s=1}^T w_{j,s}\ell(\theta; c_j, \widetilde{x}_{j,s}),$$

and its weighted expectation is exactly

$$\widetilde{\mathcal{E}}(\theta) = \sum_{j=1}^N \sum_{s=1}^T w_{j,s}\mathbb{E}_{x\sim P_{\mathcal{X}}(\cdot|c_j)}[\ell(\theta; c_j, x)].$$

Applying Lemma C.6 to the center-depth cells gives

$$\sup_{\theta\in\Theta} |\widetilde{L}(\theta) - \widetilde{\mathcal{E}}(\theta)| \le 2\mathfrak{R}_{\mathbf{w}}(\mathcal{F}) + B\sqrt{\tfrac{\log(1/\delta)}{2n_{\text{eff}}}}.$$

The result follows by the triangle inequality and a union bound over $\mathcal{A}_{\text{comp}}$ and the weighted empirical-process concentration event.

□

C.6 Alignment with the Natural Conditional Population Risk

Theorem C.7 controls the empirical risk around the surrogate target $\widetilde{\mathcal{E}}$. We now align this surrogate target with the natural conditional population risk

$$\mathcal{E}(\theta) = \mathbb{E}_{q\sim P_{\mathcal{Q}_{\text{task}}}} \mathbb{E}_{x\sim P_{\mathcal{X}}(\cdot|q)}[\ell(\theta; q, x)].$$

The empirical quantity ρ_δ measures the fraction of collected queries that fail the center-and-hash assignment used above. To compare the surrogate target with the natural population risk, we use the corresponding population-level bad mass under the fixed private hash.

Assumption C.8 (Population bad mass under the fixed hash)**.** For the fixed private hash h, define the good set

$$\mathcal{G} := \{q \in \mathcal{Q}_{\text{task}} : \exists j \in [N] \text{ such that } d_{\mathcal{Q}}(q, c_j) \le R \text{ and } h(q) = h(c_j)\}.$$

Let

$$\rho := \Pr_{q\sim P_{\mathcal{Q}_{\text{task}}}}(q \notin \mathcal{G}).$$

We assume ρ is small.

For the concentration argument controlling empirical center masses below, we take the collected task queries $q_1, \dots, q_{KT}$ to be independently sampled from $P_{\mathcal{Q}_{\text{task}}}$. This condition is used only for the alignment fluctuation in Theorem C.9.

For every $q \in \mathcal{G}$, fix an assignment

$$J(q) \in [N]$$

such that

$$d_{\mathcal{Q}}(q, c_{J(q)}) \le R, \qquad h(q) = h(c_{J(q)}).$$

Define the population center mass

$$\mu_j := \Pr_{q\sim P_{\mathcal{Q}_{\text{task}}}}(q \in \mathcal{G},\ J(q) = j), \qquad j = 1, \dots, N.$$

Define the center-level population target

$$\mathcal{E}^{\text{ctr}}(\theta) := \textstyle\sum_{j=1}^{N} \mu_j \mathbb{E}_{x\sim P_{\mathcal{X}}(\cdot|c_j)}[\ell(\theta; c_j, x)].$$

Also define the depth-aggregated empirical weight

$$\bar{w}_j := \textstyle\sum_{s=1}^{T} w_{j,s}.$$

Then

$$\widetilde{\mathcal{E}}(\theta) = \textstyle\sum_{j=1}^{N} \bar{w}_j \mathbb{E}_{x\sim P_{\mathcal{X}}(\cdot|c_j)}[\ell(\theta; c_j, x)].$$

Theorem C.9 (Surrogate-to-natural population alignment)**.** *Suppose Assumptions C.1, C.3, C.4, and C.8 hold, and suppose the collected task queries are i.i.d. from $P_{\mathcal{Q}_{\text{task}}}$. Then:*

1. *For every $\theta \in \Theta$,*
$$|\mathcal{E}(\theta) - \mathcal{E}^{\text{ctr}}(\theta)| \le L_\ell(1 + L_G)R + B\rho.$$
2. *For any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*
$$\sup_{\theta\in\Theta} |\widetilde{\mathcal{E}}(\theta) - \mathcal{E}^{\text{ctr}}(\theta)| \le BN\sqrt{\tfrac{\log(2N/\delta)}{2KT}}.$$
3. *Consequently, with probability at least $1 - \delta$,*
$$\sup_{\theta\in\Theta} |\widetilde{\mathcal{E}}(\theta) - \mathcal{E}(\theta)| \le L_\ell(1 + L_G)R + B\rho + BN\sqrt{\tfrac{\log(2N/\delta)}{2KT}}.$$

Proof. First, compare the natural target with the center-level population target. Define

$$\psi(q, \theta) := \mathbb{E}_{x\sim P_{\mathcal{X}}(\cdot|q)}[\ell(\theta; q, x)].$$

For $q \in \mathcal{G}$, let $j = J(q)$. Couple $P_{\mathcal{X}}(\cdot \mid q)$ and $P_{\mathcal{X}}(\cdot \mid c_j)$ through the same latent noise ϵ:

$$G(q, \epsilon) \sim P_{\mathcal{X}}(\cdot \mid q), \qquad G(c_j, \epsilon) \sim P_{\mathcal{X}}(\cdot \mid c_j).$$

Using Assumptions C.3 and C.4,

$$|\ell(\theta; q, G(q,\epsilon)) - \ell(\theta; c_j, G(c_j,\epsilon))| \\ \le L_\ell\big(d_{\mathcal{Q}}(q,c_j) + d_{\mathcal{X}}(G(q,\epsilon), G(c_j,\epsilon))\big) \le L_\ell(1+L_G)R.$$

Taking expectation over ϵ gives

$$\big|\psi(q,\theta) - \mathbb{E}_{x\sim P_{\mathcal{X}}(\cdot|c_j)}[\ell(\theta; c_j, x)]\big| \le L_\ell(1+L_G)R.$$

Splitting $P_{\mathcal{Q}_{\text{task}}}$ over $\mathcal{G}$ and $\mathcal{G}^c$, and using $0 \le \ell \le B$ on $\mathcal{G}^c$, yields

$$|\mathcal{E}(\theta) - \mathcal{E}^{\text{ctr}}(\theta)| \le L_\ell(1+L_G)R + B\rho.$$

Second, define

$$\phi_j(\theta) := \mathbb{E}_{x\sim P_{\mathcal{X}}(\cdot|c_j)}[\ell(\theta; c_j, x)].$$

Then $0 \le \phi_j(\theta) \le B$, and

$$\widetilde{\mathcal{E}}(\theta) - \mathcal{E}^{\text{ctr}}(\theta) = \textstyle\sum_{j=1}^N (\bar{w}_j - \mu_j)\phi_j(\theta).$$

Therefore,

$$\sup_{\theta\in\Theta} |\widetilde{\mathcal{E}}(\theta) - \mathcal{E}^{\text{ctr}}(\theta)| \le B \textstyle\sum_{j=1}^N |\bar{w}_j - \mu_j|.$$

Under the i.i.d. query-sampling condition above, for each $j \in [N]$,

$$\bar{w}_j = \tfrac{1}{KT} \textstyle\sum_{m=1}^{KT} \mathbf{1}_{\{q_m\in\mathcal{G}, J(q_m)=j\}}$$

is an average of independent Bernoulli random variables with mean μ_j. Hence Hoeffding's inequality gives, for any $t > 0$,

$$\Pr\big(|\bar{w}_j - \mu_j| \ge t\big) \le 2\exp(-2KTt^2).$$

Taking

$$t = \sqrt{\tfrac{\log(2N/\delta)}{2KT}}$$

and applying a union bound over $j = 1, \dots, N$, we obtain that, with probability at least $1-\delta$,

$$|\bar{w}_j - \mu_j| \le \sqrt{\tfrac{\log(2N/\delta)}{2KT}}, \qquad j = 1, \dots, N.$$

Hence

$$\sup_{\theta\in\Theta} |\widetilde{\mathcal{E}}(\theta) - \mathcal{E}^{\text{ctr}}(\theta)| \le BN\sqrt{\tfrac{\log(2N/\delta)}{2KT}}.$$

The third statement follows by the triangle inequality. □

C.7 Conditional LADS Bound

Combining Theorem C.7 and Theorem C.9 gives a direct bound between the empirical distillation risk and the natural conditional population risk.

Theorem C.10 (Conditional LADS bound)**.** *Under the assumptions of Theorems C.7 and C.9, for any $\delta \in (0,1)$, with probability at least $1-3\delta$,*

$$\sup_{\theta\in\Theta} |L_{\text{LADS}}(\theta) - \mathcal{E}(\theta)| \le 2L_\ell(1+L_G)R + B\rho_\delta + B\rho \\ + 2\mathfrak{R}_{\mathbf{w}}(\mathcal{F}) + B\sqrt{\tfrac{\log(1/\delta)}{2n_{\text{eff}}}} + BN\sqrt{\tfrac{\log(2N/\delta)}{2KT}}.$$

Proof. For any $\theta \in \Theta$,

$$|L_{\text{LADS}}(\theta) - \mathcal{E}(\theta)| \le |L_{\text{LADS}}(\theta) - \widetilde{\mathcal{E}}(\theta)| + |\widetilde{\mathcal{E}}(\theta) - \mathcal{E}(\theta)|.$$

Take the supremum over $\theta \in \Theta$, apply Theorem C.7 to the first term and Theorem C.9 to the second term, and combine the events by a union bound. □

C.8 Derivation of Theorem 3.2

We now connect the preceding general bound to the statement in Theorem 3.2. Proposition C.2 first establishes the semantic bucket-compression event. Conditional on this bucket structure, the remaining probabilistic steps are the weighted empirical-process fluctuation and the surrogate-to-natural alignment fluctuation. Accordingly, the confidence level $1 - 2\delta$ in Theorem 3.2 accounts for these two concentration events, while the success probability of the bucket-compression event is quantified separately in Proposition C.2.

The general analysis above allows for imperfect center-and-hash assignments through the finite-sample bad-query fraction ρ_δ and the population bad mass ρ. Theorem 3.2 corresponds to the clean-assignment specialization of this general result. In particular, we restrict to the regime in which

$$\rho_\delta = 0, \qquad \rho = 0.$$

Hence the auxiliary terms $B\rho_\delta$ and $B\rho$ appearing in the general bound of Theorem C.10 vanish.

For the center-depth empirical process, recall that the general effective sample size is

$$n_{\text{eff}} = \left(\sum_{j=1}^{N} \sum_{s=1}^{T} w_{j,s}^2 \right)^{-1}.$$

Theorem 3.2 further corresponds to the balanced specialization of the preceding weighted bound, in which all NT active center-depth cells carry equal empirical weight:

$$w_{j,s} = \frac{1}{NT}, \qquad j \in [N],\ s \in [T].$$

Since

$$\sum_{j=1}^{N} \sum_{s=1}^{T} w_{j,s} = 1,$$

this balanced regime is consistent with the clean-sample condition $\rho_\delta = 0$. Moreover,

$$n_{\text{eff}} = \left(NT \cdot \frac{1}{(NT)^2} \right)^{-1} = NT,$$

and the corresponding weighted Rademacher complexity reduces to the ordinary NT-sample Rademacher complexity:

$$\mathfrak{R}_w(\mathcal{F}) = \mathfrak{R}_{NT}(\mathcal{F}).$$

Therefore, the empirical-process contribution becomes

$$2\mathfrak{R}_{NT}(\mathcal{F}) + B\sqrt{\frac{\log(1/\delta)}{2NT}}.$$

Under the same clean-assignment specialization, the local replacement and population-alignment arguments contribute

$$2L_\ell(1 + L_G)R,$$

together with the center-mass fluctuation

$$\Delta_{\text{align}} := BN\sqrt{\frac{\log(2N/\delta)}{2KT}}.$$

Combining these terms yields

$$\sup_{\theta \in \Theta} |L_{\text{LADS}}(\theta) - \mathcal{E}(\theta)| \le 2\mathfrak{R}_{NT}(\mathcal{F}) + B\sqrt{\frac{\log(1/\delta)}{2NT}} + 2L_\ell(1 + L_G)R + \Delta_{\text{align}},$$

which is exactly the bound stated in Theorem 3.2.

C.9 Reduction to the Unconditional Case

The unconditional setting is recovered as a special case. When the query is absent, or all queries are identical, there is only one task-specific semantic bucket:

$$N = 1.$$

Moreover, there is no query-to-center replacement error, so

$$R = 0.$$

The compressed process then has effective sample size approximately

$$n_{\text{eff}} \asymp T,$$

and the dominant estimation term reduces to

$$2\mathfrak{R}_T(\mathcal{F}) + B\sqrt{\tfrac{\log(1/\delta)}{2T}}.$$

Thus the conditional theory recovers the same effective sample-size collapse as LADS-SIMPLE: although the distiller collects KT outputs, the cross-account coupling leaves only T independent latent trajectories in the unconditional case. This matches the $\mathcal{O}(1/\sqrt{T})$ rate in Proposition 3.1.