machine learning & knowledge extraction

MDPI

Article

Metrics and a Single-Pass Framework for Disentangling Feature-Level Privacy Risk in Machine Learning Models

Daniel S. Soper

Information Systems & Decision Sciences Department, California State University, Fullerton, CA 92831, USA; dsoper@fullerton.edu

Abstract

This paper introduces a diagnostic pipeline for evaluating feature-level privacy risk in machine learning (ML) models, anchored by the Coordinate Dependency Index (CDI)—a unitless, cross-modality metric that is designed to quantify the relative degree of coordinate dependence that persists following an ML model's compression process. Grounded in the Information Bottleneck principle, the CDI establishes an exact algebraic threshold—the Invariance Horizon at $e^{-1} \approx 0.37$—to detect coordinate dependence across disparate data types. To facilitate direct cross-modality comparisons of the underlying leakage, we integrate Absolute Information Gain (AIG) as a complementary metric that maps these threshold exceedances into absolute bits. To operationalize this dual-metric approach, we develop the Representational Interrogator (RI), an inverse multi-task learning architecture that resolves the long-standing Audit Scalability Gap in ML privacy auditing. Experimental results across 30 trials demonstrate that the RI framework maintains strong initialization stability and achieves single-pass efficiency by collapsing traditional, sequential privacy auditing batteries into a unified training objective. Achieving speedups of up to 91× over generative adversarial batteries while maintaining strong diagnostic rank-order alignment, the framework's robustness across deep neural networks and gradient-boosted decision trees validates the RI-CDI-AIG pipeline as a principled, architecture-independent methodology for disentangling and ranking informational vulnerabilities in ML models.

Keywords: ML privacy auditing; coordinate dependency index; absolute information gain; representational interrogator; information bottleneck; attribute inference; multi-task learning

Academic Editor: Peter Kieseberg

Received: 28 August 2026
Revised: 10 October 2026
Accepted: 23 October 2026
Published: 28 October 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The use of machine learning (ML) models in sensitive domains such as healthcare, finance, human resources, and government has created an urgent and critical need for rigorous privacy auditing of the data that those models produce [1]. The reason for this is that many ML models are trained on sensitive and legally protected data, which makes them targets for sophisticated attacks that use the models' public-facing output to reverse-engineer the sensitive data on which the models were trained [2]. This "leakage" of sensitive, private information into the hands of a malicious actor represents both an ethical failure and a significant legal and financial liability under regulatory frameworks such as the European Union's General Data Protection Regulation (GDPR) and the California

Consumer Privacy Act (CCPA) [3,4]. Systematic privacy auditing of ML models must therefore be seen not only as a technical process for protecting sensitive data, but also as a vital defensive activity that shields an organization from the reputational damage and existential legal ramifications that often follow a successful privacy breach.

1.1. The Dual Challenges of Privacy Auditing

Despite the critical importance of ML privacy auditing, two fundamental challenges currently impede organizations' efforts to perform systematic privacy auditing on their machine learning models:

1. The Measurement Challenge: A primary obstacle to effective ML privacy auditing is the absence of a unitless, model-agnostic metric that can allow information leakage across diverse types of input features to be detected, quantified, and compared. Instead, privacy risk is currently measured using disparate, modality-specific performance indicators that rely on fundamentally different mathematical foundations. For categorical features, privacy risk is typically quantified through reconstruction accuracy, representing the probability that an attacker can correctly guess the value of a discrete attribute, such as a person's race or gender [5–7]. Simple accuracy, however, is often misleading because it fails to account for class imbalances or the probability of an attacker guessing correctly by chance. More rigorous evaluations may therefore rely on Cohen's Kappa (κ), which provides a chance-corrected measure of agreement between the true and reconstructed values [8].
   By contrast, privacy risk for continuous features is typically measured using error-based metrics such as Mean Squared Error (MSE) or Mean Absolute Error (MAE) [7,9]. While common, these metrics lack interpretability because they do not account for the feature's underlying distribution. More informative assessments may therefore utilize the Coefficient of Determination (R^2) for continuous features, which quantifies the proportion of feature variance that was reconstructed by the attacker relative to a naïve, mean-guess baseline. Since these metrics utilize different scales and units, it is currently impossible for an auditor to determine whether, for example, a high κ on a categorical feature represents a greater privacy threat than a high R^2 value on a continuous feature. Put differently, these legacy metrics lack the mathematical commonality that is required to prioritize vulnerabilities in a heterogeneous dataset. To bridge this gap, this paper introduces a dual-metric approach. First, the unitless, cross-modality formulation of the Coordinate Dependency Index (CDI) provides a normalized threshold to detect which features exhibit persistent coordinate dependence. Second, to enable the direct ranking of these vulnerabilities across disparate data types, we apply Absolute Information Gain (AIG) to map the underlying reconstruction loss into absolute bits. Together, this two-step pipeline transforms ML privacy auditing from a heuristic-based exercise into a principled, mathematically grounded methodology for information disentanglement.
2. The Engineering Challenge: Beyond the lack of cross-modality metrics, a significant practical and computational bottleneck—which we refer to as the "Audit Scalability Gap"—impedes the systematic assessment of ML privacy risk. This gap is characterized by the inability of current privacy auditing methods to provide a scalable, diagnostic risk profile of all input features on a feature-by-feature basis. While Membership Inference Attacks (MIAs) can provide a coarse, instance-level "member or non-member" signal, they cannot identify *which* specific features are responsible for information leakage [2]. Conversely, Attribute Inference Attacks (AIAs) can measure feature-level risk but are inherently limited by their targeted nature [5,6]. To perform a comprehensive audit of an ML model trained on N input features, a data scientist must currently plan, implement, and execute N separate, bespoke, and

computationally expensive attribute inference attacks [10]. Even if modern GPU clusters allow the execution of these baseline attacks to be batched or parallelized, the approach creates a prohibitive operational burden. As N increases, the human engineering time required to manually design, configure, and manage these N independent adversarial architectures scales linearly, making it practically impossible to audit modern ML models with high-dimensional feature spaces in a timely manner [11,12]. As such, organizations remain effectively "blind" to their models' full privacy risk profiles, insofar as they lack a method to efficiently disentangle and rank the vulnerabilities of all N features simultaneously.

1.2. Proposed Solutions: A Dual-Metric Pipeline and the Representational Interrogator

To resolve the limitations of existing methods, this paper develops a theoretically grounded framework for detecting and measuring the persistence of coordinate information in an ML model's internal representation. As used here, "coordinate information" refers to the specific, high-resolution identifiers and individual data points that collectively constitute an ML model's training set. We first propose the Coordinate Dependency Index (CDI), an adaptive, normalized metric that quantifies the degree of coordinate dependence that is present within a machine learning model's internal representation. By normalizing for a feature's specific modality—utilizing variance for continuous features and Shannon entropy for categorical features—the CDI provides a unitless, cross-modality measure that maps the extent to which an ML model depends on a feature's coordinate data onto a unit interval. Beyond its utility in privacy auditing, the CDI serves as a fundamental measure of representational maturity, quantifying the extent to which an ML model has successfully navigated the Information Bottleneck to transform raw data points (i.e., coordinate information) into abstract knowledge. A central property of this formulation is the establishment of an Invariance Horizon at $e^{-1} \approx 0.37$, which functions as an exact algebraic consequence of the naïve baseline. Values exceeding this threshold provide a robust signal that a feature's coordinate information has failed to be optimally distilled and remains present in the model's output as reconstructible raw information. To complement this normalized detection threshold, we integrate Absolute Information Gain (AIG) into the pipeline as a secondary metric. By mapping the underlying reconstruction loss into absolute bits using standard information-theoretic capacity formulas, the AIG allows practitioners to directly stack-rank the magnitude of leaked information across features with heterogeneous data types.

To operationalize this dual-metric pipeline at scale, we also introduce the Representational Interrogator (RI) framework, a novel diagnostic method that repurposes the victim model's own internal representation as a probe for privacy risk. The RI method employs a single, inverse multi-task learning (MTL) architecture that, in a single pass, learns to map a victim model's public-facing output back to the very prototypes that the model was designed to preserve. This specific architecture is shown to be the key to solving the Audit Scalability Gap, since it interrogates the coordinate dependence within the learning process to generate a holistic, ranked vulnerability list for every input feature simultaneously, which is then mapped to absolute bits via the AIG to establish a unified, cross-modality ranking. By providing a complete privacy risk profile with single-pass efficiency, the RI framework overcomes the prohibitive bottleneck of existing privacy auditing methods that require a battery of N separate attacks, thus enabling organizations to identify and protect their most vulnerable data features with unprecedented efficiency. A comparison of the current state-of-the-art, feature-level ML privacy auditing method and the proposed RI-CDI-AIG pipeline is provided in Table 1 below.

Table 1. Comparison of Feature-Level Privacy Auditing Methods.

Attribute	Feature-Level ML Privacy Auditing Method	
	AIA Battery (Current State-of-the-Art)	**Proposed RI-CDI-AIG Framework**
Computational Efficiency	Sequential execution—requires N independent training passes. Wall-clock time scales linearly with the number of input features.	Single-pass efficiency—generates a complete risk profile for all N features in a unified training pass.
Audit Requirement	N separate, bespoke, and expensive attacks	Single multi-task learning (MTL) inversion model
Metric Type	Disparate and incommensurate (accuracy, MSE, MAE, etc.)	Dual-metric pipeline—CDI for cross-modality detection and AIG for absolute bitrate ranking.
Risk Interpretation	Ambiguous. Does not account for population priors or uncertainty	Threshold-based. $CDI > e^{-1} \approx 0.37$ flags features with high coordinate dependence

The balance of this paper is organized as follows: Section 2 establishes the background and theoretical foundations for the RI-CDI-AIG framework. Section 3 describes the Representational Interrogator algorithm and the design of the evaluative experiments that were used to compare the RI method's performance against a variety of baseline AIA-battery methods. The results of these experiments are presented and discussed in Section 4. The paper concludes in Section 5 with a summary of the paper's contributions, a description of the study's limitations, and suggestions for future research in this area.

2. Background and Theoretical Foundations

The proposed Representational Interrogator (RI) framework and its dual-metric diagnostic pipeline (comprising the CDI and AIG) are situated at the critical intersection of adversarial privacy attacks, information theory, and multi-task learning. To understand this synthesis, we must first recognize that privacy attacks and their corresponding auditing methodologies are typically classified by two dimensions: (1) their granularity (instance-level vs. feature-level) and (2) their adversarial goal (inference vs. reconstruction). By reframing model inversion not as a failure of compression, but as a diagnostic probe of a model's internal representations, we can resolve the long-standing tensions between predictive utility and data confidentiality while providing a standardized methodology to detect and rank informational persistence across diverse ML sub-fields.

2.1. A Taxonomy of Privacy Auditing Methods

The systematic auditing of ML models requires a clear understanding of the existing adversarial landscape. This section examines the spectrum of inference-based auditing methods, ranging from the coarse-grained signals of instance-level membership inference attacks to the granular, feature-level assessments that are provided by attribute inference attacks. While both represent mature fields of study, it is the latter category that establishes the comparative baselines for this study and illustrates the critical Audit Scalability Gap that the Representational Interrogator is designed to resolve.

2.1.1. Instance-Level Auditing: Membership Inference Attacks (MIAs)

Privacy auditing methods for machine learning models are typically categorized by their target granularity, with membership inference attacks (MIAs) representing the coarse-grained, instance-level baseline. The primary goal of an MIA is to determine whether a specific data record (x, y) was a member of a model's training set, often by exploiting the observation that models tend to be more confident in their predictions for "member" data than for "non-member" data [2]. While MIAs are a mature and successful

field of research, they remain a holistic tool from the perspective of privacy auditing because they provide only a single bit of information (membership status) and offer no diagnostic insight into which specific features are most vulnerable to attack or how those features might be reconstructed [6]. As such, MIAs are insufficient for the granular, feature-by-feature risk profiling that is required by modern organizations, necessitating a shift toward metrics like the proposed dual-metric pipeline that can quantify the coordinate dependence of individual input features.

2.1.2. Feature-Level Auditing: Attribute Inference Attacks (AIAs)

In comparison to MIAs, AIAs are more granular and represent the baseline method for feature-level privacy auditing of ML models [5,6]. The adversary's goal when using an AIA is to infer the value of a single, unknown, sensitive attribute x_i (e.g., a person's race or gender) given knowledge of the remaining "public" attributes $\{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N\}$ and the model's output y. To ensure a rigorous evaluation of the proposed Representational Interrogator (RI) framework, this study utilizes AIAs as the comparative basis against which the speed and quality of the RI are judged. AIAs serve as the baseline for feature-level risk assessment because they constitute the only currently available empirical tool that can isolate and measure the information leakage of each individual input feature [6,10]. To provide a scientifically robust assessment, the RI framework is compared against two distinct classes of state-of-the-art (SOTA) AIAs—one discriminative and one generative—that represent the highest levels of adversarial capability currently established in the ML privacy literature. Detailed architectural specifications for these baseline models are provided in Section 3.2.4.

2.2. Holistic Reconstruction and the Inversion Framework

While the inference-based attacks discussed above serve as essential comparative baselines, the direct architectural inspiration for the benign Representational Interrogator method is found in the field of model inversion attacks. In a traditional model inversion attack, an adversary's objective is to reconstruct a representation of the original input X using only the victim model's public-facing output Y [7,13,14]. These attacks are typically categorized into two methodological tracks:

1. Optimization-Based Attacks: These inversion attacks involve iteratively searching for an input X that maximizes the likelihood of a specific output Y.
2. Generative Attacks: These inversion attacks utilize advanced architectures such as generative adversarial networks (GANs) or variational autoencoders (VAEs) to learn the direct mapping from $Y \rightarrow X$. This generative track is particularly significant as it informs the MTL structure of the Representational Interrogator and its ability to extract the CDI at scale.

Crucially, modern research into model inversion has revealed that these attacks do not—and theoretically cannot—reconstruct exact, unique training instances from a model's input. Instead, they exploit the coordinate dependence that survives the compression process—the prototypical average of the input X conditioned on a given class or output signal Y [15,16]. The present paper builds upon this empirical observation by identifying it as the critical link for resolving the "Model Inversion Paradox." We argue that while instance-specific identifiers are purged during training, an ML model must preserve these abstract prototypes in order to maintain its predictive utility. By repurposing this adversarial mechanism as a diagnostic probe, the Representational Interrogator framework disentangles these prototypes to quantify the Coordinate Dependency Index and calculate the Absolute Information Gain for each input feature simultaneously.

2.3. The Theoretical Impetus: Reconciling Inversion and Information Theory

The theoretical viability of any inversion-based auditing framework, including the proposed Representational Interrogator, must first confront a fundamental contradiction between the design goals of ML models and the practical success of adversarial reconstruction. This conflict is referred to here as the Model Inversion Paradox.

2.3.1. The Information Bottleneck (IB) Critique

The Information Bottleneck (IB) principle provides a powerful theoretical framework for understanding deep learning not as mere function-fitting, but rather as a process of optimal, lossy compression [17,18]. Although critiques such as those by Saxe et al. [19] have demonstrated that strict compression is not universally required for generalization under certain conditions, the IB principle provides a highly effective framework for understanding representation limits. According to the IB principle, an optimal model f learns an internal representation Z (and by extension, a final output Y) by solving a constrained optimization problem, commonly expressed as a Lagrangian objective:

$$L_{IB} = \min_{p(z|x)} [I(X;Z) - \beta I(Z;Y)] \quad (1)$$

In this formulation, $I(X;Z)$ represents the mutual information between the input data X and the representation Z, while $I(Z;Y)$ represents the mutual information between the representation Z and the target variable Y. The Lagrange multiplier β controls the trade-off between two competing goals:

1. Relevance: The representation must be maximally informative about the target variable Y (i.e., $\max I(Z;Y)$).
2. Compression: The representation must be minimally informative about the original input X (i.e., $\min I(X;Z)$).

Since neural networks act as successive stages of this lossy compression process, we must acknowledge that an optimal model is explicitly designed to be a "bad inverter." It aggressively purges every bit of information from X that is not essential for predicting Y, thus presenting a direct theoretical challenge to any auditor that attempts to reconstruct X from Y.

2.3.2. The Model Inversion Paradox

What this paper defines as the Model Inversion Paradox emerges from the direct contradiction between the IB theory and the observed behavior of modern ML privacy attacks:

1. The Information Bottleneck Proposition: Theoretically, model f should destroy all unnecessary information about X, making the output Y an information-poor, non-invertible signal.
2. The Empirical Reality: Practically, model inversion attacks are a mature and successful field of AI security research, capable of reconstructing sensitive data from model outputs.

2.3.3. Resolution to the Model Inversion Paradox: Reconstructing Prototypes, Not Instances

The resolution to the Model Inversion Paradox, which forms the theoretical foundation for the Coordinate Dependency Index, is found not by challenging the Information Bottleneck's premise of compression, but rather by refining our understanding of the fundamental nature of the information that survives IB compression. While the IB critique correctly identifies that instance-specific "noise"—the unique, stochastic variations of a specific training point—is theoretically purged during the training process to ensure model generalizability, it fails to account for the coordinate dependence that a model may preserve in order to remain predictive. What model inversion attacks actually recover is

not a literal duplicate of a training instance, but instead is a high-fidelity reconstruction of the prototypical average of the input X conditioned on the class or output signal Y, which allows this process to be more accurately described as distributional inference than strict attribute inference. Although a traditional Attribute Inference Attack attempts to target a specific instance by conditioning on known public features (X_{known}) alongside the output (Y), it is mechanically forced to reconstruct a highly localized prototype. The AIA can only succeed when the targeted instance is mathematically aligned with that constrained distributional sub-space. By contrast, the Representational Interrogator conditions solely on Y to extract the global class prototype directly. This distinction reveals that predictive utility and informational persistence are mathematically coupled and must be seen as "two sides of the same coin" [20]. Put differently, an ML model's utility is derived precisely from its ability to distill raw data into these abstract, robust correlations, and this learned correlation *is* the prototype.

Crucially, this implies that model inversion does not "reverse" or "undo" the IB compression. Instead, model inversion interrogates the very prototypes that the bottleneck was designed to preserve. In this framework, the prototype serves as the structural repository for the specific identifiers and data points that constitute the training set. While the Coordinate Dependency Index (CDI) serves as a robust, normalized indicator of whether this raw coordinate information remains embedded within that prototype, the Absolute Information Gain (AIG) quantifies exactly how much information has leaked in bits. By measuring the extent to which a model has succeeded in destroying instance-specific identifiers while retaining the abstract prototypes that are required for inference, the CDI identifies the precise threshold where informational persistence becomes an adversarial vulnerability, and the AIG ranks the severity of the breach. The Representational Interrogator exploits this mathematical necessity, transforming the model's own internal representation into a dual-metric diagnostic probe that resolves the Audit Scalability Gap with single-pass efficiency.

2.4. The Representational Interrogator (RI) Framework

The resolution described above to the Model Inversion Paradox directly informs this paper's Representational Interrogator framework. By reframing inversion as the interrogation of an ML model's coordinate dependence, we can operationalize the ML privacy auditing process through the lens of architectural disentanglement. To achieve this, the RI framework utilizes a specific shared stem architecture in which a common encoder (h_θ) distills the victim model's output Y into a shared latent representation z, which effectively captures the predictive essence of the original training data. This representation then feeds into N independent decoder heads $(g_{\phi_1}, \ldots, g_{\phi_N})$, each of which is task-optimized to reconstruct a specific feature prototype $\hat{X}_i$. A conceptual representation of the RI framework is provided in Figure 1 below.

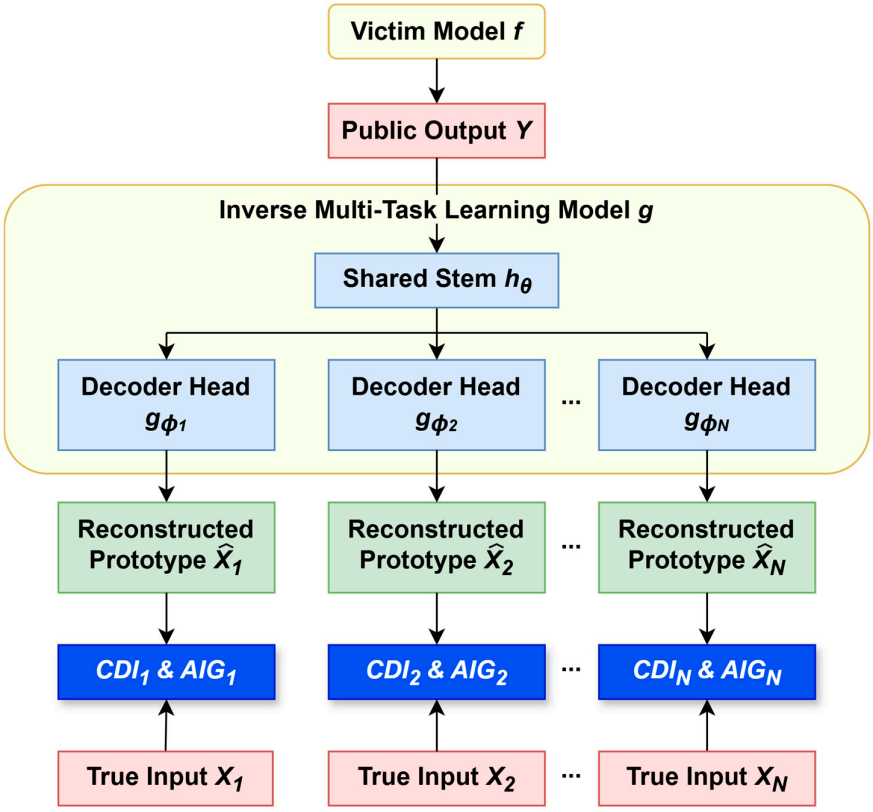

Figure 1. The Representational Interrogator (RI) architecture. The process begins with the victim model f generating public output Y. This output is interrogated by the inverse MTL model g, which processes Y through a shared stem h_θ. This stem distills the victim model's output into a shared latent representation—the compressed state of the original training data—which feeds into N parallel decoder heads $g_{\phi_1}, \ldots, g_{\phi_N}$. Each head is task-optimized to disentangle the coordinate dependence within the learning process to generate a specific reconstructed prototype $\hat{X}_i$. Finally, these prototypes are compared against the true inputs X_i to calculate the Coordinate Dependency Index (CDI), which quantifies the degree of coordinate dependence that persists within the model's representation relative to a naïve, prior-ignorance baseline, while the AIG maps the underlying reconstruction loss into absolute bits to facilitate direct cross-modality ranking.

2.5. The Coordinate Dependency Index (CDI) and Absolute Information Gain (AIG)

The proposed Coordinate Dependency Index (CDI) is formally defined as the uncertainty-normalized exponential decay of the final reconstruction loss $\mathcal{L}_i$. To ensure that this metric remains robust and comparable across diverse data types, the framework employs an adaptive definition for the normalization term U_i:

$$CDI_i = e^{-\mathcal{L}_i/U_i}, \quad (2)$$

- For Continuous Features: $\mathcal{L}_i$ is the mean squared error (MSE) and U_i is the feature variance σ_i^2. While variance is utilized here for continuous features due to its computational efficiency and direct alignment with MSE-based auditing, the CDI is functionally modular. In cases of highly non-Gaussian distributions, differential entropy can be used in place of variance for U_i to ensure that the Invariance Horizon remains a strictly relevant anchor for prior ignorance.
- For Categorical Features: U_i is the Shannon entropy (H_i) of the feature's prior distribution. To ensure mathematical calibration when calculating the final CDI ratio, the numerator $\mathcal{L}_i$ is the standard, unweighted cross-entropy loss.

- Note: Although standard cross-entropy is used for the final CDI calculation, focal loss is used to optimize the gradients during the multi-task training phase in order to prevent majority-class collapse [21].

By utilizing this exponential decay function, the CDI provides a unitless, model-agnostic metric that is mathematically constrained to the unit interval $[0, 1]$. This normalization is critical because it establishes a standardized threshold for detecting coordinate persistence across all input features, regardless of their modality (i.e., categorical or continuous). Put differently, the CDI effectively measures the relative magnitude of coordinate dependence for a specific feature compared to its own naïve baseline. The index is characterized by two theoretical extremes that define the spectrum of coordinate persistence:

- Maximum Persistence ($CDI \approx 1.0$): As the reconstruction loss $\mathcal{L}_i$ approaches zero, the CDI approaches 1.0. In a privacy context, this represents "complete leakage," indicating that the Representational Interrogator has perfectly reconstructed the specific feature instance. This state confirms that the reconstructed prototype is a literal duplicate of the training coordinate, signifying that the victim model retains an absolute and uncompressed dependency on its original input data.
- Minimum Persistence ($CDI \to 0.0$): As the reconstruction loss becomes significantly larger than the natural uncertainty ($\mathcal{L}_i \gg U_i$), the CDI asymptotically approaches 0.0. In a privacy context, this represents "no leakage," indicating that the reconstructed prototype is entirely uninformative—equivalent to pure noise—and that the feature has reached a state of maximum coordinate independence. The victim model has successfully fulfilled its compression objective by destroying the specific identifiers that are required to anchor the prototype back to its training coordinates.

2.5.1. The Invariance Horizon (e^{-1})

Crucially, the CDI formulation provides a mathematically defined baseline at $CDI = e^{-1}$, which we define as the Invariance Horizon. This boundary is derived from the "naïve adversary" scenario in which an attacker possesses no specific information about an instance and simply guesses the population mean or prior probability. In such cases, the reconstruction loss $\mathcal{L}_i$ converges to the feature's natural uncertainty U_i. Consequently, the ratio $\mathcal{L}_i/U_i$ becomes 1.0, yielding a CDI of $e^{-1} \approx 0.37$. The mathematical properties of the CDI formulation reveal a fundamental asymmetry in the interpretation of representational state:

1. A Low Coordinate Dependency Index ($CDI \leq e^{-1}$) falls within the Distillation Zone: A CDI that is less than or equal to the e^{-1} boundary indicates that the reconstruction loss meets or exceeds the input feature's natural uncertainty (i.e., $\mathcal{L}_i \geq U_i$), and that the reconstructed prototype is hence dominated by error or noise. Within our theoretical framework, this outcome represents successful distillation or optimal purging under the Information Bottleneck principle. This state is characterized by a convergence toward prior ignorance, which can arise from two distinct representational states:
    a. Discarding (Optimal Purging): The input feature X_i was irrelevant to predicting the target Y, leading the model f to completely purge the feature during the learning process. In this state, no coordinate dependence exists to be interrogated.
    b. Transformation (Irreversible Distillation): The input feature X_i was critical to predicting Y, but the model f optimally distilled the feature into a highly compressed, non-invertible representation. While the information contributes to the

model's predictive utility, its specific coordinate identifiers have been destroyed.

In the context of ML privacy auditing, the feature remains effectively private and safe from adversarial reconstruction in either of these cases.

2. A High Coordinate Dependency Index ($CDI > e^{-1}$) falls within the Persistence Zone: This is the key diagnostic signal. A high CDI indicates that the reconstructed prototype was recovered with a degree of fidelity that exceeds the naïve, prior-ignorance baseline (i.e., $\mathcal{L}_i < U_i$). This result provides unambiguous evidence of coordinate dependence, confirming that the individual data points and identifiers for input feature X_i were neither discarded nor irreversibly distilled, but rather remain embedded within the model's internal representation. In the context of ML privacy auditing, a high CDI thus confirms that the coupling between predictive utility and coordinate dependence has skewed toward the latter, thereby flagging the associated input feature as unsafe and susceptible to a privacy attack. Such persistent coordinate dependence indicates a failure of topological separability, where the model has failed to decouple its logical utility from its specific training coordinates.

The theoretical properties of the Coordinate Dependency Index are depicted in Figure 2 below.

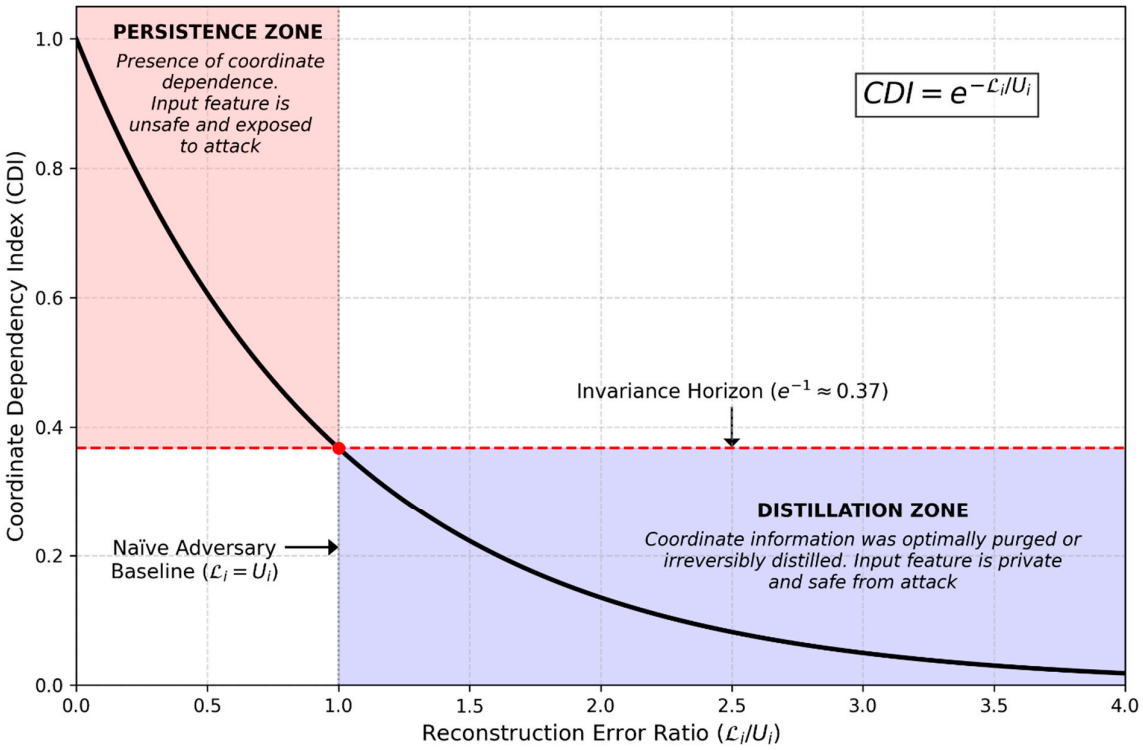

Figure 2. Theoretical Properties of the Coordinate Dependency Index (CDI). The curve illustrates the exponential decay relationship between the reconstruction error ratio ($\mathcal{L}_i/U_i$) and the resulting CDI, quantifying the degree of the input feature's coordinate dependence within an ML model's internal representation. The red dashed line marks the Invariance Horizon at $e^{-1} \approx 0.37$, an exact algebraic threshold defined by the naïve adversary baseline. Scores within the "Persistence Zone" (Red) represent a failure of the Information Bottleneck to destroy sensitive identifiers, signaling that the corresponding input feature's coordinate information remains present within the model's internal representation. Conversely, scores within the "Distillation Zone" (Blue) represent the successful fulfillment of the ML model's compression objective. In this zone, the input feature is effectively

private, either because its coordinate noise has been optimally purged or because the coordinate information has been irreversibly distilled into a non-invertible state of predictive utility.

Ultimately, the Coordinate Dependency Index establishes a rigorous, quantifiable resolution to the fundamental uncertainty of ML privacy auditing. By mapping the mathematical boundary between a model's predictive utility and its adversarial vulnerability, the CDI provides a reliable means of distinguishing between features that have been safely distilled into abstract patterns and those with persistent coordinate dependence. Although the CDI effectively flags *which* features have breached the Invariance Horizon, a comprehensive risk profile requires the ability to quantify and compare the severity of the breaches among the set of unsafe input features. To achieve this, the framework transitions from detection to ranking via the Absolute Information Gain.

2.5.2. Absolute Information Gain (AIG)

While the CDI provides a normalized threshold for detecting coordinate persistence, practitioners require a complementary metric to quantify and compare the magnitude of leaked information across input features of different modalities. To achieve this, we integrate Absolute Information Gain (AIG) to map the reconstruction loss into absolute bits. For continuous features, where U_i represents the baseline variance (σ^2) and $\mathcal{L}_i$ represents the mean squared error (MSE), AIG is derived using the standard Shannon capacity formulation for a continuous Gaussian channel:

$$AIG_i = \max\left(0, \frac{1}{2}\log_2\left(\frac{U_i}{\mathcal{L}_i}\right)\right) \tag{3}$$

For categorical features, including those represented as high-dimensional one-hot encoded vectors, AIG calculates the reduction in uncertainty across the discrete probability space. In this formulation, U_i represents the prior Shannon entropy, $\mathcal{L}_i$ represents the unweighted cross-entropy loss, and AIG is mapped to bits as:

$$AIG_i = \max\left(0, \frac{U_i - \mathcal{L}_i}{\ln(2)}\right) \tag{4}$$

By utilizing the CDI to identify which features exceed the Invariance Horizon and then using the AIG to quantify the precise informational payload of that breach in absolute bits, the framework yields a comprehensive diagnostic profile that naturally bridges continuous and discrete feature spaces.

3. Materials and Methods

Following the theoretical resolution to the Model Inversion Paradox in Section 2, this section provides a formal description of the Representational Interrogator algorithm and its role in quantifying the CDI and AIG. It then describes the design of the evaluative experiments that were used to assess the RI framework's performance against two state-of-the-art baseline AIA methods, with a focus on validating the RI's computational efficiency and the CDI's diagnostic accuracy across a diverse set of benchmark datasets.

3.1. The Representational Interrogator Algorithm

The RI method is designed to be a model-agnostic diagnostic instrument that, in this paper, is used to interrogate a victim model's coordinate dependence in order to quantify privacy risk across all input features simultaneously. The steps of the algorithm, detailed in Algorithm 1 below, represent the practical realization of the architectural disentanglement and uncertainty-normalized quantification discussed in Section 2.

Algorithm 1. The Representational Interrogator Framework.

Input:

- A pre-trained "victim" model $f: \mathbb{R}^N \rightarrow \mathbb{R}^K$, which maps an N-dimensional input feature vector X to a K-dimensional output vector $Y = f(X)$ (where $K \geq 1$).
  - Note: X is the set of N input features $\{X_1, \ldots, X_N\}$, where each X_i may be a vector (e.g., in the case of one-hot encoding).
- A reference dataset $D = \{x, y\}$, typically derived from the victim model's original training or validation set.

Step 1: Establish Uncertainty Baselines (U_i)

- For every input feature X_i in dataset D, calculate the feature's statistical uncertainty U_i according to the modality-specific formulations (variance or Shannon entropy) established in Section 2.5. This baseline serves as the prior ignorance anchor for the CDI and AIG calculations.

Step 2: Construct the Representational Interrogator (g)

- Define an inverted "auditor" model $g: \mathbb{R}^K \rightarrow \hat{\mathcal{X}}$ (where $\hat{\mathcal{X}}$ represents the reconstructed feature space) as a multi-task learning architecture designed to probe the victim model's latent state via a mirror architecture comprising two components:
  1. The Shared Stem (h_θ): A neural network that distills the K-dimensional output vector Y into a shared latent representation z, effectively acting as the "reflective surface" that captures the predictive essence of the victim model.
  2. N Independent Decoder Heads $\left(g_{\phi_1}, \ldots, g_{\phi_N}\right)$: A battery of parallel, lightweight networks, each of which is task-optimized to disentangle and reconstruct the prototypical value $(\hat{\mathcal{X}}_i)$ of a specific input feature from the shared representation z.

Step 3: Define Diagnostic Loss Functions ($\mathcal{L}_i$)

- Assign the task-specific loss functions established in Section 2.5 to each decoder head g_{ϕ_i}. Categorical decoder heads are optimized via focal loss during training to prevent the auditor from converging on a naïve, majority-class baseline [22]. This formulation ensures that the Representational Interrogator remains sensitive to high-entropy features that exhibit the most significant coordinate dependence within the victim model's internal representation.

Step 4: Train the Representational Interrogator using Uncertainty-Normalized Weighting

- Define a weighting coefficient ω_i for each head g_{ϕ_i} to normalize the loss magnitude relative to the input feature's natural scale:
  - Set $\omega_i = (U_i + \varepsilon)^{-1}$, where U_i is the statistical uncertainty from Step 1 and ε is a small constant (e.g., 1.0^{-8}) to safeguard numerical stability. This ensures that the total training objective, $\mathcal{L}_{\text{total}} = \sum_{i=1}^{N} \omega_i \mathcal{L}_i$, is not dominated by high-variance features, thereby forcing the shared stem to learn a balanced representation across the entire feature space.
  - Note: For normalized continuous features where $\sigma_i^2 \approx 1$, this naturally simplifies to $\omega_i \approx 1$.

Step 5: Quantify Coordinate Dependence (CDI) and Absolute Information Gain (AIG)

- After training converges, for each feature $i \in \{1 \ldots N\}$:
  1. Calculate the reconstruction loss $\mathcal{L}_i$ between X_i and $\hat{\mathcal{X}}_i$.
  2. Calculate CDI_i using Equation 2.
  3. Calculate AIG_i using Equation 3 (for continuous features) or Equation 4 (for categorical features).

Output:

https://doi.org/10.3390/make8110358

- Return the comprehensive risk profile:
  $P = \{(CDI_1, AIG_1), (CDI_2, AIG_2), \ldots, (CDI_N, AIG_N)\}$
- Interpretation: Features whose CDI values exceed the Invariance Horizon ($CDI > e^{-1} \approx 0.37$) are flagged as containing coordinate persistence, identifying them as being susceptible to privacy attacks. The magnitude of such coordinate persistence can be compared among features with different modalities using the AIG values.

3.2. Evaluative Experiments

A comprehensive series of experiments was conducted in order to evaluate the performance of the Representational Interrogator framework in comparison to the established baseline of executing a full battery of state-of-the-art attribute inference attacks. These experiments were designed to test two primary propositions:

1. The Representational Interrogator yields a complete risk profile with single-pass efficiency, proving significantly faster in wall-clock time than the traditional, sequential baseline methods.
2. The intra-modality Coordinate Dependency Index rankings produced by the RI demonstrate a strong and significant correlation with the coordinate persistence captured by state-of-the-art AIAs.

By validating these propositions across three diverse benchmarking datasets, we demonstrate the utility of the dual-metric pipeline as a principled diagnostic measure for representational state.

3.2.1. Benchmarking Datasets

To validate the efficacy of the Representational Interrogator across different domains and data distributions, the experiments relied on three canonical benchmarking datasets: (1) German Credit, (2) Online Shoppers Purchasing Intention, and (3) Adult Census Income. These datasets were selected not only for their ubiquity in the ML literature, but also because they represent standard "stress tests" for privacy leakage and coordinate persistence. Prior to being used in the experiments, each dataset was preprocessed using a standard pipeline, with listwise deletion being used to handle missing data, continuous input features being normalized into *z*-scores, and categorical features being one-hot encoded [23–25].

The German Credit dataset [26] is a smaller, high-dimensional dataset containing 1,000 credit applicant records, where the prediction task is to classify an applicant's credit risk. This dataset represents the "small data" challenge in privacy auditing because with only 1,000 records and 20 features, ML models trained on the German Credit dataset are highly prone to overfitting, making them theoretically more susceptible to coordinate memorization and inversion [27]. For this reason, German Credit has become a standard benchmark for testing whether a privacy metric can detect leakage in low-sample regimes. It has been used to assess how post-hoc explanations might reveal private information [28] and to validate pattern distortion techniques that have been designed to preserve privacy [29].

The Online Shoppers Purchasing Intention dataset [30] represents a behavioral e-commerce domain, tracking the session activities of approximately 12,000 users to predict whether a website visit resulted in a transaction. This dataset provides a reasonably balanced feature distribution of 10 continuous features and seven categorical features that capture each user's site engagement and technical and temporal context, and has hence become a key benchmark in ML privacy research due to its representation of behavioral clickstream data. In this context, it has been used, for example, to audit the privacy of synthetic datasets [31] and to develop models that identify whether data are inherently

vulnerable to membership inference attacks [32]. Unlike the financial focus of German Credit or the demographic focus of Adult Census Income, this dataset allows for the validation of the Representational Interrogator against behavioral coordinate dependencies and session-level intent.

The Adult Census Income dataset [33] contains approximately 45,000 records (after cleaning) that were extracted from the 1994 U.S. Census, with the prediction task being to determine whether an individual's income exceeds USD $50,000 per year. The Adult Census Income dataset is widely considered to be a standard benchmark for tabular privacy auditing [34]. The dataset's mix of categorical and continuous features, combined with its moderate size, makes it a key baseline for evaluating both MIAs and AIAs. This dataset has been used extensively to benchmark membership inference risks [2] and to audit the privacy/utility tradeoffs of synthetic data generation [35].

Collectively, these three datasets provide a heterogeneous testbed that spans a wide spectrum of sample sizes, feature modalities, and class balances. From the small-sample, high-overfitting regime of German Credit to the large-scale demographic distributions of Adult Census Income, this variety ensured that the Representational Interrogator's single-pass efficiency and the CDI's diagnostic accuracy would be validated against a diverse array of real-world data characteristics. By demonstrating stability across these varying entropy profiles, the experiments establish the CDI as a robust, universal measure of coordinate persistence regardless of the underlying data domain.

3.2.2. Victim Model Implementation

To ensure that the experimental assessment of the Representational Interrogator framework would be model-agnostic and generalizable, two distinct victim model (f) architectures were trained for each of the three benchmarking datasets. These architectures represent two dominant and mathematically distinct classes of modern machine learning methods:

1. Deep Neural Networks (DNNs): These models represent the connectionist approach to learning, characterized by continuous, differentiable layers. The DNN victim models were implemented as fully connected multilayer perceptrons (MLPs) with three hidden layers (128, 64, and 32 neurons) and ReLU activation functions. Models were implemented using the PyTorch 2.9.1 library and trained via the Adam optimizer ($lr = 0.001, \beta_1 = 0.9, \beta_2 = 0.999$) for a maximum of 300 epochs using a batch size of 64 [36,37]. An early stopping patience of 10 was utilized to ensure that the models captured the features' predictive essence. The training objective utilized focal loss ($\gamma = 2.0$) to address class imbalance in the target variable.
2. Gradient-Boosted Decision Trees (GBDTs): These models represent the ensemble approach to learning, characterized by non-differentiable, axis-aligned decision boundaries. The GBDT models were implemented using the LightGBM 4.6 library with default hyperparameters [38]. By including GBDTs, the experiments test whether the Representational Interrogator, which is itself a neural network, can successfully extract the CDI and AIG from a non-neural victim architecture.

With three benchmarking datasets and two classes of machine learning architectures, a total of six unique victim models were trained and utilized for the experimental validation.

3.2.3. Representational Interrogator Implementation

A corresponding Representational Interrogator model (g) was trained for each of the six victim models. The architecture of each Representational Interrogator was designed to perform high-speed architectural disentanglement via an inverse multi-task learning framework:

- Shared Stem (h_θ): To distill the victim model's output Y into a robust latent representation z, the stem was implemented as a three-layer MLP with hidden dimensions of 32, 64, and 128 neurons, respectively, utilizing ReLU activation functions. This configuration resulted in a shared latent representation $z \in \mathbb{R}^{128}$ that captured the predictive essence of the original training data.
- Decoder Heads (g_{ϕ_i}): Branching from the shared stem, N parallel decoder heads were instantiated, one for each input feature. Each head consisted of a single linear layer optimized for its specific modality by using a softmax output for categorical features or a single linear neuron for continuous features.

Each Representational Interrogator model was implemented in Python using the PyTorch library with a batch size of 64. To ensure a strictly fair comparison of computational effort, each Representational Interrogator was trained using the Adam optimizer ($lr = 0.001$) with the same early stopping criterion that was applied to the victim models (a patience of 10 consecutive epochs). This ensured that the recorded total audit time would represent the precise duration required for the framework to converge to its optimal reconstruction capability, thereby preventing artificial bias from underfitting or overtraining. The loss functions for the RI models were modality-specific, with MSE being used for continuous features and focal loss ($\gamma = 2.0$) being used for categorical features in order to prevent the auditor from converging on majority-class baselines. Each Representational Interrogator model was trained using uncertainty-normalized weighting, per Step 4 of Algorithm 1.

3.2.4. Baseline Methods for Comparative Analysis

The baseline sequential feature-level privacy auditing methods were implemented by executing a battery of state-of-the-art attribute inference attacks against each input feature. To provide a scientifically robust comparison for the Representational Interrogator framework, two distinct classes of attacks were selected, representing the highest levels of adversarial capability currently established in the literature. The architectural specifications used for each attack are included in order to ensure reproducibility:

1. Discriminative SOTA (Neural Inversion Attack): This AIA utilizes a supervised multi-layer perceptron to learn the inverse mapping ($Y \rightarrow X_i$) for each feature. The attack architecture used in the experiments consisted of a feed-forward network with hidden layers of 32, 64, and 128 neurons, respectively. Each layer utilized ReLU activation functions, culminating in a task-specific output layer (linear for continuous features or softmax for categorical features). Loss functions were also modality-dependent, with MSE being utilized for continuous feature reconstruction and focal loss ($\gamma = 5.0$) being used for categorical feature classification in order to mitigate the impact of class imbalance. This discriminative model was trained using the Adam optimizer ($lr = 0.001$) with a batch size of 64 and was governed by the same early-stopping criterion as the RI framework.
2. Generative SOTA (cGAN Attack): This AIA utilizes a conditional generative adversarial network (cGAN) to reconstruct prototypical averages conditioned on the victim model's output. The implementation used in the experiments involved two competing networks:

- Generator: A three-layer MLP that took a concatenated vector of a 32-dimensional latent noise signal (z) and the victim model's output (Y), processing them through hidden layers of 64 and 128 neurons with ReLU activations to produce the reconstructed feature $\widehat{X}_i$.
- Discriminator: A symmetric architecture that processed the pair (Y, X_i) or $(Y, \widehat{X}_i)$ through hidden layers of 128 and 64 neurons using LeakyReLU ($\alpha = 0.2$) and a final sigmoid activation to distinguish between real and synthetic feature values.

The cGAN attack was optimized using Adam ($lr = 0.001$) with a momentum parameter $\beta_1 = 0.5$ and a binary cross entropy adversarial loss. To prevent mode collapse and enforce high-fidelity reconstruction, the training objective included a reconstruction penalty weighted by $\lambda = 100$. The reconstruction loss utilized MSE for continuous features and focal loss ($\gamma = 5.0$) for categorical features.

While these state-of-the-art discriminative and generative methods are known to be effective at revealing privacy vulnerabilities, they are inherently targeted and modular in nature. Each neural inversion or cGAN attack is optimized to reconstruct only a single feature prototype ($\widehat{X}_i$) per training instance. As such, performing a comprehensive privacy audit on a model with N input features requires the sequential instantiation and convergence of N distinct adversarial models. This establishes the prohibitive bottleneck that characterizes current feature-level privacy auditing practices—a linear scaling of time and hardware resources that the Representational Interrogator collapses into a single-pass process.

3.2.5. Experimental Procedure and Statistical Tests

Following the implementation of the victim models and auditing frameworks, a comprehensive series of experiments was conducted in order to validate the study's theoretical propositions empirically. To evaluate the framework's robustness to random weight initialization, a total of 30 trials—each utilizing a different random seed—were carried out for each experimental condition. Rather than representing statistically independent population samples, these trials were designed to measure algorithmic convergence stability. As per the Central Limit Theorem, this sample size ensured that the resulting distributions of the performance metrics across these initialization variations would be approximately Gaussian, allowing for the application of parametric statistical tests. All experiments were executed within a controlled Ubuntu Linux hardware environment utilizing Python 3.12.3 and PyTorch 2.9.1, accelerated by a dedicated NVIDIA RTX 4060 Ti GPU. For every trial, the auditing frameworks were required to process the full record count of each benchmarking dataset. This use of complete datasets ensured that the observed performance would reflect realistic deployment scenarios in which an organization must audit its entire production training or validation set.

The primary temporal metric utilized in the experiments was Total Audit Time, which quantified the total duration required to generate a complete, feature-level risk profile for all N input features in a given dataset. The total audit time for the RI method was recorded as the wall-clock time required to train the single inverse multi-task learning model on the benchmarking dataset until the early stopping criterion was triggered. This single pass encompassed the combined forward and backward steps necessary to learn the reconstruction mappings for all N features simultaneously, representing a single training pass. Specifically, this single-pass efficiency is driven by the framework's ability to share the primary computational load. Although the total number of floating-point operations (FLOPs) required for gradient backpropagation technically maintains a linear relationship with N due to the presence of N parallel decoder heads, the framework is architecturally designed to "buy down" this cost. The shared stem, which contains the vast

majority of the framework's parameters, is computed only once per batch, so the incremental computational tax of adding a new lightweight feature head is negligible in practice. As such, the temporal cost of a comprehensive audit becomes largely independent of feature cardinality, allowing the RI to collapse the sequential auditing tax of traditional, sequential batteries into a single-pass operational profile. The total audit time for the baseline AIA methods was recorded as the cumulative sum of the execution times for N distinct, sequential attacks (neural inversion or cGAN), with one attack being instantiated and converged for each individual feature in the dataset. These audit time workflows are illustrated in Figure 3 below.

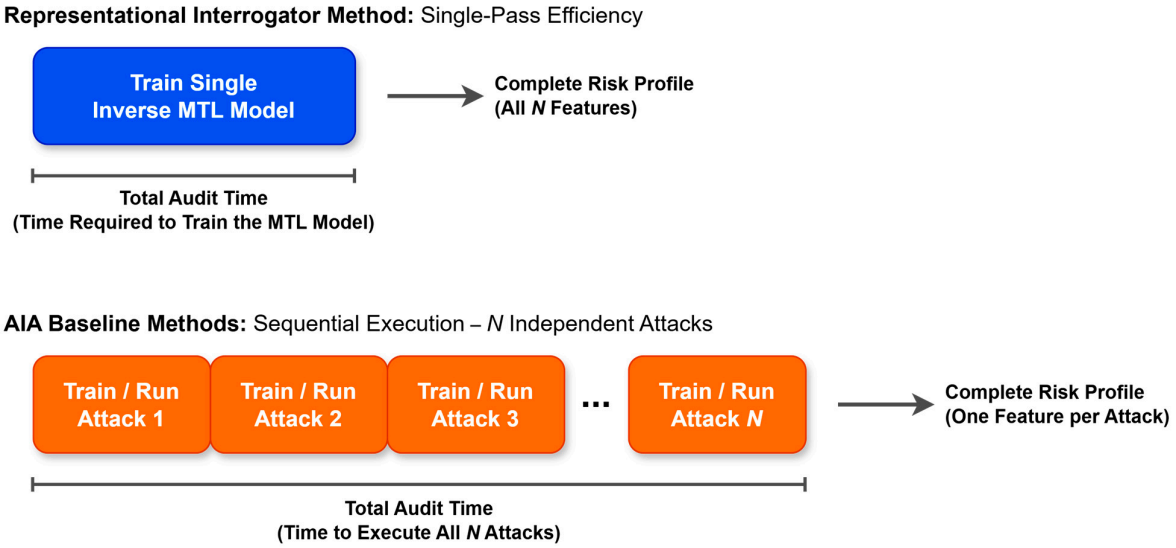

Figure 3. Comparison of computational workflows for ML privacy auditing. (**Top**) The Representational Interrogator resolves the Audit Scalability Gap with single-pass efficiency, leveraging a single inverse multi-task learning architecture to generate a complete feature-level privacy risk profile in a single training pass. (**Bottom**) The AIA baseline methods require the instantiation and convergence of N distinct adversarial models (one per feature) to achieve the same diagnostic coverage. Total Audit Time is recorded as the wall-clock time required to complete the respective workflow for all N features in a dataset.

The efficacy of the Representational Interrogator and the Coordinate Dependency Index were statistically evaluated using two primary metrics:

1. Computational Speed (Efficiency): To quantify the operational advantage of the RI method, the mean audit time across each experiment's 30 trials for the Representational Interrogator was compared against the mean cumulative audit time for each battery of baseline attacks. Differences in speed were assessed using independent, two-sample Welch's *t*-tests [39], which account for potential inequalities in variance between the two methods' timing distributions.
2. Ranking Quality (Accuracy): To evaluate the diagnostic accuracy of the Coordinate Dependency Index, we assessed the extent to which the Representational Interrogator's CDI vector achieved a consistent rank order with the risk rankings established by the state-of-the-art AIA baselines. Ranking correlations were computed separately for continuous and categorical features within each dataset. This bifurcation was methodologically essential for two reasons:
    a. For the AIAs, the adversarial risk for continuous features was quantified using the Coefficient of Determination, whereas categorical risk was quantified using

Cohen's Kappa. Since these metrics operate on fundamentally different mathematical scales, aggregating them into a single vector would introduce spurious variance and obscure the modality-specific performance of the auditor. This, as noted in the Introduction, is the Measurement Challenge that plagues current ML privacy auditing methods.

b. Tabular datasets often exhibit a significant imbalance in feature types. A particular dataset, for example, may contain many categorical features but only a few continuous features, or vice versa. Separating the analysis ensured that the diagnostic signal from one modality would not be diluted or overshadowed by the higher cardinality of the other.

Two distinct rank correlation coefficients were used to evaluate the alignment between the CDI and AIA risk vectors. First, Spearman's rank correlation (ρ) was used to measure the monotonic relationship between the CDI and the baseline risk rankings. Second, Kendall's Tau (τ) was used as a primary measure of diagnostic accuracy due to its superior statistical robustness in small-sample regimes, such as those involving the limited number of features (N) that is typical in tabular benchmarking datasets. It is important to note that the CDI and AIG share a fundamental mathematical dependency on the underlying reconstruction ratio. As such, their intra-modality rank behaviors are highly collinear and strictly mathematically identical for continuous features. Rank-order statistics were therefore computed exclusively using the CDI to avoid statistical redundancy and to maintain structural alignment with the normalized scales of the legacy AIA baselines.

For each experimental condition, Spearman and Kendall correlations were computed for each of the 30 trials. To determine the final aggregated correlation for each experiment, the trial-level correlation coefficients were converted using Fisher's z-transformation, with the sample size (n) being defined as the number of features within the respective modality track. The statistical significance of the aggregated correlation coefficients was determined using one-sample, one-tailed t-tests against the null hypothesis that $\rho \leq 0$ and $\tau \leq 0$. This directional approach aligns with the core research proposition that the Coordinate Dependency Index should exhibit a strictly positive alignment with established metrics of adversarial risk.

Finally, because $CDI > e^{-1}$ operates as a threshold diagnostic, we quantified the finite-sample stability of this boundary to ensure that any exceedances of the threshold were not artifacts of random noise or overfitting. To do so, we computed a stability score $s_i = \frac{1}{30}\sum_{t=1}^{30} 1\{CDI > e^{-1}\}$ for each feature i, representing the empirical probability that the feature consistently exceeded the Invariance Horizon across the 30 random initializations. These stability scores, along with the mean CDI and standard deviation for the most vulnerable categorical and continuous features in each benchmarking dataset are reported in the results.

3.2.6. Boundary Sensitivity and Interpretability Analysis

Finally, to validate the mathematical stability and cross-modality consistency of the CDI, we conducted a synthetic stress test that was designed to map the relationship between reconstruction error, coordinate persistence, and Absolute Information Gain. Using a controlled sample size of $n = 15{,}000$, we established two feature prototypes: (1) a continuous variable drawn from a standard normal distribution ($X \sim \mathcal{N}(0, 1)$) and (2) a categorical variable with three perfectly balanced classes ($n = 5{,}000$ cases per class; $H(X) \approx 1.098$ nats). By systematically injecting noise to vary the error ratio ($\mathcal{L}_i/U_i$) from 0.000 (perfect reconstruction) to 2.000 (high distillation), we calculated the corresponding CDI and its associated AIG in absolute bits, with the AIG values being computed using the standard, modality-specific Shannon capacity formulas defined in Section 2.5.2. This analysis

allowed for a precise mapping of the unitless CDI to a physical bitrate, thereby ensuring the interpretability of the diagnostic zones across disparate data modalities.

In summary, the statistical validation framework was designed to provide a rigorous, multi-modal assessment of the Representational Interrogator's diagnostic accuracy. By bifurcating the analysis into continuous and categorical tracks, we ensured that the Coordinate Dependency Index was evaluated against the specific mathematical properties of both regression-based and classification-based adversarial risks. The utilization of Kendall's τ alongside Spearman's ρ provided a conservative and robust measure of rank-order agreement, particularly in the finite-feature regimes of tabular benchmarks. Finally, the aggregation of results across the 30 initialization trials (using the Fisher z-transformation and one-tailed t-tests) ensured that the reported correlations and p-values would provide stable and statistically robust insights into the framework's ability to faithfully proxy state-of-the-art privacy auditing methods in a single pass.

4. Results and Discussion

The experimental evaluation of the Representational Interrogator demonstrates a significant advancement in the scalability and precision of ML privacy auditing. By transforming the auditing objective from a series of independent adversarial attacks into a unified multi-task learning problem, the RI framework successfully addresses the computational bottleneck that is inherent in current state-of-the-art ML privacy auditing methods. The results across 30 initialization trials establish that the Coordinate Dependency Index provides a diagnostically accurate measure of intra-modality adversarial reconstruction risk, while the framework's single-pass efficiency profile eliminates the sequential auditing tax. Specifically, the findings presented in this section validate three core principles:

1. Computational Superiority: The RI framework can provide a complete ML privacy profile much more quickly than generative and discriminative baseline batteries, with efficiency gains scaling alongside dataset dimensionality.
2. Diagnostic Fidelity: The CDI maintains strong and highly statistically significant rank-order correlations (Spearman's ρ and Kendall's τ) with established adversarial risk markers, ensuring that speed is not achieved at the expense of diagnostic reliability.
3. Model-Agnosticism: The RI framework's utility remains consistent across disparate victim model architectures, including both deep neural networks (DNNs) and gradient-boosted decision trees (GBDTs), demonstrating its effectiveness regardless of the underlying model's differentiability.

Collectively, these results position the Representational Interrogator as a robust fidelity guardrail that is capable of identifying coordinate dependencies in complex, high-dimensional datasets that would otherwise be computationally prohibitive to audit using conventional methods.

4.1. Computational Efficiency and Scaling Dynamics

The results obtained from the experiments with respect to the comparative computational efficiency of the RI framework are summarized in Table 2 below:

Table 2. Comparative Computational Efficiency of ML Privacy Auditing Methods.

Dataset	Victim Model	Mean Total Audit Time (Seconds)			Representational Interrogator Speedup	
		Representational Interrogator	Neural Inversion Attack	cGAN Attack	vs. Neural Inversion	vs. cGAN
German Credit	DNN	3.23	9.37	294.71	2.90× ***	91.25× ***
	GBDT	7.01	18.06	295.07	2.58× ***	42.11× ***
Online Shoppers	DNN	102.87	270.82	2962.42	2.63× ***	28.80× ***
	GBDT	92.07	254.60	2920.52	2.77× ***	31.72× ***
Adult Census Income	DNN	399.42	1135.77	8975.37	2.84× ***	22.47× ***
	GBDT	393.20	1124.58	8974.51	2.86× ***	22.82× ***

*** $p < 0.001$.

The experimental evaluation provides statistical confirmation that the Representational Interrogator effectively addresses the computational "auditing tax" that is characteristic of traditional ML privacy auditing methods. By shifting the complexity of the auditing task from a sequential, linear growth model—where each feature must be attacked individually—to a single-pass operation, the RI framework enables high-fidelity privacy auditing at a fraction of the traditional cost. Across 30 initialization trials, the Representational Interrogator demonstrated decisive and statistically significant speedups compared to both the discriminative Neural Inversion and generative cGAN attack batteries. Specifically, all temporal comparisons reached a significance level of $p < 0.001$, confirming that the observed efficiency gains are highly reliable and robust to variations in random weight initialization.

When evaluated against the discriminative Neural Inversion AIA baseline, the Representational Interrogator consistently achieved speedup factors ranging from 2.6× to 2.9× across the entire experimental suite. For the large-scale Adult Census Income dataset, the framework completed the comprehensive feature privacy audit in approximately 393 to 399 s per trial, while the corresponding Neural Inversion AIA battery required over 1100 s to reach completion. This consistency indicates that even when compared to relatively lightweight MLP-based attacks, the multi-task learning architecture of the Representational Interrogator provides a substantial and reliable reduction in total auditing time.

The most dramatic efficiency gains were observed relative to the generative cGAN AIA battery, where the computational overhead of training individual networks for each feature becomes particularly acute. In the German Credit domain, the Representational Interrogator achieved its highest relative efficiency, performing the audit 91.25 times faster than the generative baseline for the deep neural network (DNN) victim model. In the more complex behavioral and demographic domains of the Online Shoppers and Adult Census Income datasets, these speedups remained significant, ranging from 22.4 to 31.7 times faster. These results collectively demonstrate that as the complexity of the adversarial attack increases, the relative advantage of the single-pass RI framework grows exponentially, since it bypasses the massive temporal costs associated with training individual generative adversarial networks for every target attribute.

The efficiency of the RI framework was also observed to be remarkably robust across disparate victim model architectures. The Representational Interrogator maintained consistent performance advantages when auditing deep neural networks, notably in the high-dimensional Online Shoppers dataset where it was 28.8 times faster than the cGAN AIA baseline. These gains were similarly pronounced for non-differentiable gradient-boosted decision tree (GBDT) victim models. For example, in the German Credit domain, the RI was 42.1 times faster than the cGAN battery when auditing the GBDT-based victim, confirming that the framework's inherent architectural efficiency is fundamentally model-agnostic.

These findings underscore a critical shift in the scalability of ML privacy audits. In traditional linear auditing, the computational and human engineering costs of a comprehensive audit scale directly with the number of features, rendering the process prohibitively expensive for high-dimensional ML models in production environments. The Representational Interrogator's shared-stem design instead allows for the simultaneous extraction of coordinate dependencies for all features in a single training pass, effectively flattening the cost curve of the audit. As theorized in Section 3.2.5, the empirical results confirm that the framework achieves a state of single-pass operational efficiency. While the gradient operations technically scale with N, the shared architectural stem ensures that the marginal cost of additional feature heads is negligible in practice. This transformation is best illustrated by the Adult Census Income dataset, where a full generative audit was observed to require nearly 2.5 h per trial, while the Representational Interrogator provided a statistically equivalent diagnostic in just over six minutes. Such efficiency suggests that the framework is well-suited for real-time model monitoring pipelines where traditional adversarial batteries would be too slow to be practical.

4.2. Diagnostic Fidelity and Risk Proxying

The results obtained from the experiments with respect to the diagnostic fidelity of the Coordinate Dependency Index values generated by the RI framework are summarized in Table 3 below.

Table 3. Diagnostic Fidelity of Coordinate Dependency Index for ML Privacy Auditing.

Dataset	Victim Model	CDI vs. Neural Inversion Correlations				CDI vs. cGAN Correlations			
		Continuous		Categorical		Continuous		Categorical	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ
German Credit	DNN	0.981 ***	0.949 ***	0.659 ***	0.573 ***	0.985 ***	0.959 ***	0.388 ***	0.299 ***
	GBDT	0.967 ***	0.917 ***	0.838 ***	0.745 ***	0.985 ***	0.959 ***	0.603 ***	0.492 ***
Online Shoppers	DNN	0.983 ***	0.947 ***	0.691 ***	0.614 ***	0.882 ***	0.750 ***	0.525 ***	0.427 ***
	GBDT	0.983 ***	0.956 ***	0.328 ***	0.289 ***	0.907 ***	0.801 ***	0.277 ***	0.230 ***
Adult Census Income	DNN	0.994 ***	0.987 ***	0.868 ***	0.709 ***	0.998 ***	0.996 ***	0.883 ***	0.730 ***
	GBDT	0.990 ***	0.978 ***	0.879 ***	0.721 ***	0.998 ***	0.996 ***	0.881 ***	0.732 ***

*** $p < 0.001$, $\boldsymbol{\tau}$ = Kendall's Tau, $\boldsymbol{\rho}$ = Spearman's Rho.

To evaluate whether the Representational Interrogator's efficiency comes at the expense of its diagnostic utility, we assessed the rank-order alignment between the CDI and the traditional reconstruction risk metrics established by the adversarial baselines. The results across all 30 initialization trials confirm that the CDI serves as a faithful and highly stable proxy for state-of-the-art adversarial risk, consistently converging to a highly similar diagnostic hierarchy regardless of the starting seed. By utilizing not only Spearman's ρ, but also Kendall's τ as a more conservative measure of rank correlation, we demonstrate that the information leakage hierarchy identified by the Representational Interrogator in a single pass strongly corresponds to the hierarchy produced by exhaustive, feature-by-feature attack batteries.

The alignment observed in the continuous feature track between the CDI and the discriminative and generative baselines was particularly striking, with the conservative Kendall's τ correlations ranging from 0.750 to 0.996 across all datasets and victim model architectures (median $\tau = 0.958$). At the low end of the range, the τ correlations with the generative cGAN baseline on the Online Shoppers dataset were 0.750 and 0.801 for the DNN and GBDT victim models, respectively. On the high end of the range, the τ correlation with the generative cGAN baseline on the Adult Census Income dataset was nearly perfect at 0.996 for both the DNN and GBDT victims. These results provide strong

empirical evidence that the CDI accurately captures the coordinate dependencies of continuous variables and identifies which features are most vulnerable to privacy attacks.

Among the categorical features, the Representational Interrogator also maintained a stable and highly statistically significant diagnostic signal across initialization variations for all domains, with the conservative Kendall's τ correlations ranging from 0.230 to 0.745 (median $\tau = 0.593$). This relative reduction in Kendall's τ compared to the continuous features is statistically driven by the high frequency of zero-risk scores within the adversarial reference batteries, a phenomenon that we define as "adversarial clumping". Analysis of the categorical feature trials revealed that approximately 62.7% of all attributes across the benchmarking suite yielded a Neural Inversion baseline risk score of exactly 0.0, while the generative cGAN trials yielded slightly more variegated dependencies with 48.8% of the attributes exhibiting scores of exactly 0.0. Specific examples of this clumping were evident in the German Credit dataset, for which 10 out of 13 categorical features showed zero adversarial risk, and in the Online Shoppers dataset, where five out of seven categorical features clustered at this same zero-risk baseline. Correlating the continuous, high-resolution CDI metric against these highly invariant, binary success/failure thresholds effectively flattens the reference hierarchy, imposing an artificial mathematical ceiling on the magnitude of the measured rank-order agreement.

It is important to recognize that this observed adversarial clumping is entirely correct within the limited, single-target scope of legacy attribute inference methods. Given that traditional AIAs are designed to interrogate specific instances in isolation, they perform strict attribute inference but remain architecturally blind to the complex inter-feature correlations and hidden dependencies that the victim model would typically have preserved as distributional prototypes. By contrast, the RI's shared-stem architecture distills the victim model's output into a unified latent representation, enabling a phenomenon we define as "leakage triangulation." Essentially, the multi-task learning framework allows the Representational Interrogator to capture leakage via inter-feature associations that single-target attacks must naturally miss. Traditional AIA risk scores therefore serve as a valid but coarse-grained indicator of distillation, while the RI provides a significantly more sensitive risk profile. Ultimately, we believe that the lower correlation scores for categorical features do not represent a failure of the RI framework, but instead highlight its superior ability to detect granular coordinate dependencies in the Persistence Zone that isolated adversarial probes are simply too architecturally limited to perceive.

4.3. Finite-Sample Stability and Robustness

To validate that features crossing the Invariance Horizon exhibited genuine coordinate dependence rather than finite-sample variance, we analyzed the empirical stability (s_i) of the Invariance Horizon threshold exceedance. As shown in Table 4 below, all of the features identified by the Representational Interrogator as possessing the highest mean CDI across both modalities consistently maintained stability scores of 1.0. This indicates that for the most exposed attributes, such as *capital-gain* in the Adult Census dataset or *PageValues* in the Online Shoppers dataset, the measured coordinate dependence exceeded the exact algebraic baseline in 100% of the 30 random initializations. The small standard deviations (e.g., $\sigma \leq 0.01$ for most top-ranked features) further confirm that while borderline features may naturally fluctuate near the Invariance Horizon, the framework's core diagnostic signal for high-risk features is robust to finite-sample noise and architectural initialization.

Table 4. Finite-Sample Stability of Highest-Risk Features.

Dataset	Victim Model	Feature Modality	Most Vulnerable Feature	Mean CDI (Std. Dev.)	Stability (s_i)
German Credit	DNN	Categorical	status	0.383 (<0.001)	1.0
		Continuous	duration	0.385 (<0.001)	1.0
	GBDT	Categorical	status	0.413 (0.004)	1.0
		Continuous	duration	0.424 (0.004)	1.0
Online Shoppers	DNN	Categorical	Month	0.394 (0.005)	1.0
		Continuous	PageValues	0.529 (0.022)	1.0
	GBDT	Categorical	Month	0.463 (0.007)	1.0
		Continuous	PageValues	0.611 (0.002)	1.0
Adult Census Income	DNN	Categorical	relationship	0.441 (0.003)	1.0
		Continuous	capital-gain	0.466 (0.010)	1.0
	GBDT	Categorical	relationship	0.480 (0.001)	1.0
		Continuous	capital-gain	0.521 (0.009)	1.0

As shown in Table 4, the magnitude of the measured CDI for each specific feature varies between the DNN and GBDT victim models. This variation confirms that the RI framework correctly measures the unique representational state and coordinate dependencies retained by the specific victim architecture, rather than merely reflecting intrinsic, model-agnostic properties of the dataset.

4.4. Sensitivity Analysis and Informational Anchoring

The results of the synthetic stress test are reported in Table 5 below. These results confirm that the Invariance Horizon ($e^{-1} \approx 0.37$) serves as a universal anchor for identifying coordinate persistence. Regardless of the underlying data modality or loss function, the CDI behaves as theoretically predicted when normalized against the natural uncertainty (U_i) of the feature.

Table 5. Sensitivity Analysis of the Coordinate Dependency Index (CDI).

Error Ratio ($\mathcal{L}_i/U_i$)	Continuous CDI	Categorical CDI	Continuous AIG (Bits)	Categorical AIG (Bits)	Auditing Zone
0.000	1.0000	1.0000	Max	1.5850	Persistence
0.125	0.8825	0.8825	1.5000	1.3868	
0.250	0.7788	0.7788	1.0000	1.1887	
0.375	0.6873	0.6873	0.7075	0.9906	
0.500	0.6065	0.6065	0.5000	0.7925	
0.625	0.5353	0.5353	0.3390	0.5944	
0.750	0.4724	0.4724	0.2075	0.3962	
0.875	0.4169	0.4169	0.0963	0.1981	
1.000	0.3679 (e^{-1})	0.3679 (e^{-1})	0.0000	0.0000	Invariance Horizon
1.500	0.2231	0.2231	0.0000	0.0000	Distillation
2.000	0.1353	0.1353	0.0000	0.0000	

As shown in Table 5, the CDI provides high diagnostic resolution within the Persistence Zone. Specifically, the analysis highlights the distinct informational capacity limits of different data modalities, with the continuous Gaussian feature possessing a theoretically infinite capacity at perfect reconstruction, while the three-class categorical variable is strictly bounded at $\log_2(3) \approx 1.5850$ bits. Despite the differing decay trajectories, the

results reported in Table 5 demonstrate that both modalities converge perfectly at the Invariance Horizon. At this boundary, where $\mathcal{L}_i = U_i$ and the error ratio is 1.0, the information gain for both the continuous and categorical features drops to exactly zero bits. This consistent mathematical alignment empirically justifies the classification of any CDI score less than or equal to $e^{-1} \approx 0.37$ as being within the Distillation Zone. Collectively, these results confirm the CDI as a high-fidelity instrument for identifying coordinate dependencies that traditional ML privacy auditing methods may overlook.

In summary, the consistent performance of the RI framework across disparate victim model learning algorithms confirms that its capacity to quantify coordinate dependencies is independent of decision boundary differentiability. These findings, further bolstered by the synthetic sensitivity analysis, validate the Coordinate Dependency Index as a robust metric that is empirically grounded in information theory. Although specific bitrate capacities vary by data modality, translating the CDI into absolute bits via the AIG provides a rigorous, non-arbitrary interpretation of adversarial risk that consistently converges to zero information gain at the Invariance Horizon. Ultimately, by resolving the Audit Scalability Gap via single-pass architectural disentanglement and solving the measurement challenge through a dual-metric diagnostic pipeline, the Representational Interrogator establishes itself as a robust and reliable tool for proactive ML privacy auditing in complex, multi-modal data environments.

5. Conclusions

5.1. Summary of Contributions

This paper introduced a theoretically grounded and computationally efficient framework for quantifying coordinate persistence in machine learning models. By reframing model inversion not as a localized adversarial failure, but as a diagnostic interrogation of an ML model's coordinate dependencies, we have successfully reconciled the Information Bottleneck principle with the practical success of attribute inference attacks. Our primary contribution, the Coordinate Dependency Index (CDI), provides the first unitless, model-agnostic metric that is capable of normalizing ML privacy risk across disparate data modalities. Beyond its theoretical elegance, the framework utilizes Absolute Information Gain (AIG) to translate the CDI's normalized risk into information-theoretic bitrates, providing a physical anchor for the Invariance Horizon at $e^{-1} \approx 0.37$. This formulation transforms ML privacy auditing from a fragmented, heuristic-based exercise into a mathematically precise science of information disentanglement, establishing the CDI as a robust indicator of representational maturity.

Furthermore, this paper resolved the long-standing Audit Scalability Gap through the development of the Representational Interrogator (RI). By employing a novel, inverse multi-task learning architecture, the RI framework collapses the computational complexity of full-feature auditing from a prohibitive linear scale—which makes modern high-dimensional ML models practically impossible to audit—into a single-pass efficiency process. Experimental results across 30 initialization trials empirically validated the magnitude of this innovation, demonstrating that the RI can generate a complete, feature-level risk profile up to 91 times faster than state-of-the-art generative AIA batteries. Crucially, the observed efficiency gains were achieved without sacrificing diagnostic fidelity. The intra-modality CDI rankings produced by the RI demonstrate strong and statistically significant alignment with the coordinate persistence captured by state-of-the-art adversarial baselines across all victim model architectures. By providing a complete and accurate privacy profile in a single training pass, the RI framework empowers organizations to identify and protect their most vulnerable data features with unprecedented speed and mathematical certainty, thereby establishing a new standard for informational transparency

that effectively transforms the prohibitive "auditing tax" into a proactive and scalable science of representational defense.

5.2. Limitations and Future Research

While the RI-CDI-AIG framework represents a significant advancement in the systematic assessment of machine learning privacy, several inherent limitations provide compelling opportunities for future research in this domain. One primary limitation is that the current study focused exclusively on tabular data benchmarks, which, while canonical for privacy auditing, do not represent the full spectrum of modern ML applications. This presents a clear opportunity to extend the framework into unstructured data domains—such as computer vision or natural language processing—to evaluate how coordinate dependencies manifest and persist within convolutional or transformer-based latent spaces.

Additionally, the current implementation of the Representational Interrogator operates under the assumption of a shadow modeling paradigm, requiring a reference dataset to establish uncertainty baselines and train the multi-task learning architecture. Developing "zero-shot" or data-free inversion techniques that can estimate the CDI without representative data would further lower the operational barrier to entry, particularly for third-party regulators who may lack access to an ML model's original training data. The temporal nature of the current metric also suggests a fertile area for investigation, as the CDI currently provides a post-hoc diagnostic snapshot of a model's representational state. Future research could explore the integration of the CDI directly into the ML model training process as a regularization term, potentially allowing developers to dynamically monitor and penalize the formation of coordinate dependencies as the model converges.

Finally, while the framework demonstrated remarkable stability, the performance of the Representational Interrogator remains fundamentally linked to the architectural capacity of the shared stem. This limitation highlights the potential for future work involving automated architecture searches to ensure that the Representational Interrogator's shared stem is always optimally scaled to the specific complexity and dimensionality of the victim model that is being interrogated. By addressing these challenges, future iterations of this framework can continue to bridge the gap between information theory and the practical demands of secure artificial intelligence.

5.3. Concluding Remarks

The development of the framework's dual metric pipeline establishes a mathematically rigorous and empirically validated threshold—anchored by the AIG at exactly zero bits of information gain—between an ML model's predictive utility and its informational vulnerability. Beyond its immediate application in ML privacy auditing, it has not gone unnoticed that the CDI has the potential to transform several adjacent fields within machine learning. In the realm of Explainable AI (xAI), for example, the CDI can be utilized to distinguish between features to which an ML model is genuinely sensitive and those that represent mere statistical noise.

In addition to its applications in xAI and ML privacy auditing, the CDI can also act as a "fidelity guardrail" for ML model pruning and compression, providing a granular diagnostic to ensure that essential predictive information is preserved while non-essential coordinate dependencies are successfully purged. Finally, as a measure of representational maturity, the CDI offers a new lens through which to view ML model training dynamics, allowing researchers to quantify precisely when a model has successfully navigated the Information Bottleneck to transform coordinate dependence into abstract knowledge. By establishing this unitless and empirically anchored science of information disentanglement, it is hoped that this paper will provide a foundation for a more

transparent, secure, and theoretically grounded era of machine learning and artificial intelligence.

Funding: This research was generously supported by a senior intramural grant from the Office of Research & Sponsored Programs at California State University, Fullerton (award number 3512632).

Data Availability Statement: All of the datasets used in the study and the source code for the RI-CDI framework and the experiments described in this paper are available at: https://github.com/daniel-soper/Representational_Interrogator (access on 20 October 2026).

Conflicts of Interest: The author declares no competing interests.

References

1. Namatevs, I.; Sudars, K.; Nikulins, A.; Ozols, K. Privacy Auditing in Differential Private Machine Learning: The Current Trends. *Appl. Sci.* **2025**, *15*, 647.
2. Shokri, R.; Stronati, M.; Song, C.; Shmatikov, V. Membership Inference Attacks against Machine Learning Models. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP)*; IEEE: Piscataway, NJ, USA, 2017.
3. Veale, M.; Binns, R.; Edwards, L. Algorithms That Remember: Model Inversion Attacks and Data Protection Law. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2018**, *376*, 20180083.
4. Fakeyede, O.G.; Okeleke, P.A.; Hassan, A.; Iwuanyanwu, U.; Adaramodu, O.R.; Oyewole, O.O. Navigating Data Privacy through It Audits: Gdpr, Ccpa, and Beyond. *Int. J. Res. Eng. Sci.* **2023**, *11*, 45–58.
5. Liu, B.; Ding, M.; Shaham, S.; Rahayu, W.; Farokhi, F.; Lin, Z. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–36.
6. Yeom, S.; Giacomelli, I.; Fredrikson, M.; Jha, S. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *Proceedings of the 2018 IEEE 31st Computer Security Foundations Symposium (CSF)*; IEEE: Piscataway, NJ, USA, 2018.
7. Fredrikson, M.; Jha, S.; Ristenpart, T. Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures. In *CCS'15: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*; Association for Computing Machinery: New York, NY, USA, 2015.
8. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46.
9. Annamalai, M.S.M.S.; Gadotti, A.; Rocher, L. A Linear Reconstruction Approach for Attribute Inference Attacks against Synthetic Data. In *SEC'24: Proceedings of the 33rd USENIX Conference on Security Symposium*; USENIX Association: Berkeley, CA, USA, 2024.
10. Jia, J.; Gong, N.Z. Attriguard: A Practical Defense against Attribute Inference Attacks via Adversarial Machine Learning. In *SEC'18: Proceedings of the 27th USENIX Conference on Security Symposium*; USENIX Association: Berkeley, CA, USA, 2018.
11. Rigaki, M.; Garcia, S. A Survey of Privacy Attacks in Machine Learning. *ACM Comput. Surv.* **2023**, *56*, 1–34.
12. Stadler, T.; Oprisanu, B.; Troncoso, C. Synthetic Data—Anonymisation Groundhog Day. In Proceedings of the 31st USENIX Security Symposium, Boston, MA, USA, 10–12 August 2022. Available online: https://www.usenix.org/conference/usenixsecurity22/presentation/stadler (accessed on 1 June 2026).
13. Nguyen, N.-B.; Chandrasegaran, K.; Abdollahzadeh, M.; Cheung, N.-M. Re-Thinking Model Inversion Attacks against Deep Neural Networks. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2023.
14. Song, J.; Junzhe, S.; Namiot, D.E. A Survey of Model Inversion Attacks and Countermeasures. *Proc. Inst. Syst. Anal. Russ. Acad. Sci.* **2023**, *73*, 82–93.
15. Kahla, M.; Chen, S.; Just, H.A.; Jia, R. Label-Only Model Inversion Attacks Via Boundary Repulsion. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2022.
16. Zhou, S.; Zhu, T.; Ye, D.; Yu, X.; Zhou, W. Boosting Model Inversion Attacks with Adversarial Examples. *IEEE Trans. Dependable Secur. Comput.* **2023**, *21*, 1451–1468.
17. Goldfeld, Z.; Polyanskiy, Y. The Information Bottleneck Problem and Its Applications in Machine Learning. *IEEE J. Sel. Areas Inf. Theory* **2020**, *1*, 19–38.
18. Tishby, N.; Zaslavsky, N. Deep Learning and the Information Bottleneck Principle. In *Proceedings of the 2015 IEEE Information Theory Workshop (ITW)*; IEEE: Piscataway, NJ, USA, 2015.

19. Saxe, A.M.; Bansal, Y.; Dapello, J.; Advani, M.; Kolchinsky, A.; Tracey, B.D.; Cox, D.D. On the Information Bottleneck Theory of Deep Learning. *J. Stat. Mech. Theory Exp.* **2019**, *2019*, 124020.
20. Zhang, Y.; Jia, R.; Pei, H.; Wang, W.; Li, B.; Song, D. The Secret Revealer: Generative Model-Inversion Attacks against Deep Neural Networks. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2020. https://doi.org/10.1109/CVPR42600.2020.00033.
21. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
22. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2017.
23. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
24. Zheng, A.; Casari, A. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*; O'Reilly Media: Sebastopol, CA, USA, 2018.
25. Little, R.J.A.; Rubin, D.B. *Statistical Analysis with Missing Data*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2019.
26. Hofmann, H. *German Credit Dataset*; UCI Machine Learning Repository: Irvine, CA, USA, 1994. https://doi.org/10.24432/C5NC77.
27. Wu, B.; Chen, C.; Zhao, S.; Chen, C.; Yao, Y.; Sun, G.; Wang, L.; Zhang, X.; Zhou, J. Characterizing Membership Privacy in Stochastic Gradient Langevin Dynamics. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2026. https://doi.org/10.1609/aaai.v34i04.6107.
28. Sonna, B.; Grastien, A.; Benn, C. Beyond Verification: Abductive Explanations for Post-Ai Assessment of Privacy Leakage. In Proceedings of the 40th Annual AAAI Conference on Artificial Intelligence, Singapore, 20–27 January 2026.
29. Hajian, S.; Domingo-Ferrer, J.; Farràs, O. Generalization-Based Privacy Preservation and Discrimination Prevention in Data Publishing and Mining. *Data Min. Knowl. Discov.* **2014**, *28*, 1158–1188.
30. Sakar, C.O.; Polat, S.O.; Katircioglu, M.; Kastro, Y. Real-Time Prediction of Online Shoppers' Purchasing Intention Using Multilayer Perceptron and LSTM Recurrent Neural Networks. *Neural Comput. Appl.* **2019**, *31*, 6893–6908.
31. Trudslev, F.M.; Lissandrini, M.; Rodriguez, J.M.; Bøgsted, M.; Dell'Aglio, D. Priveval: A Tool for Interactive Evaluation of Privacy Metrics in Synthetic Data Generation. *Proc. VLDB Endow.* **2025**, *18*, 5271–5274.
32. Moore, H.D.; Stephens, A.; Scherer, W. An Understanding of the Vulnerability of Datasets to Disparate Membership Inference Attacks. *J. Cybersecur. Priv.* **2022**, *2*, 882–906.
33. Becker, B.; Kohavi, R. *Adult Census Income Dataset*; UCI Machine Learning Repository: Irvine, CA, USA, 1996. https://doi.org/10.24432/C5XW20.
34. Vijayan, N.; Gururajan, R.; Chan, K.C. A Comparative Analysis of Defense Mechanisms against Model Inversion Attacks on Tabular Data. *J. Cybersecur. Priv.* **2025**, *5*, 80.
35. Jayaraman, B.; Evans, D. Evaluating Differentially Private Machine Learning in Practice. In *SEC'19: Proceedings of the 28th USENIX Conference on Security Symposium*; USENIX Association: Berkeley, CA, USA, 2019.
36. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L. Pytorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2019. https://doi.org/10.48550/arXiv.1912.01703.
37. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
38. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017. Available online: https://api.semanticscholar.org/CorpusID:3815895 (accessed on 1 June 2026).
39. Welch, B.L. The Generalization of ‘Student's' Problem When Several Different Population Varlances Are Involved. *Biometrika* **1947**, *34*, 28–35.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.