idSCD: Identifying Training Datasets through Semantic Correlation Descriptors

Andrada Gobeajă
POLITEHNICA University of Bucharest
andrada.gobeaja@stud.acs.upb.ro

Ionuț Hodoroagă
POLITEHNICA University of Bucharest
ionut.hodoroaga@stud.acs.upb.ro

Marius Leordeanu
POLITEHNICA University of Bucharest
Institute of Mathematics of the Romanian Academy
marius.leordeanu@upb.ro

Elena Burceanu
POLITEHNICA University of Bucharest
elena.burceanu@gmail.com

Abstract

Can a dataset be recognized from the spurious correlations it induces during training? We argue that datasets leave dataset-specific traces in a model's learned semantic correlation structure: incidental regularities that are predictive within a dataset, but not causal for the underlying task, can be internalized during training. We use this insight to study dataset-level membership inference, moving beyond existing methods that rely on behavioral or distributional evidence such as confidence scores, losses, margins, generated samples, or query responses. We introduce a white-box semantic fingerprinting approach based on semantic correlation descriptors (SCDs), which capture the semantic correlation structure learned by a model and make it comparable across dataset mixtures. In a controlled leave-one-dataset-out diagnostic, SCDs recover dataset-specific changes and perfectly separate matching from non-matching dataset pairs. We then propose a practical SCD-based membership score that tests whether a target dataset is part of a model's training mixture using only the model's SCD and the target dataset's standalone SCD, without requiring leave-one-dataset-out models. Across three diverse experimental settings, with dataset groups for native language identification, emotion classification, and medical text classification, we test both the advantages and limitations of SCD-based membership inference with different degrees of semantic separation and keyword support between dataset splits. On average, the classifier id_{SCD} based on this score achieves the highest performance and the lowest std, outperforming black-box baselines RMIA, Attack-P, and LiRA, as well as the white-box SIF baseline. These results show that dataset membership can be traced through internal semantic correlations, with the largest relative gain exceeding 60% in ROC-AUC when dataset groups expose distinct semantic particularities. Code publicly available HERE.

1 Introduction

Determining whether a dataset was used to train a model is increasingly important for model auditing. Modern models are trained on mixtures of public, private, licensed, synthetic, and domain-specific data, often with limited visibility into which sources were used. Dataset usage is therefore not merely a training detail: it affects privacy, licensing, benchmark contamination, reproducibility, and accountability. Prior work has studied related questions through membership inference, dataset-level membership inference, dataset inference, data provenance, and training-data attribution [Carlini et al., 2022, Maini et al., 2024, Sun et al., 2025]. Most existing evidence for dataset membership is

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

output-driven. Methods typically rely on confidence scores, losses, prediction margins, generated samples, query responses, or influence estimates [Carlini et al., 2022, Ye et al., 2022, Cohen and Giryes, 2022]. These signals can be effective, but they primarily ask whether a model behaves as if it has seen a dataset. This makes them vulnerable to cases where different datasets induce similar output behavior, or where membership evidence is expressed less in predictions than in the internal structure learned by the model. We ask a complementary question: can a dataset be recognized from the semantic correlations it induces during training?

Our starting point is that datasets differ in incidental regularities arising from their construction, domain, annotation protocol, and label expression patterns. These regularities may be predictive within a dataset without being causal for the underlying task. When a model is trained on a mixture of datasets, such regularities can be internalized as spurious semantic correlations and persist in the trained model. Thus, even datasets mapped to the same label space may leave different fingerprints in how semantic features align with classes. This suggests a different route to dataset-level membership inference: instead of probing the model only from the outside, we audit the semantic association structure learned inside the model. We introduce Semantic Correlation Descriptors (SCDs), which summarize this structure as dataset-level fingerprints and allow models trained on different mixtures to be compared in a shared correlation space. The goal is not to explain individual keywords, but to test whether the global semantic fingerprint of a target dataset is present in a model under inspection.

First, in a controlled leave-one-dataset-out diagnostic, SCDs recover dataset-specific changes and perfectly separate matching from non-matching dataset pairs, showing that the signal is not a generic training artifact. Second, in a more realistic membership setting, we use an SCD-based score that requires only the inspected model and a standalone reference descriptor for the target dataset, avoiding leave-one-dataset-out models. Across native language identification, emotion classification, and medical text classification, this classifier achieves the best average performance and lowest standard deviation against recent strong baselines, with the largest gains when dataset splits expose distinct semantic particularities.

Our **main contributions** are the following:

1) A semantic view of dataset-level membership inference. Existing dataset-level membership methods primarily rely on behavioral or distributional evidence, such as confidence, loss, prediction margins, generated samples, or query responses. We instead formulate dataset membership as a semantic tracing problem: can we recover the presence of a target dataset from the high-level keyword-class correlations it induces in the trained model?

2) Descriptor for dataset fingerprinting. We introduce a Semantic Correlation Descriptor (SCDs), denoted $\mathbf{d}_{\mathrm{SCD}}$, a compact representation for keyword-class semantic correlation maps. Through a shared keyword vocabulary and zero-fill alignment, SCDs place models trained on different dataset mixtures in the same correlation space, enabling direct comparison of their learned association structure. In a controlled leave-one-dataset-out analysis, this representation perfectly separates matching from non-matching dataset pairs.

3) A practical SCD-based membership score. We propose a membership score that tests whether a target dataset D_i is part of the training mixture of a model under inspection, using only the model's SCD and the standalone SCD of D_i. Unlike the controlled analysis, it does not require leave-one-dataset-out models. Across three diverse experimental settings for native language identification, emotion classification, and medical text classification, we test both the advantages and limitations of SCD-based membership inference with different degrees of semantic separation and keyword support between dataset splits. On average, the classifier id_{SCD} induced by this score achieves the highest performance and the lowest standard deviation, outperforming recent strong baselines: the black-box methods RMIA, Attack-P, and LiRA, and the white-box SIF baseline.

2 Related work

Relation to Spurious Correlations. Spurious correlations arise when models rely on superficial, non-causal cues, such as backgrounds, textures, or secondary objects in vision tasks [Geirhos et al., 2020, Wang et al., 2021, Yang et al., 2022, Lin et al., 2024] and keywords in language tasks [Gururangan et al., 2018, McCoy et al., 2019, Niven and Kao, 2019, Du et al., 2023], instead of causal concepts. These shortcuts encode dataset-specific patterns that generalize poorly beyond the training distribution

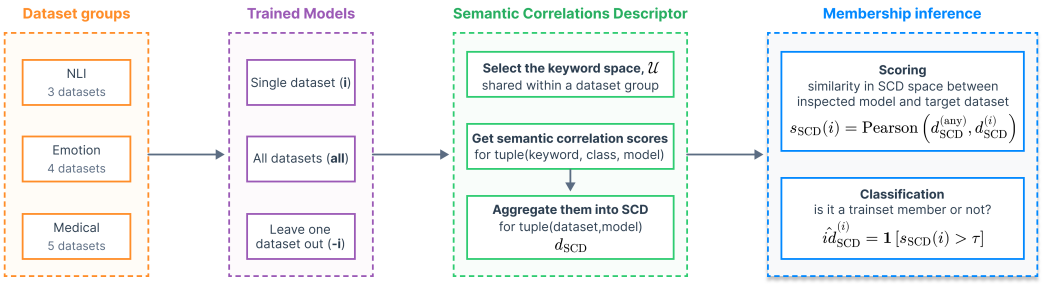


Figure 1: Dataset membership through semantic correlation fingerprints. The pipeline evaluates whether dataset-specific regularities leave recoverable traces in the semantic correlation structure of trained models. We extract semantic correlations over a keyword vocabulary shared within each dataset group, aggregate them into Semantic Correlation Descriptors (SCDs), and use the SCD classifier over the membership score to test whether a target dataset was part of the training mixture.

[Geirhos et al., 2020, Arjovsky et al., 2020], often yielding higher confidence on samples containing familiar cues and greater uncertainty on out-of-distribution or clean test samples [Izmailov et al., 2022, Ye et al., 2024]. This confidence gap can be exploited by membership inference attacks [Nasr et al., 2018, Choquette-Choo et al., 2021]. We show that such shortcut patterns can also be recovered at the dataset level as semantic correlation fingerprints, enabling dataset-level membership inference.

Relation to Membership Inference Attacks. Membership inference attacks (MIAs) determine whether a sample was used to train a machine learning model [Shokri et al., 2017, Carlini et al., 2022]. They exploit systematic differences between model behavior on training and unseen examples, and are typically classified by attacker access. White-box MIAs use internal signals such as activations, gradients, losses, confidence scores, parameters, or updates [Papernot et al., 2018, Song and Mittal, 2020, Rigaki and Garcia, 2023], including gradient/activation attacks [Nasr et al., 2019], Bayes-optimal loss tests [Sablayrolles et al., 2019], memorized feature associations [Leino and Fredrikson, 2020], and self-influence scores [Cohen and Giryes, 2022]. Black-box MIAs rely only on query outputs such as labels, confidence scores, logits, entropy, or losses [Li and Zhang, 2021, Liu et al., 2022, Carlini et al., 2022, Zarifzadeh et al., 2024], covering shadow-model attacks [Shokri et al., 2017], threshold attacks [Yeom et al., 2018, Salem et al., 2018], label-only attacks [Choquette-Choo et al., 2021, Peng et al., 2024], and calibrated likelihood-ratio attacks [Carlini et al., 2022, Bertran et al., 2023]. Prior MIAs decide membership for individual examples using per-example signals. In contrast, we study dataset-level membership: given a target dataset D_i, we ask whether D_i appeared during training, using the semantic correlation structure induced by the dataset as membership signal.

Relation to Training Data Attribution. Training data attribution methods identify training examples that most affect model outputs using influence functions to approximate upweighting or removal effects [Koh and Liang, 2020], datamodels to predict counterfactual outputs from subset membership [Ilyas et al., 2022], and scalable variants such as TRAK [Park et al., 2023], In-Run Data Shapley [Wang et al., 2025], TrackStar [Chang et al., 2024], and AirRep [Sun et al., 2025]. These reduce attribution costs through kernel approximations, single-run contribution estimates, LLM-scale gradient influence, or learned attribution representations [Chang et al., 2024].Our method targets dataset-level membership inference, asking whether an entire dataset was part of a model's training mixture, using Semantic Correlation Descriptors (SCDs) built from keyword-class correlation maps to capture dataset-induced spurious correlations as dataset-specific signatures.

3 Our approach

Datasets differ in the incidental regularities that arise from their specific construction, domain, annotation protocol, and label expression patterns. During training, models may internalize these regularities as spurious correlations: surface keyword–class associations that are predictive within a dataset but noncausal for the underlying task. In models trained on dataset mixtures, these correlations may become embedded in the learned decision structure and persist after training. We ask whether the presence of a target dataset can be recovered from the spurious correlations it induces during training. We study this question in a controlled setting with a group of datasets $D_1, \ldots, D_N$, each

mapped to a shared label space $\mathcal{Y}$, and a fixed model architecture. Given a target dataset D_i, our goal is to infer whether D_i was present in the training mixture. To formalize this, let $\mathcal{U}$ denote the global keyword vocabulary associated with the dataset collection. For each trained model m, we construct a keyword-class correlation matrix:

$$\mathbf{C}^{(m)} \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{Y}|},$$

where each entry $\mathbf{C}^{(m)}_{u,y}$ quantifies the strength of the correlation between keyword $u \in \mathcal{U}$ and label $y \in \mathcal{Y}$ in model m. Rather than interpreting individual keyword-class correlations, we treat the full correlation map as a dataset-level fingerprint and use it to infer whether D_i was present in the training mixture through its dataset-specific spurious correlation patterns.

3.1 From semantic correlations to dataset fingerprints

For each dataset group, we train three model variants designed to make dataset-specific contributions observable. For each target dataset D_i, we first train a single-dataset model m_i using only D_i. We also train a full model m_{all} on the union of all datasets, $D_{\text{all}} = \bigcup_{i=1}^{N} D_i$, and a leave-one-dataset-out model m_{-i} on the union of all datasets except D_i, $D_{-i} = \bigcup_{j\neq i} D_j$.

Semantic correlation maps. Our approach requires a scoring function that estimates the degree of keyword-class correlation across trained models. In our implementation, we obtain the correlation matrices $\mathbf{C}$ using BEE [Paduraru et al., 2026], which extracts keyword-class correlations by comparing keyword embeddings and class weight vectors in a shared embedding space. In BEE, each trained model is converted into a keyword-class correlation matrix. For a model m, each keyword $u \in \mathcal{U}$ is embedded with the model's last-token hidden state, giving a keyword embedding $\mathbf{e}^{(m)}_u$. Let $\mathbf{w}^{(m)}_y$ be the classifier's vector for class y, extracted after training, from the sequence-classification head. The keyword-class alignment is computed by cosine similarity, $\text{sim}_m(u, y) = \cos\Big(\mathbf{e}^{(m)}_u, \mathbf{w}^{(m)}_y\Big)$. The cosine alignment is then refined into the BEE scoring function for keyword-class correlations - which yields entries of the correlation matrix where rows correspond to keywords and columns to labels:

$$\mathbf{C}^{(m)}_{u,y} = \text{BEE}_m(u, y) = \text{sim}_m(u, y) - \min_{y'\in\mathcal{Y}} \text{sim}_m(u, y'). \tag{1}$$

Zero-fill alignment. Keyword extraction tools generally produce keyword sets of different sizes across datasets. As a result, the number of rows of the corresponding correlation matrices $\mathbf{C}^{(m_i)}$ may differ. To bring these matrices to the same size we use a shared global keyword vocabulary $\mathcal{U}$ within each dataset group. For each model m_i, we construct $\mathbf{C}^{(m_i)}$ over the same vocabulary $\mathcal{U}$. If keyword u is present in the keyword set extracted for model m_i, its BEE scores are inserted into the corresponding row of $\mathbf{C}^{(m_i)}$. If u is absent, that row is set to zero:

$$\mathbf{C}^{(m_i)}_{u,y} = \begin{cases} \text{BEE}_{m_i}(u, y), & \text{if keyword } u \text{ is present for model } m_i, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

This zero-fill alignment yields correlation matrices with a common shape, $|\mathcal{U}| \times |\mathcal{Y}|$, enabling direct comparison of trained models from the perspective of their learned keyword-class correlations.

Semantic Correlation Descriptor (SCD). For each dataset D_i, we define its correlation descriptor as the vectorized correlation map of the corresponding single-dataset model m_i: $d^{(i)}_{\text{SCD}} = \text{vec}\big(\mathbf{C}^{(m_i)}\big)$ By construction, $d^{(i)}_{\text{SCD}}$ captures the keyword-class correlation pattern learned from D_i in isolation. This pattern includes the spurious correlations induced by the dataset, and serves as a dataset-specific reference for comparing learned correlation structures.

3.2 SCD sensitivity analysis in a controlled setting

Before defining the final membership score, we first test whether SCDs capture dataset-specific correlation patterns in a controlled setting. For each target dataset D_i, we compare the full model m_{all} with the leave-one-dataset-out model m_{-i}. The goal of this analysis is not to define the final membership inference protocol, but to verify whether the change between these two models aligns

with the standalone descriptor of D_i, rather than with the descriptors of other datasets. We estimate the dataset-specific change associated with D_i as

$$\Delta_i = d_{\mathrm{SCD}}^{(\mathrm{all})} - d_{\mathrm{SCD}}^{(-i)}. \tag{3}$$

Intuitively, Δ_i captures the change in the learned keyword-class correlation structure associated with D_i. If D_i induces a distinctive correlation pattern, then this estimated change should be more similar to the standalone descriptor of D_i than to the descriptors of other datasets. We measure this alignment using Pearson correlation. For each target dataset D_i and each candidate dataset D_j, we define

$$s_\Delta(i,j) = \mathrm{Pearson}\left(\Delta_i, d_{\mathrm{SCD}}^{(j)}\right) = \mathrm{Pearson}\left(d_{\mathrm{SCD}}^{(\mathrm{all})} - d_{\mathrm{SCD}}^{(-i)}, d_{\mathrm{SCD}}^{(j)}\right). \tag{4}$$

This score induces a binary test over dataset pairs (i,j): pairs with $j = i$ correspond to matching descriptors, while pairs with $j \neq i$ correspond to non-matching descriptors. A simple threshold classifier can therefore be defined as $\hat{z}_{i,j} = \mathbf{1}\left[s_\Delta(i,j) > \tau\right]$, where $\hat{z}_{i,j} = 1$ predicts that Δ_i matches the descriptor of D_j, and $\hat{z}_{i,j} = 0$ otherwise.

The resulting matrix $S_\Delta \in \mathbb{R}^{N \times N}$ measures how well each estimated change matches each standalone dataset descriptor. Diagonal dominance, $s_\Delta(i,i) > s_\Delta(i,j)$ for all $j \neq i$, indicates that the change associated with D_i is best explained by the fingerprint of D_i itself. In Sec. 4, we use this controlled analysis to validate the dataset-specificity of SCDs: the threshold classifier based on s_Δ perfectly separates matching from non-matching pairs, achieving a ROC-AUC of 100% in Fig. 2. This controlled analysis is not intended as the final membership inference protocol, since it assumes access to both m_{all} and m_{-i}. Rather, it serves as a diagnostic test of whether SCDs respond to dataset-specific changes in the learned correlation structure. In the next section, we consider a more realistic setting and define a membership score that does not require leave-one-dataset-out models.

3.3 SCD-based dataset membership inference

Instead of requiring paired models trained with and without a target dataset D_i, we assume access only to a trained model m_{any} and to the single-dataset reference descriptor of D_i. The goal is to decide whether D_i was included in the training data of m_{any}.

The key idea is that, if m_{any} was trained on D_i, then the dataset-specific correlation pattern captured by the descriptor of D_i should also be visible in the correlation descriptor of m_{any}. We therefore compare the two descriptors directly using Pearson correlation:

$$s_{\mathrm{SCD}}(i) = \mathrm{Pearson}\left(d_{\mathrm{SCD}}^{(\mathrm{any})}, d_{\mathrm{SCD}}^{(i)}\right). \tag{5}$$

Higher values of $s_{\mathrm{SCD}}(i)$ indicate stronger evidence that the correlation pattern associated with D_i is present in m_{any}. This score essentially measures how strongly the descriptor of the target dataset matches the descriptor of the model under inspection. Unlike s_Δ, it does not require training or accessing models explicitly constructed with D_i and without D_i, making it suitable for real-life membership inference settings. We use s_{SCD} to define a threshold-based membership classifier:

$$\hat{id}_{\mathrm{SCD}}^{(i)} = \mathbf{1}\left[s_{\mathrm{SCD}}(i) > \tau\right], \tag{6}$$

where $\hat{id}_{\mathrm{SCD}}^{(i)} = 1$ predicts that D_i is included in the training data of m_{any}, and $\hat{id}_{\mathrm{SCD}}^{(i)} = 0$ otherwise. In Sec. 4, we evaluate this classifier by measuring the ROC-AUC obtained from s_{SCD} in Fig 3.

4 Experiments

Dataset groups. We evaluate dataset-level attribution and membership inference on three dataset groups: native language identification (NLI), emotion classification, and medical section classification. The emotion group contains four heterogeneous datasets: EmoLit [Rei and Mladenić, 2023], EmotionDataset20 [Pulle, 2025], GoEmotions [Demszky et al., 2020], and a merged XED [Öhman et al., 2020] /SMED [Ganguly, 2023] corpus. The NLI group contains three datasets of English text written by non-native speakers: Reddit-L2, EFCAMDAT, and iTalki. The medical section classification group is derived from the PubMed 200k RCT corpus [Dernoncourt and Lee, 2017].

In every group, all datasets are mapped to a shared set of common classes so that keyword-class association maps are comparable across models trained on different dataset subsets.

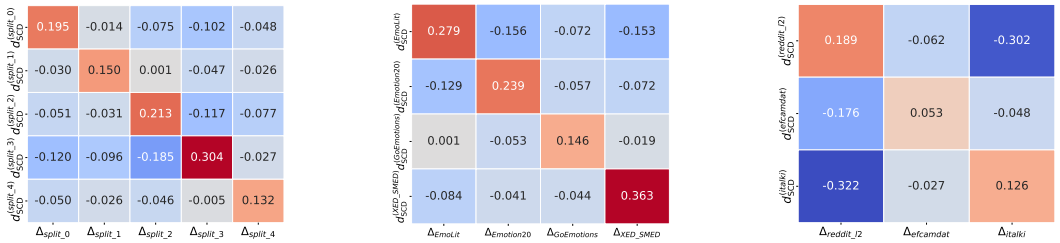
Emotion dataset group. The emotion group combines four sources that differ in genre, granularity, and annotation protocol. **EmoLit** is treated as multi-label over fine-grained emotion dimensions: after deduplicating sentences, we assign each example the emotion dimension with the largest real-valued score and then map that winner to a coarse label using a fixed many-to-one scheme. **GoEmotions** merges the train, development, and test splits; we drop duplicate comments, and for multi-labeled rows we take the *first* emotion id in the comma-separated list as the provisional fine-grained tag before mapping to the same coarse inventory. **EmotionDataset20** uses the CSV text field and single per-sentence emotion string; labels are normalized (lowercased) and mapped with the same coarse taxonomy. **XED** (English) supplies integer ids that may enumerate several emotions per line. We expand these into one row per id, map ids to string emotion names, and concatenate the result with **SMED**, after normalizing SMED's string labels (strip, title case). The XED/SMED union is deduplicated on raw text. All four mapped corpora are concatenated and deduplicated again on `text` (first occurrence kept) to form a single pooled emotion dataset, while leave-one-corpus-out training files are built by concatenating any three of the four mapped sources with the same text-level deduplication (see App. A.2 for details). Throughout, fine-grained original labels are collapsed into nine shared classes: *Angry*, *Anticipation*, *Disgust*, *Fear*, *Happy*, *Neutral*, *Sad*, *Surprise*, and *Trust*, so that keyword-class association rules remain comparable across models trained on different subsets.

NLI datasets. The NLI evaluation relies on three foundational datasets: Reddit-L2 [Rabinovich et al., 2018], EFCAMDAT [Geertzen et al., 2014], and iTalki [Hudson and Jaf, 2018]. To avoid trivial membership inference signals caused by unique labels, we filter the corpora to retain only the five first-language (L1) classes shared by all three sources: French, German, Italian, Spanish, and Turkish. To ensure the models are not biased toward the larger corpora, we use stratified subsampling to Reddit-L2 and EFCAMDAT, reducing them to the size of the smallest dataset (iTalki), as detailed in App. A.2, while preserving original class distributions. Due to its prohibitive size, Reddit-L2 is subsampled by a chunked, two-pass approach and EFCAMDAT is processed in-memory.

Medical datasets. Unlike the emotion and NLI groups, which inherently consist of multiple distinct datasets, the medical dataset group is a five-way split created from PubMed 200k RCT, a single large biomedical corpus containing over 2 million sentences. Following prior work on clustering-based splits [Firca et al., 2025], we partition it into semantically distinct topic clusters expected to induce clearer dataset fingerprints. First, to accommodate the computational footprint of the 14-million parameter LLM, Pythia [Biderman et al., 2023], used in our experiments, we extract a 12% stratified subsample of the original dataset. This subsample preserves the original class distribution of the rhetorical roles (i.e., BACKGROUND, METHODS, RESULTS, CONCLUSIONS, OBJECTIVE). To ensure the data is split based on overarching medical topics rather than sentence-level syntax, we aggregate the individual sentences using their `abstract_id` to reconstruct the full text of each abstract. We apply TF-IDF [Spärck Jones, 1972] (Term Frequency-Inverse Document Frequency) to these reconstructed abstracts to extract features, removing standard English stop words and limiting the vocabulary to the top 10,000 terms. We then apply the MiniBatch K-Means [Sculley, 2010] algorithm to the resulting TF-IDF matrix to group the abstracts into five distinct topic clusters. Crucially, we map these cluster assignments back to the original sentence-level data. This ensures that all sentences belonging to a single abstract are kept together in the same cluster, strictly preventing data leakage across different splits.

Keyword extraction. Keywords are extracted independently for each trained-model configuration using YAKE [Campos et al., 2020]. We use YAKE with maximum n-gram length $n = 3$, `dedupLim=0.9`, and 5000 keywords per model configuration, the resulting keyword pool containing unigrams, bigrams, and trigrams. We filter class-name keywords and empty or invalid tokens. For each dataset group, we construct a global keyword pool $\mathcal{U}$ as the union of all extracted keywords across the full, leave-one-out, and single-dataset models in that group.

Implementation details. We use the Pythia-14M architecture (`EleutherAI/pythia-14m`) as the base model. Each model is first finetuned as a causal language model and then used to initialize a sequence-classification model. The classifier uses the `GPTNeoXForSequenceClassification` head, exposed as `model.score`; the class directions used by BEE are the rows of `model.score.weight`. Both stages are trained for three epochs with learning rate 2×10^{-5}, weight decay 0.01, epoch-level evaluation and checkpointing, and mixed precision when CUDA



(a) Medical datasets group. (b) Emotion datasets group. (c) NLI datasets group.

Figure 2: Dataset-specificity of correlation descriptors in the controlled difference-based setting. Each heatmap shows the S_Δ matrix defined in Sec. 3.2, with rows corresponding to target datasets and columns corresponding to standalone dataset descriptors, defined in Eq. 3. Across different dataset groups, the strongest score in every row lies on the diagonal, showing that each estimated dataset-specific change is best matched by the descriptor of the corresponding dataset. Treating diagonal entries as positives and off-diagonal entries as negatives yields perfect separation, with a ROC-AUC of 100%.

is available. All preprocessing, sampling, training, and evaluation use fixed random seed 42. For each dataset group, we train individual, leave-one-dataset-out, and full-corpus models, generating 27 training configurations in total, as reported in App. A.2.

Deployment assumptions. Deploying idSCD assumes: *(i) Model access*: white-box access to the inspected model, specifically its last-layer embeddings and classification-head weights, required to construct its SCD. In our implementation, these correspond to the last-token hidden state $\mathbf{e}_u^{(m)}$ for each keyword u and the class directions $\mathbf{w}_y^{(m)}$ from the sequence-classification head. *(ii) Target reference*: access to the target dataset to construct its standalone SCD by training a reference model from the same pretrained checkpoint as the inspected model, using the same SCD construction procedure. The deployment score s_{SCD} directly compares the inspected and target-reference SCDs, without requiring a leave-one-dataset-out version of the inspected model. *(iii) Shared label space*: candidate datasets must share, or be mapped to, a common label space, making their SCD similarity scores directly comparable and enabling calibration and interpretation of the target dataset's membership score. *(iv) Semantic support*: candidate datasets must provide sufficient shared keyword support, as low overlap reduces the semantic evidence and may weaken the membership signal. Keyword overlap can be measured a priori, before model training, from the raw datasets using the keyword extraction procedure (see App. A.3).

Computational costs. Model training was run on NVIDIA A100 GPUs with 40GB VRAM on Google Colab, while dataset processing and keyword extraction used high-RAM CPU workers with 40GB RAM. Reproducing the main experiments requires approximately 42 GPU-hours: 19 hours for the emotion group, 9 hours for NLI, and 14 hours for the medical group. Reproducing the baseline comparisons requires an additional 30 GPU-hours.

Ablations and design choices. We explored several variants along the pipeline before settling on the final setup. At the preprocessing level, we varied YAKE keyword extraction by changing the n-gram length ($n = \overline{1,5}$), the number of extracted keywords per dataset ($200, 500, 1000, 2000, 5000$), and class-conditioned extraction strategies such as selecting the top $\overline{10,100}$ keywords per class. We also tested alternative medical split constructions, including PCA followed by K-Means, before adopting TF-IDF with K-Means, and initially treated XED and SMED as separate emotion sources before merging them into a single dataset. At the descriptor construction level, we evaluated variants without zero-fill alignment, variants that did not merge matrices into a shared vocabulary space, and alternative BEE scoring choices, including softmax-temperature variants with $T = \overline{1,3}$. At the scoring level, we tested cosine similarity instead of Pearson correlation, as well as direct Pearson correlation over logits rather than BEE-based descriptors. These ablations guided the final design: shared-vocabulary zero-fill alignment, BEE-based semantic correlation descriptors, and Pearson-based SCD scoring.

4.1 Correlation Descriptors are dataset-specific

We evaluate here whether the correlation descriptors defined in Sec. 3.1 are sensitive to dataset-specific changes in the training mixture. This experiment is designed as a controlled diagnostic for the intuition in Sec. 3.2: if a dataset D_i contributes a characteristic correlation pattern, then the difference-based descriptor Δ_i from Eq. 3 should align most strongly with the standalone descriptor of D_i, rather than with the descriptors of other datasets.

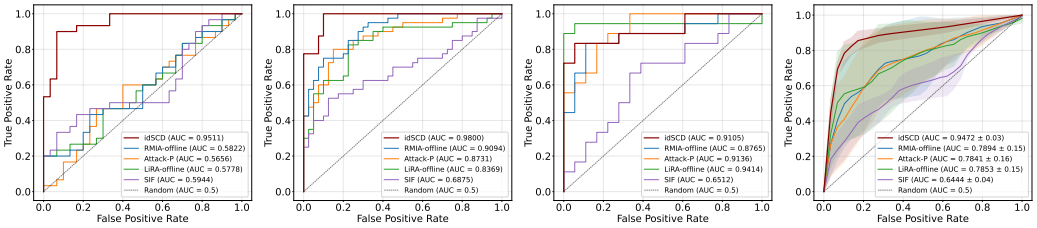
Results. Fig. 2 shows the resulting S_Δ matrices for the NLI, emotion, and medical dataset groups. The pattern is consistent across all groups: the largest score in each row appears on the diagonal. This means that the estimated change associated with each target dataset is best matched by that dataset's own correlation descriptor. Off-diagonal entries are substantially lower and often negative, indicating that these changes do not align with unrelated dataset descriptors. We further quantify this separation by treating diagonal entries ($i = j$) as positives and off-diagonal entries ($i \neq j$) as negatives. A simple threshold classifier over $s_\Delta(i, j)$ perfectly separates the two classes, achieving a ROC-AUC of 100%. This confirms that, in this controlled setting, correlation descriptors are not merely capturing generic training artifacts. They encode dataset-specific correlation structure strongly enough to recover which dataset accounts for the observed change in the model. The experiments were repeated with five random seeds, and the results are highly stable across all dataset groups as seen in A.4

Takeaway. This experiment validates the core mechanism behind our approach: dataset-specific correlation patterns are measurable, separable, and recoverable from trained models. At the same time, this is a controlled diagnostic setting because it assumes access to the leave-one-dataset-out model m_{-i}. In the next section, we move to a more realistic membership inference setting, where we do not rely on leave-one-dataset-out models.

4.2 SOTA Comparison for Membership Inference

We next compare the practical SCD-based membership classifier from Sec. 3.3 against established membership inference baselines. Unlike the controlled diagnostic, this setting does not rely on leave-one-dataset-out differences. The goal of this evaluation is to test whether id_{SCD} can distinguish models that were trained with a target dataset from models where that dataset was absent. For each target dataset D_i, member observations are computed with the full model m_{all} and non-member observations with the leave-one-dataset-out model m_{-i}. In both cases, we obtain one dataset-level observation by correlation the full per-sample score vectors, and one class-level observation for each label $y \in Y_i$ by restricting the score vectors to the samples of class y. Each target dataset therefore contributes $1 + |Y_i|$ member observations and $1 + |Y_i|$ non-member observations. We compute ROC-AUC by varying the decision threshold of id_{SCD}, and compare against recent strong membership inference baselines: black-box methods RMIA, Attack-P, and LiRA, which rely on output-derived membership signals such as loss, confidence and likelihood-ratio, and the white-box SIF baseline, which uses model-internal information to compute self-influence scores. For a fair comparison, all baselines use the same dataset-level fingerprint protocol, with each model's per-sample membership score vector serving as its fingerprint $s(m)$. For each target dataset D_i, we compute Pearson correlations over both the full vectors and each class-restricted subset. Member and non-member observations are $\mathrm{Pearson}(s(m_{all}), s(m_i))$ and $\mathrm{Pearson}(s(m_{-i}), s(m_i))$, respectively, where m_{all}, m_i, and m_{-i} denote the full, single-dataset, and corresponding leave-one-dataset-out models.

Results. Fig. 3 shows that our SCD-based classifier achieves the strongest performance on the medical and emotion groups, and remains competitive on NLI. The largest gains appear on the medical group, where id_{SCD} improves over all baselines by more than 60% (see the discussion below for why this is the case). On emotion, id_{SCD} also achieves the best performance, improving over the strongest baseline by about 8% and reaching a ROC-AUC of 0.980. On NLI, id_{SCD} remains competitive, but is no longer the top method: it is about 3% below the best baseline. As discussed below, this gap is explained by the much smaller shared semantic support in the NLI setup, a condition that can be detected before applying our method and used to assess whether SCD-based membership inference is suitable for a given dataset group. Aggregated across groups, id_{SCD} achieves the highest mean ROC-AUC and the lowest std, indicating a stronger and more stable membership signal. This suggests that SCDs are most effective when datasets induce distinct semantic correlation patterns, while baselines are more sensitive to non-semantic dataset artifacts.



(a) Medical datasets group. (b) Emotion group. (c) NLI datasets group. (d) Mean across groups.

Figure 3: SOTA comparison of ROC curves for dataset membership inference. (a-c) show performance on each dataset group, while (d) summarizes the mean and standard deviation (std) across groups. We compare against black-box baselines RMIA, Attack-P, and LiRA, and white-box SIF baseline. We achieve the top performance on the medical and emotion groups, are competitive on NLI, and obtain the highest average performance with lowest std. In contrast, the baselines fluctuate more across groups, suggesting they are more sensitive to non-semantic dataset-specific artifacts.

Why medical benefits from SCDs. The strongest improvement (+60% in ROC-AUC) appears on the medical dataset group, and this is expected from the construction of the benchmark. Unlike NLI and emotion, the medical group is built by splitting each class in PubMed 200k RCT into five topic clusters using high-level semantic features. This creates datasets that differ substantially in their semantic content while preserving the same rhetorical-role label space. Such a construction directly favors methods that use internal semantic correlations, because the resulting splits induce more separated feature-class semantic association patterns. In contrast, output-based attacks may not fully capture these high-level semantic differences if the model's predictive behavior remains similar across clusters. The medical result therefore highlights the main advantage of SCDs: when datasets differ semantically, their internal correlation fingerprints can be more discriminative than output behavior.

Limited SCD suitability on NLI. The smaller margin on NLI is also informative. Our method compares semantic correlations over a keyword vocabulary shared across datasets, so the effective decision signal depends on how much semantic overlap is exposed through shared keywords between the target dataset and the inspected model's descriptor. In the NLI group, the keyword overlap between two datasets is only 15%, while in the other groups is 20%. This leaves the SCD classifier with less shared semantic evidence and explains why the SCD signal is less dominant on NLI: the method is designed to exploit correlations between shared keywords and classes, and its discriminative power decreases when the common keyword support is sparse. See App. A.3 for details.

Takeaway. These results suggest that semantic correlation structure can substantially change how dataset-level membership is detected: when datasets induce distinct semantic fingerprints, gains over output-based and influence-based baselines can be large. Across our three evaluation groups, SCDs are strongest when datasets are semantically separable and share enough keyword support, while the baselines oscillate much more, suggesting higher sensitivity to non-semantic dataset-specific artifacts. Importantly, shared-keyword support can be checked a priori, providing a simple diagnostic for whether SCD-based inference is sufficiently semantically informed for a given dataset group.

5 Limitations

White-box access. Our strongest instantiation of the method uses internal model information, including class weights and embedding-space associations. This is a limitation for fully closed models, where such quantities are not directly available. However, this assumption matches many realistic auditing settings, including internal model audits, open-weight models, model handoffs, and regulated provenance checks where the auditor can inspect model parameters. Moreover, the framework is not intrinsically tied to this exact signal: the correlation descriptor can be instantiated with alternative scoring functions, including output-based probes or descriptors extracted from distilled surrogate models. These variants may provide weaker evidence, but they offer a path toward applying the same dataset-level tracing principle beyond strict white-box access.

Dataset identity versus dataset similarity. Our method operates at the level of dataset descriptors, not individual training examples. A strong match should therefore be interpreted as evidence that the model contains a dataset-level correlation pattern associated with D_i, not as a sample-level proof that every example in D_i was used during training. This distinction matters when datasets share domain, style, label semantics, or annotation artifacts: a high score may reflect genuine use of D_i, but it may also arise from a closely related dataset with a similar correlation structure. The leave-one-dataset-out analysis tests whether the recovered contribution is better explained by D_i than by other datasets, which reduces this ambiguity within the evaluated dataset groups. Still, absolute dataset identity is difficult when two datasets are near duplicates or intentionally constructed to share the same artifacts.

Dependence on shared keyword support. SCD-based inference relies on comparing semantic correlations over keywords that are shared across datasets. When the overlap between dataset-specific keyword vocabularies is sparse, the descriptor comparison has less semantic evidence to exploit, and the membership signal can become weaker. This limitation is partly diagnosable before running the method: low shared-keyword overlap indicates that the dataset group may be less suitable for SCD-based membership inference. In our experiments, this effect is visible in the NLI group, where lower keyword overlap reduces the discriminative power of the SCD classifier.

Class alignment (evaluation framework limitation). Our evaluation requires datasets within a group to be mapped to a shared set of canonical labels before fingerprinting. This is primarily a constraint of the current test framework, rather than a fundamental requirement that all datasets share identical labels. In practice, many scenarios already involve groups of datasets for the same task, such as NLI, emotion classification, or medical text classification, where such mappings are meaningful. We evaluate the method across three dataset groups to avoid relying on a single label taxonomy. But the quality of canonicalization matters: ambiguous, hierarchical, or partially overlapping labels can blur correlation descriptors and make dataset-specific patterns harder to separate.

Scalability of controlled validation (evaluation framework limitation). The leave-one-dataset-out protocol used in our controlled sensitivity analysis requires training one model per removed dataset, which becomes expensive as the number of candidate datasets grows. This cost is a limitation of the validation protocol, not of the final membership score itself. The controlled setting is useful because it isolates the contribution of each dataset and tests whether correlation descriptors respond to dataset-specific changes. For larger collections, this ablation-based analysis could be replaced or approximated by more efficient strategies, such as subset ablations, influence-style approximations, low-rank descriptor updates, or surrogate models. Importantly, our more realistic membership scoring setting does not require training a full leave-one-dataset-out suite.

Semantic similarity vs. exact dataset membership Our method detects dataset-specific semantic correlation patterns rather than establishing training dataset identity with certainty. Leave-one-dataset-out diagnostics favor the true member within the evaluated groups, but semantically similar or near-duplicate non-members may also score highly. Without access to the original training data, dataset-level membership inference relies on statistical and semantic traces rather than exact identity.

6 Conclusion

We show that a dataset can be recognized from the semantic correlations it induces during training. Our work reframes dataset-level membership inference as a white-box semantic fingerprinting problem: instead of relying on output behavior, we trace datasets through the internal correlation structure learned by the model. We instantiate this idea with Semantic Correlation Descriptors (SCDs), show in a controlled leave-one-dataset-out diagnostic that they recover dataset-specific changes with perfect separation, and introduce a practical SCD-based membership score that only requires the inspected model's SCD and the target dataset's standalone SCD. Across NLI, emotion classification, and medical text classification, we evaluate SCDs under different levels of semantic separation and keyword support. On average, the resulting id_{SCD} classifier achieves the highest performance and the lowest std among recent strong baselines, including RMIA, Attack-P, LiRA, and SIF. The largest gains appear when dataset groups expose distinct semantic particularities, reaching up to $+60\%$ ROC-AUC. These results establish internal semantic correlations as a strong and stable signal for dataset provenance, complementary to behavioral and influence-based membership inference.

Acknowledgments and Disclosure of Funding

This work is supported in part by projects “Romanian Hub for Artificial Intelligence - HRIA”, Smart Growth, Digitization and Financial Instruments Program, 2021-2027 (MySMIS no. 334906) and "European Lighthouse of AI for Sustainability - ELIAS", Horizon Europe program (Grant No. 101120237).

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. URL https://arxiv.org/abs/1907.02893.

Martin Bertran, Shuai Tang, Michael Kearns, Jamie Morgenstern, Aaron Roth, and Zhiwei Steven Wu. Scalable membership inference attacks via quantile regression, 2023. URL https://arxiv.org/abs/2307.03694.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023. URL https://arxiv.org/abs/2304.01373.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2020.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles, 2022. URL https://arxiv.org/abs/2112.03570.

Tyler A. Chang, Dheeraj Rajagopal, Tolga Bolukbasi, Lucas Dixon, and Ian Tenney. Scalable influence and fact tracing for large language model pretraining, 2024. URL https://arxiv.org/abs/2410.17413.

Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks, 2021. URL https://arxiv.org/abs/2007.14321.

Gilad Cohen and Raja Giryes. Membership inference attack using self influence functions, 2022. URL https://arxiv.org/abs/2205.13680.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A Dataset of Fine-Grained Emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Franck Dernoncourt and Ji Young Lee. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. In *International Joint Conference on Natural Language Processing*, 2017. URL https://arxiv.org/abs/1710.06071.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. Shortcut learning of large language models in natural language understanding, 2023. URL https://arxiv.org/abs/2208.11857.

Liviu Nicolae Firca, Antonio Barbalau, Dan Oneata, and Elena Burceanu. Not all splits are equal: Rethinking attribute generalization across unrelated categories. *NeurIPSW*, 2025.

Amrita Ganguly. SMED: Social media emotion dataset. Kaggle, 2023. URL https://www.kaggle.com/datasets/gangulyamrita/social-media-emotion-dataset. Five-class emotion dataset curated from social media.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge open language database (EFCamDat). In Ryan T. Miller et al., editors, *Selected Proceedings of the 2012 Second Language Research Forum*, pages 240–254, Somerville, MA, 2014. Cascadilla Proceedings Project. Document #3100, https://www.lingref.com.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *CoRR*, abs/2004.07780, 2020. URL https://arxiv.org/abs/2004.07780.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data, 2018. URL https://arxiv.org/abs/1803.02324.

Thomas G Hudson and Sardar Jaf. On the development of a large scale corpus for native language identification. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), December 13–14, 2018, Oslo University, Norway*, volume 155 of *Linköping Electronic Conference Proceedings*, pages 115–129. Linköping University Electronic Press, 2018.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data, 2022. URL https://arxiv.org/abs/2202.00622.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations, 2022. URL https://arxiv.org/abs/2210.11369.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions, 2020. URL https://arxiv.org/abs/1703.04730.

Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated white-box membership inference, 2020. URL https://arxiv.org/abs/1906.11798.

Zheng Li and Yang Zhang. Membership leakage in label-only exposures, 2021. URL https://arxiv.org/abs/2007.15528.

Manxi Lin, Nina Weng, Kamil Mikolaj, Zahra Bashir, Morten BS Svendsen, Martin G Tolsgaard, Anders N Christensen, and Aasa Feragen. Shortcut learning in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 623–633. Springer, 2024.

Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. Membership inference attacks by exploiting loss trajectory, 2022. URL https://arxiv.org/abs/2208.14933.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. Llm dataset inference: Did you train on my dataset?, 2024. URL https://arxiv.org/abs/2406.06443.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference, 2019. URL https://arxiv.org/abs/1902.01007.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization, 2018. URL https://arxiv.org/abs/1807.05852.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, page 739–753. IEEE, May 2019. doi: 10.1109/sp.2019.00065. URL http://dx.doi.org/10.1109/SP.2019.00065.

Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments, 2019. URL https://arxiv.org/abs/1907.07355.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. XED: A multilingual dataset for sentiment analysis and emotion detection. In *The 28th International Conference on Computational Linguistics (COLING 2020)*, 2020.

Cristian Daniel Paduraru, Antonio Barbalau, Radu Filipescu, Andrei Liviu Nicolicioiu, and Elena Burceanu. Bridging explainability and embeddings: Bee aware of spuriousness. In *ICLR*, 2026.

Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P. Wellman. Sok: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 399–414, 2018. doi: 10.1109/EuroSP.2018.00035.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale, 2023. URL https://arxiv.org/abs/2303.14186.

Yuefeng Peng, Jaechul Roh, Subhransu Maji, and Amir Houmansadr. Oslo: One-shot label-only membership inference attacks, 2024. URL https://arxiv.org/abs/2405.16978.

Shreyas Pulle. 20-emotion text classification dataset, 2025. URL https://huggingface.co/datasets/shreyaspulle98/emotion-dataset-20-emotions.

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. Native language cognate effects on second language lexical choice, 2018. URL https://arxiv.org/abs/1805.09590.

Luis Rei and Dunja Mladenić. Detecting fine-grained emotions in literature. *Applied Sciences*, 13(13), 2023. ISSN 2076-3417. doi: 10.3390/app13137502. URL https://www.mdpi.com/2076-3417/13/13/7502.

Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *ACM Computing Surveys*, 56(4):1–34, November 2023. ISSN 1557-7341. doi: 10.1145/3624010. URL http://dx.doi.org/10.1145/3624010.

Alexandre Sablayrolles, Matthijs Douze, Yann Ollivier, Cordelia Schmid, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference, 2019. URL https://arxiv.org/abs/1908.11229.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models, 2018. URL https://arxiv.org/abs/1806.01246.

David Sculley. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1177–1178. ACM, 2010.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models, 2017. URL https://arxiv.org/abs/1610.05820.

Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models, 2020. URL https://arxiv.org/abs/2003.10595.

Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

Weiwei Sun, Haokun Liu, Nikhil Kandpal, Colin Raffel, and Yiming Yang. Enhancing training data attribution with representational optimization, 2025. URL https://arxiv.org/abs/2505.18513.

Jiachen T. Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. Data shapley in one training run, 2025. URL https://arxiv.org/abs/2406.11011.

Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition, 2021. URL https://arxiv.org/abs/2108.08782.

Yao-Yuan Yang, Chi-Ning Chou, and Kamalika Chaudhuri. Understanding rare spurious correlations in neural networks, 2022. URL https://arxiv.org/abs/2202.05189.

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models, 2022. URL https://arxiv.org/abs/2111.09679.

Wenqian Ye, Luyang Jiang, Eric Xie, Guangtao Zheng, Yunsheng Ma, Xu Cao, Dongliang Guo, Daiqing Qi, Zeyu He, Yijun Tian, et al. The clever hans mirage: A comprehensive survey on spurious correlations in machine learning. *arXiv preprint arXiv:2402.12715*, 2024.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting, 2018. URL https://arxiv.org/abs/1709.01604.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks, 2024. URL https://arxiv.org/abs/2312.03262.

A Appendix

A.1 Broader Impact

This work supports training-data auditing by providing a white-box signal for dataset attribution and dataset-level membership inference. Potential positive uses include dataset provenance analysis, benchmark contamination detection, licensing audits, and accountability for models trained on mixed data sources. The method is also dual-use. The same signal could be used to infer confidential or commercially sensitive information about a model's training data. For this reason, the results should be interpreted as statistical evidence rather than definitive proof of dataset use, especially in legal, privacy-sensitive, or compliance settings.

A.2 Dataset configurations and training regimes

Tab. 1 reports the dataset configurations used in our experiments. For each dataset group, we construct three types of training sets: individual datasets, leave-one-dataset-out mixtures, and the full combined corpus. The individual configurations provide standalone dataset fingerprints, the leave-one-dataset-out configurations support the controlled sensitivity analysis, and the full corpus configurations are used for dataset-level membership evaluation.

Overall, this yields 27 training configurations across the three experimental groups: 9 for emotion classification, 7 for native language identification, and 11 for medical text classification. The table also reports the number of training rows in each configuration, making explicit the scale of each model training regime used in the main experiments.

A.3 Per-class keyword overlap between single-source datasets

To better understand why our membership inference attack performs differently across dataset groups, we measure how much the discriminative vocabularies of individual source datasets overlap. A low overlap indicates that each source contributes a largely unique set of class-discriminative features, leaving less shared signal for the fingerprint-based detector to exploit when comparing models trained on different subsets.

Procedure. For each dataset group (NLI, Emotion, Medical), we extract the top 5,000 keywords independently from each single-source training set using YAKE (with n-gram size 3 and deduplication threshold 0.9). For each single-source configuration i, we then identify which of its keywords appear in the concatenated (lowercased) texts of each class c, yielding a set of per-class keyword occurrences. We denote the number of keywords from configuration i that appear in class c of its own dataset as part of this set.

To quantify the overlap between two single-source configurations i and j, we count, for each class c, how many keywords are found in the texts of class c in *both* configurations, and sum this count across all classes. This gives the raw overlap value reported in the tables below as the off-diagonal entries. The diagonal entries represent the total number of per-class keyword occurrences within each configuration's own data (i.e., self-overlap).

To obtain a normalized measure, we divide each pairwise overlap value by the average of the two corresponding diagonal (self-overlap) values. This yields a ratio between 0 and 1, where 1 would indicate identical per-class keyword coverage and 0 would indicate completely disjoint vocabularies. We then report the mean of all off-diagonal normalized values as the average pairwise overlap for each dataset group.

Results. Tab. 2 shows the normalized pairwise overlap matrix for the NLI group, Tab. 3 for the Emotion group, and Tab. 4 for the Medical group. Tab. 5 summarizes the average pairwise overlap across groups alongside the id_{SCD} ROC-AUC.

Implications. The NLI group exhibits the lowest average pairwise overlap (14.96%), meaning that its constituent sources (reddit_l2, efcamdat, italki) share only about 15% of their class-discriminative vocabulary. By contrast, the Emotion and Medical groups share approximately 20% of theirs. This aligns with the lower id_{SCD} ROC-AUC observed for NLI: when single-source datasets have less

Table 1: Dataset configurations and sample sizes across the three experimental groups. For each group, we construct individual, leave-one-dataset-out, and full-corpus training sets, yielding 27 training configurations in total.

Dataset group	Training configuration	Sample size (rows)
Emotion	EmoLit (Individual)	200,727
	EmotionDataset20 (Individual)	79,595
	GoEmotions (Individual)	53,993
	XED & SMED (Individual)	92,978
	Without EmoLit (Leave-One-Out)	226,563
	Without EmotionDataset20 (Leave-One-Out)	347,695
	Without GoEmotions (Leave-One-Out)	373,300
	Without XED & SMED (Leave-One-Out)	334,315
	Full Combined Corpus	**427,290**
NLI	iTalki (Individual)	45,204
	Reddit-L2 (Individual)	45,201
	EFCAMDAT (Individual)	45,200
	Without iTalki (Leave-One-Out)	90,401
	Without Reddit-L2 (Leave-One-Out)	90,404
	Without EFCAMDAT (Leave-One-Out)	90,405
	Full Balanced Corpus	**135,605**
Medical	Cluster 0 (Individual)	21,100
	Cluster 1 (Individual)	31,791
	Cluster 2 (Individual)	33,251
	Cluster 3 (Individual)	160,516
	Cluster 4 (Individual)	25,776
	All Except Cluster 0 (Leave-One-Out)	251,334
	All Except Cluster 1 (Leave-One-Out)	240,643
	All Except Cluster 2 (Leave-One-Out)	239,183
	All Except Cluster 3 (Leave-One-Out)	111,918
	All Except Cluster 4 (Leave-One-Out)	246,658
	Full Combined Corpus (12% Subsample)	**272,434**

Table 2: Normalized per-class keyword overlap between single-source datasets (NLI).

	reddit_l2	efcamdat	italki
reddit_l2	1.00	0.14	0.14
efcamdat	0.14	1.00	0.17
italki	0.14	0.17	1.00

overlapping discriminative features, the aligned keyword matrices used by the detector contain more zero-filled (uninformative) positions, which dilutes the similarity signal that distinguishes member from non-member configurations. In other words, less lexical overlap between sources means less information is available to inform the dataset fingerprint, making the membership inference task harder.

Importantly, this overlap can be measured *a priori*, before training any models or running the attack, using only the raw text data and an unsupervised keyword extraction method. This suggests that keyword overlap between constituent sources could serve as a practical diagnostic for anticipating the difficulty of fingerprint-based dataset membership inference in multi-source settings.

Table 3: Normalized per-class keyword overlap between single-source datasets (Emotion).

	EmoLit	Emotion20	GoEmotions	XED_SMED
EmoLit	1.00	0.12	0.24	0.17
Emotion20	0.12	1.00	0.20	0.20
GoEmotions	0.24	0.20	1.00	0.25
XED_SMED	0.17	0.20	0.25	1.00

Table 4: Normalized per-class keyword overlap between single-source datasets (Medical).

	split_0	split_1	split_2	split_3	split_4
split_0	1.00	0.17	0.12	0.16	0.13
split_1	0.17	1.00	0.25	0.27	0.25
split_2	0.12	0.25	1.00	0.18	0.19
split_3	0.16	0.27	0.18	1.00	0.27
split_4	0.13	0.25	0.19	0.27	1.00

Table 5: Average pairwise normalized keyword overlap (single-source pairs only) and id_{SCD} membership inference ROC-AUC for each dataset group.

Dataset Group	Single Sources	Avg. Pairwise Overlap (%)	id_{SCD} ROC-AUC
NLI	3	14.96	0.9105
Emotion	4	19.62	0.9800
Medical	5	20.01	0.9511

A.4 Multi-seed experiments

To test the stability of our results, we repeat the main experiment over 5 independent random seeds (0, 1, 2, 3, 42). The seed controls model initialization, data shuffling, and the sampling of member/non-member datasets. All other hyperparameters are kept identical to those in Section 4. Table 6 reports the mean and standard deviation across seeds.

The results are consistent across seeds, confirming that the conclusions of Section 4 do not depend on a particular random seed.

Table 6: Results over 5 random seeds (mean $\pm$ std).

Dataset group	ROC-AUC
Emotion	0.990 ± 0.006
NLI	0.914 ± 0.007
Medical	0.955 ± 0.021

A.5 Quantitative Descriptor Size Ablations

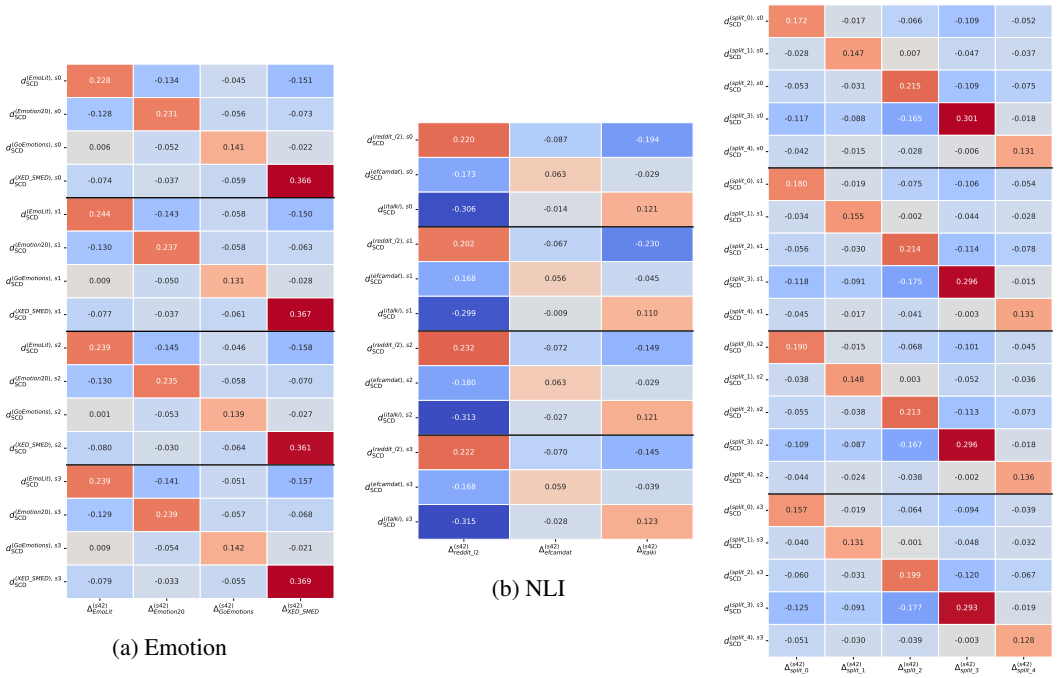
We vary the number of extracted keywords per dataset (k), which changs the size of the shared keyword vocabulary ($|U_k|$) and consequently the descriptor dimensionality ($|U_k| \times |Y|$). Across all three dataset groups, increasing the descriptor dimensionality consistently improves ROC-AUC before largely saturating, as reported in Table 7

Table 7: Descriptor dimensionality ablation results

Dataset group	k	Descriptor dimension	ROC-AUC
Emotion	10	396	0.740
Emotion	50	1818	0.913
Emotion	100	3528	0.968
Emotion	500	17757	0.965
Emotion	1000	35136	0.977
Emotion	**5000**	**169425**	**0.980**
NLI	10	190	0.824
NLI	50	915	0.818
NLI	100	1835	0.855
NLI	500	8650	0.880
NLI	1000	16995	0.886
NLI	**5000**	**81550**	**0.910**
Medical	10	265	0.737
Medical	50	1145	0.856
Medical	100	2255	0.878
Medical	500	10935	0.877
Medical	1000	21545	0.918
Medical	**5000**	**104630**	**0.951**

A.6 Independently seeded descriptor controls

To evaluate the fingerprint robustness cross-seed, we compute reference descriptors from independently seeded single-dataset models with multiple independent random seeds while keeping the full and leave-one-dataset-out models fixed. We then recomputed the controlled diagnostic by replacing each reference descriptor with the descriptor obtained from the independently seeded model and regenerated the corresponding matrices. We use the descriptors obtained with seed 42 as a fixed reference and compare them with descriptors learned independently using seeds $s \in \{0, 1, 2, 3\}$. The resulting correlation structure is highly consistent across the four comparison seeds. Both the sign and magnitude of the correlations vary only slightly from one seed to another. In particular, the diagonal



(a) Emotion

(b) NLI

(c) Medical

Figure 4: Cross-seed descriptor reproducibility relative to seed 42. Each cell reports the correlation between a descriptor learned with a comparison seed $s \in \{0, 1, 2, 3\}$ (rows) and a descriptor learned with the reference seed 42 (columns). Diagonal entries compare corresponding descriptors across seeds, while off-diagonal entries capture cross-descriptor relationships. Horizontal lines separate the four comparison seeds.

correlations remain positive for all descriptors, while the off-diagonal relationships exhibit nearly identical patterns across seeds. This suggests that, although the descriptors obtained from different random initializations are not identical, their relationship to the seed-42 reference representation is reproducible. This shows that the alignment between Δ_i and $d^{(i)}_{SCD}$ is robust to independent training initializations and reflects stable dataset-specific semantic fingerprints rather than a seed-specific construction artifact.

A.7 IID and Non-IID Partitions

We test whether the SCD fingerprint depends on the specific training examples or on the data distribution. To do so, we split each dataset D_i into two disjoint parts and train a model on each part. Training uses the same procedure, hyperparameters and seed as the single-dataset models. We use two partitioning schemes:

- **IID:** a random 50/50 split. Both halves preserve the class distribution of D_i.
- **Non-IID:** each class c is sampled at its own rate r_c, spaced linearly from 0.02 to 0.98 across the nine classes. A fraction r_c of class c goes to the first part and $1 - r_c$ to the second. The two parts therefore have strongly skewed, roughly mirrored class distributions; for example, *Trust* makes up under 1% of the second part. Because the rates average 0.5, the two parts remain similar in size.

To keep all descriptors in the same space, each part model reuses its parent dataset's keyword list. We then correlate each part model's descriptor with the difference descriptors Δ_j computed from the original full-data models.

Figures 5 and 6 show the results. Under both schemes, every part remains diagonally dominant for all four datasets, and the off-diagonal structure is essentially unchanged. The mean diagonal correlation

is 0.257 for the full data. For the IID halves it is 0.241 and 0.239. For the class-skewed non-IID parts it is 0.240 and 0.236, close to the IID results even though some classes are almost absent. The fingerprint therefore behaves mainly as a property of the data distribution rather than of individual examples. Note that m_{all} was trained on the full D_i, so both parts are members with respect to Δ_i. The experiment shows that a subset reproduces the dataset's signature; it does not compare examples a model was trained on against examples it never saw.

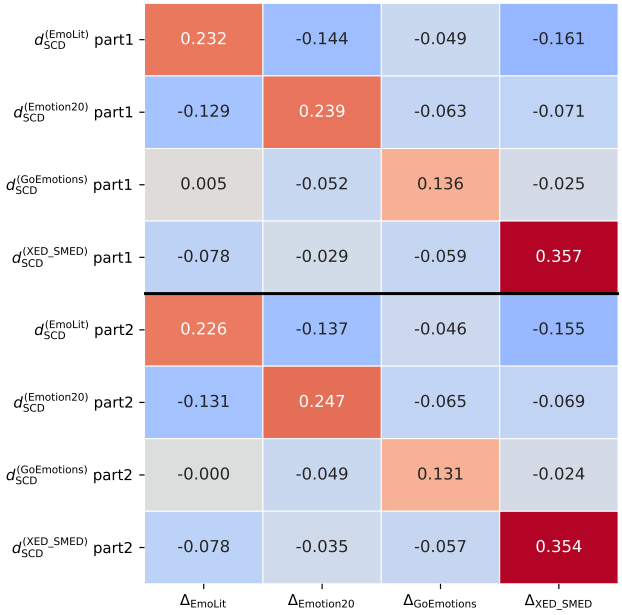


Figure 5: IID split. Correlation between Δ_i (columns) and the descriptors of models trained on the full dataset (top block) or on each random half (middle and bottom blocks).

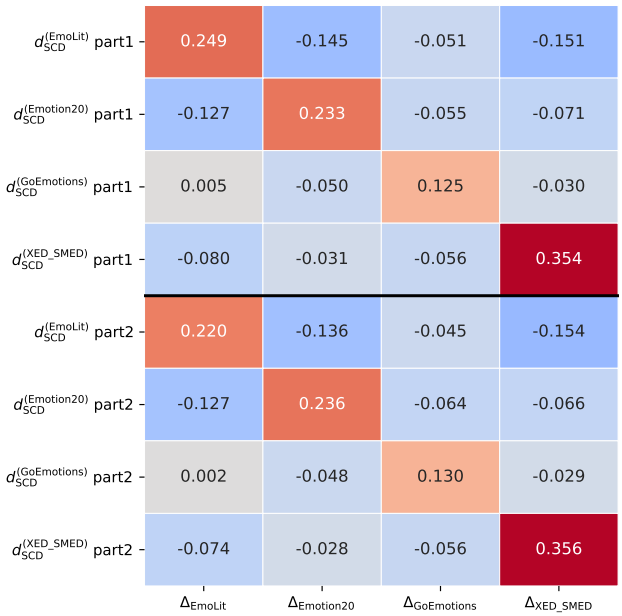


Figure 6: Non-IID split. Same as Figure 5, but the middle and bottom blocks use models trained on the two class-skewed parts of each dataset.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state the main contribution as dataset tracing through learned association fingerprints, and the experiments evaluate attribution and dataset-level membership inference under this scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The paper includes a Limitations section discussing the controlled auditing setting, class alignment, fingerprint distinctiveness, keyword coverage, white-box access, scalability, and scope of evaluation.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A].

Justification: The paper does not present formal theoretical results or proofs. The proposed method is defined through correlation descriptors and empirical scoring procedures, and its claims are evaluated experimentally.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: The paper and supplemental material specify the datasets, preprocessing, model architecture, training protocol, keyword extraction method, correlation descriptor construction, scoring functions, and evaluation metrics needed to reproduce the main experimental results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: The experiments use publicly available datasets, and the paper and supplemental material describe the experimental protocol, preprocessing, model training, descriptor construction, and evaluation metrics. The code is publicly available at the repository linked in the paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The paper specifies the dataset groups, label-space alignment, model architecture, training setup, keyword extraction procedure, construction of correlation descriptors, scoring metrics, and evaluation protocols.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No].

Justification: The paper reports aggregate metrics such as ROC-AUC and pairwise score matrices, but does not currently report error bars, confidence intervals, or statistical significance tests across repeated runs. We discuss the controlled nature of the evaluation and the scope of the conclusions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the compute resources used for the experiments, including the GPU type, memory, and approximate aggregate compute for the main experiments and baseline comparisons.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work uses public datasets and studies model auditing and dataset provenance. We do not collect sensitive data or conduct human-subject experiments.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: The paper includes a Broader Impact section discussing positive uses for dataset provenance, contamination detection, licensing audits, and accountability, as well as dual-use risks related to inferring confidential or commercially sensitive training data.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release high-risk models, scraped datasets, image generators, or other assets requiring special safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes].

Justification: The paper cites the existing datasets, models, and tools used in the experiments. All assets are used consistently with their intended research use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A].

Justification: The paper does not release new datasets, models, or code at submission time. It evaluates a method using existing public datasets and tools.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A].

Justification: The paper does not involve crowdsourcing or human-subject experiments.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A].

Justification: The paper does not involve human subjects, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A].

Justification: The core method does not rely on an LLM as an important or non-standard research component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.