Privacy Risks of Intermediate Representations: Attribute Inference in Distributed LLM Inference

Anonymous ACL submission

Abstract

Distributed LLM inference avoids sending raw inputs by transmitting intermediate hidden states, a practice widely assumed to preserve privacy. We challenge this assumption and demonstrate that *intermediate representations alone* are sufficient to leak sensitive user attributes. This setting poses a fundamental obstacle for existing attribute inference attacks, which typically rely on auxiliary embedding–attribute pairs. To characterize this previously underexplored privacy risk, we reformulate attribute inference as zero-shot semantic similarity matching directly in the intermediate representation space, and introduce a purely intermediate-representation-based attribute inference attack, termed IR-AIA. To address two structural challenges that hinder attribute inference from intermediate representations, we propose SG-APCR to address layer-dependent anisotropy in intermediate embeddings and a sliding-window similarity matching strategy to handle subword-level semantic fragmentation. Experiments across three LLMs and three real-world datasets show that sensitive attributes can be reliably inferred using only intermediate representations, achieving Top-1 accuracy of up to 0.997 on CMS, 0.980 on Skytrax, and 0.986 on ECHR. These results reveal that intermediate states commonly considered safe to share can expose sensitive personal attributes on their own.

1 Introduction

Large language models (LLMs) are increasingly deployed in privacy-sensitive domains such as healthcare, finance, and online services (Ravenda et al., 2025; Fan et al., 2024). To reduce computational cost and improve accessibility in these deployments, distributed inference frameworks have been proposed (Huang et al., 2025; Borzunov et al., 2023a). In such frameworks, model execution is partitioned across multiple participants, requiring the exchange of intermediate hidden states, which is commonly assumed to preserve user privacy by avoiding the transmission of raw inputs (Jiang et al., 2024). However, as these systems operate closer to end users, input prompts often contain sensitive personal attributes, such as identity or medical conditions, raising urgent concerns about the privacy risks of exposing intermediate representations (Shanmugarasa et al., 2025; Kwesi et al., 2025).

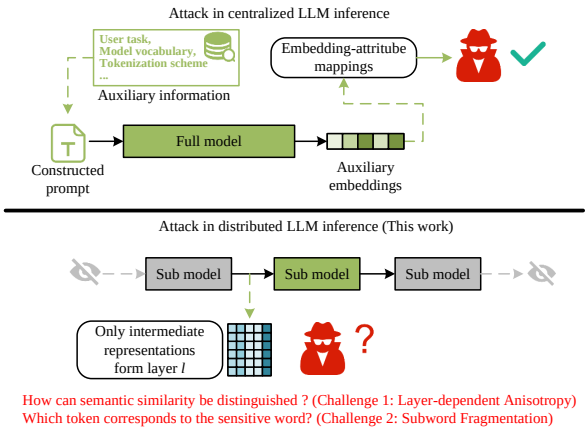

Figure 1: Comparison between centralized and distributed LLM inference for attribute inference. Prior attacks rely on full-model access and auxiliary data (top), whereas distributed inference exposes only intermediate representations, leading to layer-dependent anisotropy and subword-level semantic fragmentation (bottom).

Against this backdrop, prior works in distributed LLM inference have mainly focused on reconstruction attacks, which attempt to recover the entire prompt from intermediate representations (Luo et al., 2025; Qu et al., 2025). Such attacks are inherently goal-agnostic: they quantify how much of the prompt can be reconstructed, but do not directly indicate whether specific sensitive information is disclosed. In contrast, we study attribute inference attacks, which directly characterize the leakage risk of sensitive information (Liu et al., 2025). Here, an

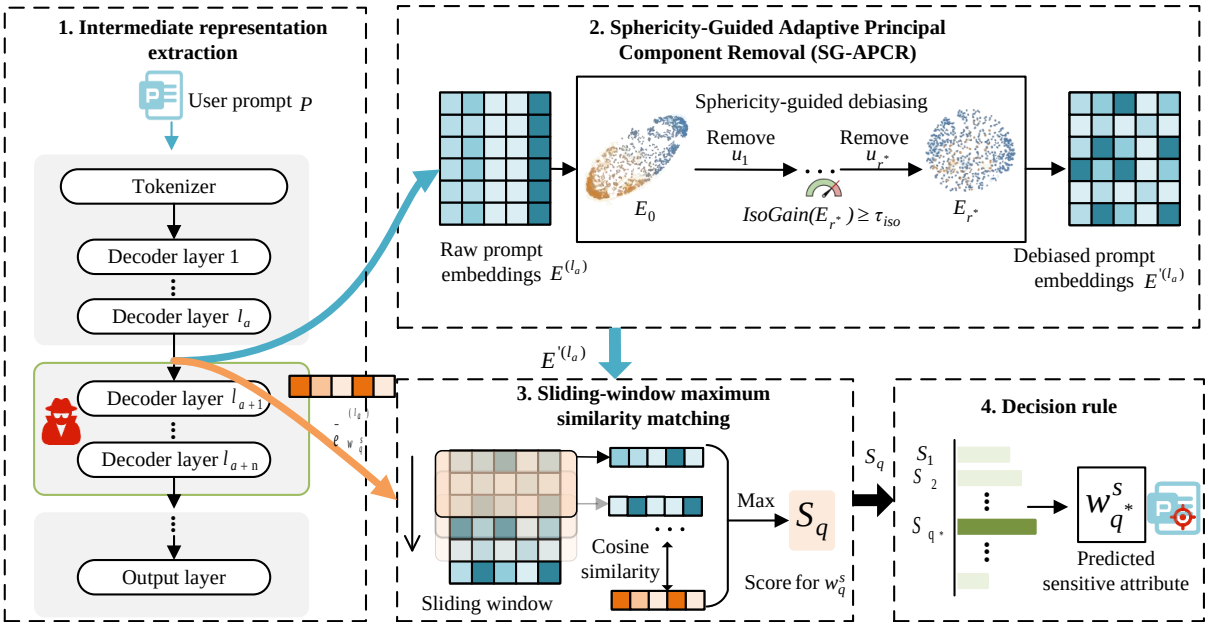

Figure 2: Overview of the proposed attack IR-AIA. The attacker observes intermediate token embeddings at layer ℓ_a, applies SG-APCR to remove dominant principal components and obtain debiased embeddings for reliable similarity comparison, and then performs sliding-window similarity matching to compute detection scores, which are finally used to predict the sensitive attribute.

attribute corresponds to the sensitive information itself in the input (e.g., a specific drug name or location) (Pan et al., 2020; Song and Raghunathan, 2020), and the adversary directly infers whether a cherry-picked target attribute is present (Jayaraman and Evans, 2022). Consequently, attribute inference reveals sensitive information even when full-text reconstruction is inaccurate or infeasible.

However, the intrinsic characteristics of distributed LLM inference fundamentally hinder existing attribute inference attacks, which are primarily developed for centralized settings. As illustrated in Figure 1, this distributed inference setting fundamentally differs from the centralized scenarios assumed in prior attribute inference attacks. In centralized LLM inference, an adversary has access to the full model and therefore typically relies on auxiliary data to construct embedding–attribute pairs for conducting attribute inference attacks (Wei et al., 2025; Sutton et al., 2025; Gu et al., 2023). By contrast, in distributed LLM inference, the adversary holds only a subset of model layers and observes only the intermediate hidden states exchanged between participants (Borzunov et al., 2023b), which precludes access to auxiliary information needed to establish embedding–attribute mappings. As a result, the privacy risks associated with intermediate representations have not yet been fully revealed in this setting. This work aims to reveal the privacy risks inherent in intermediate hidden states under distributed LLM inference, showing that they can be exploited to leak sensitive user attributes.

Given access only to intermediate representations during inference, attribute inference attacks in distributed LLM inference face the following two major challenges. **Challenge 1: Layer-dependent anisotropy of intermediate representations.** Intermediate representations exhibit strong and layer-dependent anisotropy (Skean et al., 2025; Machina and Mercer, 2024), meaning that the degree to which the embedding space is dominated by a few principal directions varies substantially across layers. This geometric distortion compresses meaningful semantic variation in a non-uniform manner, making it difficult to reliably distinguish attribute-related signals at different intermediate layers. **Challenge 2: Subword-level fragmentation of attribute semantics.** Attribute information is scattered across multiple discrete token embeddings, as words encoding sensitive attributes are often split into several subword tokens by the tokenizer (Rust et al., 2021). This fragmentation eliminates a one-to-one correspondence between embedding vectors and attribute semantics, preventing direct attribute localization.

To address the above challenges, we introduce a purely intermediate-representation-based attribute

inference attack, termed *IR-AIA*. In contrast to prior approaches, we reformulate attribute inference as a zero-shot similarity matching task, thereby avoiding the need for auxiliary data to learn embedding–attribute mappings. To counter the layer-dependent anisotropy of intermediate embeddings, we propose an unsupervised, geometry-driven mechanism termed Sphericity-Guided Adaptive Principal Component Removal (SG-APCR), which layer-adaptively restores meaningful semantic structure for reliable similarity comparison. Furthermore, to overcome semantic fragmentation across subword tokens, we introduce a window-based maximum similarity strategy that reconstructs word-level signals from local embedding windows, enabling attribute detection without relying on token boundary information. Our contributions are threefold:

- We demonstrate that sensitive user attributes can be inferred using only intermediate representations in distributed LLM inference, and systematically characterize this privacy risk through a proposed purely intermediate-representation-based attribute inference attack, termed IR-AIA.
- We identify two structural obstacles that fundamentally prevent intermediate-layer attribute inference: embedding anisotropy and multi-token semantic fragmentation. To address these challenges, we propose two corresponding mechanisms, SG-APCR and Sliding-Window Maximum Similarity Matching, which together make similarity-based attribute inference feasible.
- We empirically show that attribute leakage from intermediate hidden states is persistent across multiple layers and model families, rather than being confined to isolated depths. These results underscore non-trivial privacy risks in distributed LLM inference and challenge the common assumption that intermediate representations are inherently safe to share.

2 Threat Model

We consider an adversary who participates as an honest but curious node in a distributed inference system. The model is partitioned across multiple parties, and the adversary is assigned a subset of model layers $\{\ell_{a+1}, \ldots, \ell_{a+n}\}$. Beyond these layers, the adversary cannot modify the model or directly observe the user's input. The adversary also holds a cherry-picked set of sensitive words $W^s = \{w_q^s\}_{q=1}^M$ representing attributes of interest. Additionally, we assume that the adversary can obtain intermediate-layer representations of W^s at layer ℓ_a via limited queries to the model, without accessing any user data, auxiliary datasets, or embedding–attribute pairs. This assumption is consistent with realistic distributed LLM inference settings (Huang et al., 2025; Borzunov et al., 2023a). The adversary's objective is to output an index $q^* \in \{1, \ldots, M\}$ such that $w_{q^*}^s$ is most semantically aligned with the hidden representations observed during inference. Aside from representations of its assigned layers, the adversary has no access to user inputs, task distributions, or other model components. In summary, the adversary is semi-honest, observes intermediate hidden states required for legitimate computation, and aims to infer a sensitive attribute without auxiliary data or access to the user's input.

3 Problem Formulation

To address the challenge of inferring sensitive attributes in distributed LLM inference without access to auxiliary data for constructing embedding–attribute mappings, we formulate sensitive word inference as a semantic retrieval task in the embedding space. Given a user prompt P and a predefined sensitive word set $W^s = \{w_q^s\}_{q=1}^M$, the adversary aims to identify the word in W^s that is most semantically aligned with the hidden representations observed during distributed inference.

Let the prompt P consist of words $W = \{w_i\}_{i=1}^N$. Each word is tokenized into subword tokens,

$$w_i \xrightarrow{\text{tokenize}} \{t_1, t_2, \ldots, t_{n_i}\}, \tag{1}$$

and each token t is mapped to an embedding $e_t \in \mathbb{R}^d$. After forward propagation through the first a layers, the embedding becomes $e_t^{(\ell_a)}$, which we refer to as the intermediate embedding at layer ℓ_a.

We define $S_w^{(\ell_a)}$ as the semantic vector of word w at layer ℓ_a, obtained either by a single token representation or by aggregating intermediate embeddings of its subword tokens. Under the threat model in Section 2, which captures distributed LLM inference, the adversary has access to $S_{w_i}^{(\ell_a)}$ for tokens in the prompt and can precompute $S_{w_q^s}^{(\ell_a)}$ for all w_q^s in W^s.

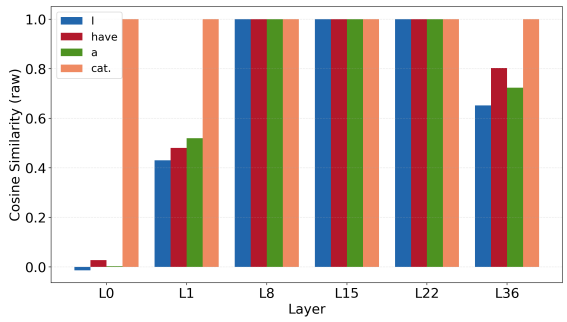

Figure 3: Layer-wise cosine similarity between each token in "I have a cat." and the token "cat" on Qwen-3B across model layers.

The cosine similarity between two semantic vectors is defined as

$$\mathrm{sim}(S_{w_i}^{(\ell_a)}, S_{w_q^s}^{(\ell_a)}) = \frac{\langle S_{w_i}^{(\ell_a)}, S_{w_q^s}^{(\ell_a)} \rangle}{\|S_{w_i}^{(\ell_a)}\|_2 \, \|S_{w_q^s}^{(\ell_a)}\|_2}. \quad (2)$$

For a given prompt, we compute a matching score for each sensitive word w_q^s as

$$\mathrm{score}(w_q^s) = \max_{w_i \in W} \mathrm{sim}\left(S_{w_q^s}^{(\ell_a)}, S_{w_i}^{(\ell_a)}\right), \quad (3)$$

where the max operator selects the most semantically aligned word in the prompt for each candidate attribute in W^s.

The final prediction selects the sensitive word with the highest matching score:

$$q^* = \arg\max_{w_q^s \in W^s} \mathrm{score}(w_q^s), \quad (4)$$

$$\mathrm{score}^* = \max_{w_q^s \in W^s} \mathrm{score}(w_q^s). \quad (5)$$

In practice, applying this similarity-based formulation in distributed LLM inference is non-trivial, due to the anisotropy of intermediate representations and the fragmentation of semantic information across subword tokens.

4 Structural Obstacles to Attribute Inference from Intermediate Representations

The formulation in Section 3 assumes that semantic similarity can be reliably measured between word representations at intermediate layers. We show that this assumption does not hold due to two structural properties of intermediate representations: layer-dependent anisotropy and subword-level semantic fragmentation.

4.1 Layer-dependent Anisotropy of Intermediate Representations

Figure 3 provides a diagnostic example illustrating the layer-dependent effect of anisotropy. Although cosine similarity is expected to reflect token-level semantic alignment, the observed collapse at certain intermediate layers implies that similarity scores are dominated by shared global directions rather than lexical content. As a result, raw cosine similarity becomes an unreliable measure for semantic matching on intermediate representations, rendering fixed, layer-agnostic similarity calibration ineffective across layers.

4.2 Fragmentation of Semantic Information Across Subword Tokens

Subword tokenization introduces a second obstacle. Many attribute-bearing words are split into multiple tokens, distributing word-level semantics across several intermediate embeddings. For example, in both GPT-2 and Qwen models, *"Amoxicillin"* is tokenized as `["Am", "oxic", "illin"]`, with no single token corresponding to the full semantic meaning of the drug. In distributed inference, this fragmentation renders direct token-level matching inherently ill-defined, particularly for multi-token attributes.

Together, anisotropy and semantic fragmentation invalidate naïve similarity matching on intermediate representations and motivate the need for explicit geometry correction and word-level reconstruction, which we address in the next section.

5 Proposed Method: IR-AIA

Motivated by the structural obstacles identified above, we propose IR-AIA, a purely intermediate-representation-based attribute inference attack that performs data-free semantic similarity matching on intermediate hidden states. The proposed IR-AIA performs data-free semantic similarity matching on intermediate hidden states and consists of four stages: (1) intermediate representation extraction, (2) Sphericity-Guided Adaptive Principal Component Removal, (3) sliding-window maximum similarity matching, and (4) a final decision rule. An overview of the attack pipeline is shown in Figure 2.

5.1 Intermediate Representation Extraction

As the first step of IR-AIA, the adversary extracts token-level intermediate representations at a desig-

nated layer ℓ_a during distributed inference. Given a prompt P, the observed embeddings are denoted by $E^{(\ell_a)} \in \mathbb{R}^{m\times d}$ for its m tokens.

5.2 Sphericity-Guided Adaptive Principal Component Removal

To address the challenge of layer-dependent anisotropy in intermediate representations, we propose *Sphericity-Guided Adaptive Principal Component Removal* (SG-APCR), an unsupervised, layer-adaptive, and model-agnostic debiasing method that removes a small number of dominant principal directions to restore meaningful semantic structure for similarity matching. Let the input prompt be P, and let $E^{(\ell_a)} \in \mathbb{R}^{m\times d}$ denote its intermediate embedding matrix at layer ℓ_a. Each row $e_{t_i}^{(\ell_a)} \in \mathbb{R}^{1\times d}$ is the embedding of token t_i. We perform SVD directly on $E^{(\ell_a)}$:

$$E^{(\ell_a)} = USV^\top, \tag{6}$$

where $V = [u_1, u_2, \ldots, u_d]$ and each $u_j \in \mathbb{R}^d$ is a principal component direction. In practice, we compute a single SVD per prompt and reuse the resulting basis for all similarity operations. Unlike standard PCA-based post-processing (Mu et al., 2017), we do not perform mean-centering before SVD. In pretrained language models, the global mean often captures semantic baselines; subtracting it can distort the geometric structure of embeddings. Working directly with $E^{(\ell_a)}$ preserves this reference while still removing dominant anisotropic directions.

To avoid manually fixing the debiasing dimension, we quantify the isotropy of the embedding space after removing the top-r principal components and use it to adaptively select r.

For a candidate r, we define:

$$U_r = [u_1, \ldots, u_r], \quad E_r = E^{(\ell_a)} - E^{(\ell_a)} U_r U_r^\top. \tag{7}$$

We compute the covariance matrix

$$\mathrm{Cov}_r = (E_r)^\top E_r / m, \tag{8}$$

and let $\{\lambda_i\}_{i=1}^d$ be the eigenvalues of Cov_r. The *Participation Ratio* (PR) is defined as

$$\mathrm{PR}_r = \frac{\left(\sum_{i=1}^d \lambda_i\right)^2}{\sum_{i=1}^d \lambda_i^2}, \tag{9}$$

and we normalise it into an isotropy score

$$\mathrm{IsoGain}(E_r) = \frac{\mathrm{PR}_r - 1}{d-1} \in [0,1]. \tag{10}$$

We search $r \in [0, R_{\max}]$, where $R_{\max} = \alpha \cdot \min(m,d)$ and $\alpha \in (0,1)$ is a search ratio, and choose the smallest r such that $\mathrm{IsoGain}(E_r) \geq \tau_{\mathrm{iso}}$ for a prescribed threshold τ_{iso}. If no r satisfies the threshold, we choose the r that maximises $\mathrm{IsoGain}(E_r)$. Since SVD is performed only once on $E^{(\ell_a)}$, all candidate E_r share the same basis U and can be obtained via inexpensive projections, yielding an unsupervised and computationally efficient debiasing procedure. This procedure also ensures reproducibility since all debiased vectors are computed from the same orthonormal basis.

Let $r^\star$ denote the selected number of components and $U_{r^\star}$ the corresponding basis. For each token embedding $e_{t_i}^{(\ell_a)}$, the debiased representation is:

$$e_{t_i}^{\prime(\ell_a)} = e_{t_i}^{(\ell_a)} - \sum_{j=1}^{r^\star} (u_j^\top e_{t_i}^{(\ell_a)}) u_j. \tag{11}$$

In matrix form, the debiased embedding matrix is

$$E^{\prime(\ell_a)} = E^{(\ell_a)} - E^{(\ell_a)} U_{r^\star} U_{r^\star}^\top. \tag{12}$$

5.3 Sliding-Window Maximum Similarity Matching

To address the challenge of subword-level semantic fragmentation in intermediate representations, we propose a *sliding-window maximum similarity matching* strategy that aggregates local token sequences to recover word-level semantic signals.

For a sensitive word $w_q^s \in \mathcal{W}_s$, let L_q denote the number of tokens in w_q^s, and let $e_j^{(\ell_a)}$ be the intermediate embedding of its j-th token at layer ℓ_a. We represent w_q^s by the averaged embedding

$$\bar{e}_{w_q^s}^{(\ell_a)} = \frac{1}{L_q} \sum_{j=1}^{L_q} e_j^{(\ell_a)}. \tag{13}$$

We then apply SG-APCR to obtain the debiased sensitive-word vector $\bar{e}_{w_q^s}^{\prime(\ell_a)}$.

Let $T = \{t_j\}_{j=1}^m$ be the token sequence of the prompt, and let $E^{\prime(\ell_a)} = \{e_{t_j}^{\prime(\ell_a)}\}_{j=1}^m$ denote the corresponding debiased intermediate embeddings. To approximate word-level semantics without segmentation, we slide a continuous window of length L_q across the sequence with stride 1.

Let the starting index of the window be $k \in \{1, \ldots, m - L_q + 1\}$. The averaged embedding of the k-th window is

$$\bar{e}_{T_k}^{\prime(\ell_a)} = \frac{1}{L_q} \sum_{j=0}^{L_q-1} e_{t_{k+j}}^{\prime(\ell_a)} \in \mathbb{R}^d, \tag{14}$$

and we compute the cosine similarity between the window average and the sensitive-word vector:

$$\mathrm{sim}\left(\bar{e}'^{(\ell_a)}_{T_k}, \bar{e}'^{(\ell_a)}_{w_q^s}\right) = \frac{\left\langle \bar{e}'^{(\ell_a)}_{T_k}, \bar{e}'^{(\ell_a)}_{w_q^s} \right\rangle}{\left\| \bar{e}'^{(\ell_a)}_{T_k} \right\|_2 \left\| \bar{e}'^{(\ell_a)}_{w_q^s} \right\|_2}. \quad (15)$$

The maximum similarity over all windows is taken as the semantic matching score for w_q^s:

$$S_q = \max_{k \in [1, m-L_q+1]} \mathrm{sim}\left(\bar{e}'^{(\ell_a)}_{T_k}, \bar{e}'^{(\ell_a)}_{w_q^s}\right). \quad (16)$$

Intuitively, S_q reflects the degree of semantic match between w_q^s and the most similar local region in the prompt.

5.4 Decision Rule

The final prediction is obtained by selecting the sensitive attribute with the highest similarity score. Let S_q denote the matching score for w_q^s, then $\hat{q} = \arg\max_q S_q$. This rule follows directly from the zero-shot similarity matching formulation and introduces no additional parameters.

6 Experiments

6.1 Experimental Setup

Models and Datasets. To assess the generality of our approach across different model scales and tokenization strategies, we evaluate three pretrained large language models: GPT-small (12 layers), GPT-large (36 layers), and Qwen3-4B-Thinking-2507 (36 layers with a tokenizer distinct from GPT). We conduct experiments on three representative real-world datasets drawn from distinct application domains: an airline review dataset (Skytrax (Bunchongchit and Wattanacharoensil, 2021)), a medical reimbursement dataset (CMS–DE (Han et al., 2023)), and a legal judgment dataset (ECHR (Chalkidis et al., 2019)).

For these datasets, we respectively treat locations, drug names, and legal Articles as sensitive attributes. Drug names are typically long and frequently segmented into multiple subword tokens, whereas legal Articles often differ only by their numerical identifiers, resulting in low lexical distinctiveness.

Evaluation metrics. We measure matching performance using the following standard metrics: Top-k hit rate (Top-1 / Top-3 / Top-5), AUC, and F1. Top-k metrics evaluate semantic ranking quality, while AUC and F1 quantify overall identification performance.

Model	Dataset	Layer	Top-1	AUC	F1
GPT-small	Skytrax	**2**	**0.899**	**0.953**	**0.871**
		6	0.879	0.952	0.834
		10	0.535	0.772	0.446
	CMS	2	0.821	0.982	0.805
		6	0.926	0.996	0.917
		10	**0.937**	**0.976**	**0.934**
	ECHR	2	0.901	0.712	0.873
		6	**0.986**	**0.791**	**0.989**
		10	0.507	0.804	0.527
GPT-large	Skytrax	**6**	**0.980**	**0.989**	**0.960**
		18	0.899	0.962	0.879
		30	0.737	0.853	0.723
	CMS	6	0.663	0.945	0.603
		18	0.958	0.989	0.958
		30	**0.997**	**0.993**	**0.990**
	ECHR	**6**	**0.986**	**0.806**	**0.989**
		18	0.930	0.729	0.914
		30	0.901	0.774	0.883
Qwen3	Skytrax	**6**	**0.808**	**0.969**	**0.849**
		18	0.232	0.734	0.113
		30	0.828	0.860	0.796
	CMS	6	0.832	0.973	0.743
		18	0.642	0.873	0.588
		30	**0.990**	**0.994**	**0.989**
	ECHR	**6**	**0.789**	**0.795**	**0.812**
		18	0.451	0.657	0.446
		30	0.578	0.793	0.638

Table 1: Overall attack performance at representative shallow, middle, and deep layers for each model on the Skytrax, CMS, and ECHR datasets.

6.2 Overall Performance

We evaluate the proposed attack on five uniformly sampled layers per model and report representative shallow, middle, and deep-layer results in Table 1. Across all models, sensitive-word leakage is observed at multiple intermediate layers, while the depth at which the strongest leakage occurs varies across datasets. On the Skytrax dataset, leakage is most pronounced at shallow or early-middle layers. In contrast, on the CMS dataset, the deepest layers consistently yield the strongest attack performance across all evaluated models. On the ECHR dataset, the layer at which the strongest leakage occurs differs across models.

Findings. Intermediate representations alone are sufficient to leak sensitive attributes across multiple layers, even when the sensitive attributes are difficult to distinguish.

6.3 Ablation Studies

We ablate the two core components of IR-AIA: SG-APCR and sliding-window semantic matching,

Model	Dataset	Layer	Top-1↓	AUC↓	F1↓
GPT-small	Skytrax	2	0.747	0.455	0.845
	CMS	10	0.832	0.482	0.915
GPT-large	Skytrax	6	0.828	0.550	0.934
	CMS	30	0.895	0.497	0.981
Qwen3	Skytrax	6	0.707	0.359	0.831
	CMS	30	0.895	0.523	0.981

Table 2: Performance differences when **removing SG-APCR**, computed as ↓= full − ablated. Positive values indicate degradation; negative values indicate improvement.

Model	Dataset	Layer	Top-1↓	AUC↓	F1↓
GPT-small	Skytrax	2	0.061	-0.006	0.069
	CMS	10	0.421	0.120	0.504
GPT-large	Skytrax	6	0.041	0.005	0.003
	CMS	30	0.284	0.037	0.360
Qwen3	Skytrax	6	-0.020	0.005	-0.013
	CMS	30	0.074	0.006	0.126

Table 3: Performance differences when **removing the sliding-window module**. The differences are computed in the same manner as in Table 2.

keeping all other settings identical to the main experiments. Removing SG-APCR causes substantial performance degradation across all evaluated models and datasets (Table 2), indicating that without anisotropy correction, similarity-based attribute inference on intermediate representations becomes highly unreliable. In contrast, removing the sliding-window module yields domain-dependent effects (Table 3): the impact is minor on Skytrax but substantial on CMS, where sensitive attributes often span multiple tokens.

Findings. SG-APCR is consistently necessary for effective attribute inference from intermediate representations, while sliding-window semantic matching is critically important in domains with multi-token sensitive attributes.

6.4 Attack Effectiveness Under Defense Methods

We evaluate how an existing forward-pass privacy defense affects both downstream model utility and the effectiveness of our attack. Specifically, we consider DP-Forward (Du et al., 2023), a differential privacy defense designed to protect intermediate representations during the forward pass while preserving model utility, and apply it during inference of GPT-large on the CMS dataset. The attack is evaluated at layer 30 using the complete

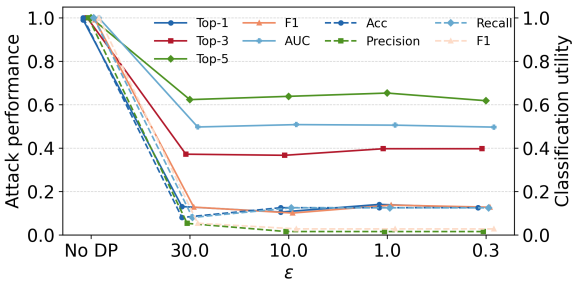

Figure 4: Attack performance and model utility under DP-Forward with different privacy budgets ε. Solid lines report attack performance metrics, while dashed lines correspond to downstream model utility metrics.

IR-AIA pipeline. Figure 4 reports both model utility metrics and attack performance as the privacy budget ε varies. Across the evaluated privacy budgets, enabling DP-Forward substantially degrades downstream classification performance while reducing attribute inference toward random guessing as the privacy budget decreases. Even at comparatively large privacy budgets, downstream utility remains significantly impaired for the multi-class task, with no evaluated setting simultaneously preserving strong model utility and effectively suppressing attribute inference.

Findings. Noise-based forward-pass defenses are unable to achieve a favorable privacy–utility trade-off: the degradation in downstream utility often exceeds the suppression of attribute inference.

7 Defense Insight

We investigate whether existing forward-pass privacy defenses can mitigate attribute inference from intermediate representations by examining DP-Forward (Du et al., 2023). Our analysis reveals a fundamental limitation of such noise-based defenses in this setting: although DP-Forward applies selective and utility-aware perturbations, it does not explicitly account for how sensitive attributes are encoded in the structured geometry of intermediate representations.

These observations suggest that effective protection of intermediate representations requires moving beyond purely noise-based perturbations toward structure-aware defense mechanisms. One promising direction is to regularize the geometry of intermediate representations such that attribute-aligned semantic directions are decorrelated from task-relevant subspaces, thereby reducing attribute separability while preserving utility. Another complementary direction is to explicitly model attribute

inference as an adversarial objective, penalizing intermediate representations that exhibit strong semantic alignment with sensitive attribute prototypes, while maintaining downstream performance. Together, these directions point toward structure-aware and inference-aware defenses as a more principled alternative to existing noise-based forward-pass privacy mechanisms.

8 Related Work

Attacks on Distributed LLM Inference. A growing body of work has studied privacy threats in split and collaborative inference frameworks. Early work on vision models shows that intermediate activations, sometimes referred to as smashed data, can reveal high fidelity input information (Li et al., 2023; Yin et al., 2023). These findings suggest that partial model execution alone offers limited privacy protection. Recent studies extend these attacks to large language models. Prompt inference and inversion techniques (Luo et al., 2025; Qu et al., 2025) demonstrate that adversaries observing hidden states in distributed LLM inference can reconstruct user prompts. Split based fine tuning research further indicates that both forward and backward activations expose sensitive training samples (Chen et al., 2024). Taken together, these works establish that distributed LLM deployments remain vulnerable to input reconstruction from shared intermediate representations.

Attribute Inference Attacks on Language Models. Attribute inference has been extensively investigated in centralized settings. Prior studies show that embedding spaces encode sensitive demographic or document level attributes (Song and Raghunathan, 2020) and that language models are vulnerable to various inference attacks when outputs or auxiliary data can be queried (Pan et al., 2020). More recent work finds that sentence embeddings produced by pretrained language models carry rich attribute signals (Gu et al., 2023) and that LLMs may reveal user or content attributes even without explicit memorization (Kandpal et al., 2024; Staab et al., 2023).

Taken together, prior works primarily focus on two distinct threats: input reconstruction from intermediate activations and attribute inference based on final-layer embeddings. In contrast, the setting of attribute inference from intermediate hidden states in distributed LLM inference, without relying on auxiliary data, remains less explored. Our work complements existing studies by examining whether sensitive attributes can be inferred under this threat model, where the adversary observes only intermediate representations.

9 Conclusion

In this work, we study the privacy risks of intermediate hidden states in distributed LLM inference. We show that sensitive attributes can be inferred even when an adversary relies solely on intermediate representations. To this end, we introduce IR-AIA, a purely intermediate-representation-based attribute inference attack, together with SG-APCR and a sliding-window similarity matching strategy to address layer-dependent anisotropy and subword-level semantic fragmentation. Experiments across multiple LLMs and real-world datasets demonstrate that intermediate representations alone can leak sensitive user attributes, highlighting non-trivial privacy risks in distributed LLM deployments. Future work includes exploring defenses for intermediate activations and extending the analysis to broader settings.

Limitations

This work demonstrates the feasibility of data-free attribute inference from intermediate representations under a distributed inference threat model. While our evaluation is conducted in controlled settings, validating the attack in large-scale production systems remains an important direction for future work. Moreover, we focus on a single adversarial participant observing intermediate activations. Scenarios involving multiple adversaries, acting independently or collaboratively, are left for future investigation.

Ethical Considerations

This work examines privacy risks in distributed LLM inference by analyzing how sensitive attributes may be implicitly encoded in intermediate representations. The study is conducted under a realistic honest-but-curious threat model and does not assume access to raw user inputs, private data, or model parameters beyond those legitimately available in distributed inference. Our goal is to support defensive analysis and responsible deployment by identifying potential leakage mechanisms, rather than enabling misuse. All experiments rely on publicly available datasets or synthetic inputs and do not involve personal or identifiable user data.

References

Alexander Borzunov, Dmitry Baranchuk, Tim Dettmers, Maksim Riabinin, Younes Belkada, Artem Chumachenko, Pavel Samygin, and Colin Raffel. 2023a. Petals: Collaborative inference and fine-tuning of large models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 558–568, Toronto, Canada.

Alexander Borzunov, Max Ryabinin, Artem Chumachenko, Dmitry Baranchuk, Tim Dettmers, Younes Belkada, Pavel Samygin, and Colin A Raffel. 2023b. Distributed inference and fine-tuning of large language models over the internet. *Advances in neural information processing systems*, 36:12312–12331.

Kritya Bunchongchit and Walanchalee Wattanacharoensil. 2021. Data analytics of skytrax's airport review and ratings: Views of airport quality by passengers types. *Research in Transportation Business & Management*, 41:100688.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. *arXiv preprint arXiv:1906.02059*.

Guanzhong Chen, Zhenghan Qin, Mingxin Yang, Yajie Zhou, Tao Fan, Tianyu Du, and Zenglin Xu. 2024. Unveiling the vulnerability of private fine-tuning in split-based frameworks for large language models: A bidirectionally enhanced attack. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 2904–2918.

Minxin Du, Xiang Yue, Sherman SM Chow, Tianhao Wang, Chenyu Huang, and Huan Sun. 2023. Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 2665–2679.

Wei Fan, Haoran Li, Zheye Deng, Weiqi Wang, and Yangqiu Song. 2024. Goldcoin: Grounding large language models in privacy laws via contextual integrity theory. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3321–3343.

Kang Gu, Ehsanul Kabir, Neha Ramsurrun, Soroush Vosoughi, and Shagufta Mehnaz. 2023. Towards sentence level inference attack against pre-trained language models. *Proceedings on Privacy Enhancing Technologies*.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Yuxiang Huang, Mingye Li, Xu Han, Chaojun Xiao, Weilin Zhao, Ao Sun, Hao Zhou, Jie Zhou, Zhiyuan Liu, and Maosong Sun. 2025. APB: Accelerating distributed long-context inference by passing compressed context blocks across GPUs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10708–10727, Vienna, Austria. Association for Computational Linguistics.

Bargav Jayaraman and David Evans. 2022. Are attribute inference attacks just imputation? In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 1569–1582.

Youhe Jiang, Ran Yan, Xiaozhe Yao, Yang Zhou, Beidi Chen, and Binhang Yuan. 2024. Hexgen: Generative inference of large language model over heterogeneous environment. In *International Conference on Machine Learning*, pages 21946–21961. PMLR.

Nikhil Kandpal, Krishna Pillutla, Alina Oprea, Peter Kairouz, Christopher A Choquette-Choo, and Zheng Xu. 2024. User inference attacks on large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18238–18265.

Jabari Kwesi, Jiaxun Cao, Riya Manchanda, and Pardis Emami-Naeini. 2025. Exploring user security and privacy attitudes and concerns toward the use of {General-Purpose}{LLM} chatbots for mental health. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 6007–6024.

Ziang Li, Mengda Yang, Yaxin Liu, Juan Wang, Hongxin Hu, Wenzhe Yi, and Xiaoyang Xu. 2023. Gan you see me? enhanced data reconstruction attacks against split inference. *Advances in neural information processing systems*, 36:54554–54566.

Zihan Liu, Yizhen Wang, Rui Wang, and Sai Wu. 2025. Dualguard: A parameter space transformation approach for bidirectional defense in split-based llm fine-tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17065–17080.

Xinjian Luo, Ting Yu, and Xiaokui Xiao. 2025. Prompt inference attack on distributed large language model inference frameworks. *arXiv preprint arXiv:2503.09291*.

Anemily Machina and Robert Mercer. 2024. Anisotropy is not inherent to transformers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4892–4907.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2017. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331. IEEE.

Wenjie Qu, Yuguang Zhou, Yongji Wu, Tingsong Xiao, Binhang Yuan, Yiming Li, and Jiaheng Zhang. 2025. Prompt inversion attack against collaborative inference of large language models. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 1695–1712. IEEE.

Federico Ravenda, Seyed Ali Bahrainian, Andrea Raballo, Antonietta Mira, and Noriko Kando. 2025. Are LLMs effective psychological assessors? leveraging adaptive RAG for interpretable mental health screening through psychometric practice. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8975–8991, Vienna, Austria. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135.

Yashothara Shanmugarasa, Ming Ding, Chamikara Mahawaga Arachchige, and Thierry Rakotoarivelo. 2025. Sok: The privacy paradox of large language models: Advancements, privacy risks, and mitigation. In *Proceedings of the 20th ACM Asia Conference on Computer and Communications Security*, pages 425–441.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*.

Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 377–390.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

Adam Sutton, Xi Bai, Kawsar Noor, Thomas Searle, and Richard Dobson. 2025. Named entity inference attacks on clinical llms: Exploring privacy risks and the impact of mitigation strategies. In *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*, pages 42–52.

Yuecen Wei, Xingcheng Fu, Lingyun Liu, Qingyun Sun, Hao Peng, and Chunming Hu. 2025. Prompt-based unifying inference attack on graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12836–12844.

Yupeng Yin, Xianglong Zhang, Huanle Zhang, Feng Li, Yue Yu, Xiuzhen Cheng, and Pengfei Hu. 2023. Ginver: Generative model inversion attacks against collaborative inference. In *Proceedings of the ACM Web Conference 2023*, pages 2122–2131.

A Additional Details for Experimental Setup

A.1 Dataset Details

CMS–PDE. The Prescription Drug Event (PDE) dataset released by the Centers for Medicare & Medicaid Services (CMS) contains more than 50,000 prescription drug reimbursement records. Each record includes attributes such as beneficiary identifier, prescribing physician, prescription date, drug name, and days of supply. In our study, the *drug name* field is used as the sensitive attribute because it directly reflects disease categories and thus carries medically sensitive information.

Skytrax. The Skytrax airline review dataset comprises approximately 41,000 user-generated airline reviews. Each entry contains structured metadata including flight route, reviewer name, reviewer country, review date, cabin class, and free-form textual feedback. We treat the *reviewer country* field as the sensitive attribute, corresponding to users' geographic identity.

Sampling Procedure. For both datasets, we construct a controlled evaluation set by selecting ten sensitive-word categories (e.g., ten drug names or ten country names). From each category, we uniformly sample 100 records using stratified sampling, resulting in a balanced evaluation set with equal representation across sensitive-word classes. This avoids distributional skew and ensures that differences in model performance are attributable to model behaviour rather than dataset imbalance.

Token Statistics. To illustrate the effect of different tokenization schemes, we report token counts under two tokenizers. Under the GPT-2 tokenizer, the CMS dataset has an average sequence length of 505 tokens and the Skytrax dataset averages 191 tokens. Under the Qwen tokenizer, the averages are 493 tokens for CMS and 190 tokens for Skytrax. These statistics indicate that the CMS records are substantially longer and more structured, whereas Skytrax reviews are shorter and less homogeneous.

Model	Dataset	Layer	Top-1	Top-3	Top-5	AUC	F1
GPT-small	Skytrax	2	0.8990	0.9495	0.9697	0.9533	0.8710
		4	0.8586	0.9192	0.9293	0.9415	0.8027
		6	0.8788	0.9293	0.9596	0.9519	0.8337
		8	0.7778	0.8889	0.9394	0.8955	0.6978
		10	0.5354	0.7374	0.8586	0.7721	0.4464
	CMS	2	0.8211	1.0000	1.0000	0.9823	0.8053
		4	0.9579	1.0000	1.0000	0.9974	0.9568
		6	0.9263	1.0000	1.0000	0.9961	0.9171
		8	0.9684	1.0000	1.0000	0.9983	0.9693
		10	0.9368	1.0000	1.0000	0.9760	0.9341
GPT-large	Skytrax	6	0.9798	1.0000	1.0000	0.9894	0.9595
		12	0.9394	0.9899	1.0000	0.9888	0.9203
		18	0.8990	0.9798	0.9798	0.9624	0.8794
		24	0.8687	0.8990	0.9394	0.9139	0.8405
		30	0.7374	0.8384	0.8889	0.8534	0.7231
	CMS	6	0.6632	0.9895	1.0000	0.9454	0.6032
		12	0.9053	1.0000	1.0000	0.9898	0.8549
		18	0.9579	1.0000	1.0000	0.9893	0.9583
		24	1.0000	1.0000	1.0000	0.9926	1.0000
		30	1.0000	1.0000	1.0000	0.9933	1.0000
Qwen3-4B	Skytrax	6	0.8081	0.9495	1.0000	0.9688	0.8488
		12	0.6061	0.8384	0.8485	0.8018	0.5806
		18	0.2323	0.6667	0.8990	0.7338	0.1132
		24	0.5960	0.8485	0.9091	0.8150	0.5190
		30	0.8283	0.8485	0.8485	0.8595	0.7962
	CMS	6	0.8316	1.0000	1.0000	0.9734	0.7430
		12	0.8947	1.0000	1.0000	0.9714	0.8667
		18	0.6421	0.8000	0.9789	0.8729	0.5881
		24	1.0000	1.0000	1.0000	0.9736	1.0000
		30	1.0000	1.0000	1.0000	0.9943	1.0000

Table 4: Full layer-wise attack performance across all evaluated layers, models, and datasets.

A.2 Model Configurations

We evaluate three pretrained large language models to assess robustness across architectural depth and tokenization mechanisms:

- **GPT-small**: a 12-layer Transformer using the GPT-2 tokenizer.
- **GPT-large**: a 36-layer Transformer sharing the same tokenizer as GPT-small but with greater depth.
- **Qwen3-4B-Thinking-2507**: a 36-layer Transformer using a distinct tokenizer and vocabulary from GPT, representing a different tokenization paradigm.

This combination allows us to study two key axes of variation: (1) *same tokenization but different depth* (GPT-small vs. GPT-large), and (2) *same depth but different tokenization* (GPT-large vs. Qwen). Together, these settings test the adaptability of our method to both architectural and tokenization differences in modern LLMs.

A.3 Hyperparameter Settings

In all experiments, we fix the isotropy-related parameter τ_{iso} to 0.1 for GPT-family models and to 0.02 for Qwen models, reflecting differences in the scale and distribution of their intermediate representations. The parameter α is set to 0.5 across all experiments. We found that the method is not overly sensitive to small variations of these hyperparameters, and thus we use fixed values for consistency.

A.4 Evaluation Metric Details

We adopt standard metrics to evaluate sensitive-word matching accuracy:

- **Top-k hit rate (Top-1 / Top-3 / Top-5)**: measures whether the correct sensitive word appears among the top-k most similar predictions, capturing ranking and retrieval quality.
- **AUC (Area Under the ROC Curve)**: evaluates overall discrimination capability across thresholds.
- **F1 score**: reflects the balance between preci-

Model	Dataset	Layer	Top-1	AUC	F1
GPT-small	Skytrax	2	0.152 (-0.747)	0.498 (-0.455)	0.026 (-0.845)
GPT-small	CMS	10	0.105 (-0.832)	0.494 (-0.482)	0.019 (-0.915)
GPT-large	Skytrax	6	0.152 (-0.828)	0.439 (-0.550)	0.026 (-0.934)
GPT-large	CMS	30	0.105 (-0.895)	0.496 (-0.497)	0.019 (-0.981)
Qwen	Skytrax	6	0.101 (-0.707)	0.610 (-0.359)	0.018 (-0.831)
Qwen	CMS	30	0.105 (-0.895)	0.471 (-0.523)	0.019 (-0.981)

Table 5: Full ablation results for removing SG-APCR. Each entry shows the score of the ablated variant and, in parentheses, the change relative to the full attack at the same layer.

Model	Dataset	Layer	Top-1	AUC	F1
GPT-small	Skytrax	2	0.838 (-0.061)	0.959 (+0.006)	0.802 (-0.069)
GPT-small	CMS	10	0.516 (-0.421)	0.856 (-0.120)	0.430 (-0.504)
GPT-large	Skytrax	6	0.939 (-0.041)	0.984 (-0.005)	0.957 (-0.003)
GPT-large	CMS	30	0.716 (-0.284)	0.956 (-0.037)	0.640 (-0.360)
Qwen	Skytrax	6	0.828 (+0.020)	0.964 (-0.005)	0.862 (+0.013)
Qwen	CMS	30	0.926 (-0.074)	0.988 (-0.006)	0.874 (-0.126)

Table 6: Full ablation results for removing sliding-window maximum similarity matching. Each entry shows the ablated variant and the change relative to the full attack.

sion and recall for the final predicted sensitive word.

These metrics collectively assess both classification performance (AUC, F1) and semantic ranking quality (Top-k), providing a comprehensive view of attack effectiveness across domains and models.

B Full Layer-Wise Performance Results

This appendix reports the complete layer-wise performance across all five evaluated layers for each model and dataset. These results expand upon the representative shallow, middle, and deep layers presented in Table 4 of the main text.

C Full Ablation Results

This appendix provides the complete numerical results and detailed analyses for the ablation studies introduced in Section 6.3 of the main text. For each model–dataset pair, we report results at the layer that achieves the highest F1 score under the full attack. Two components are ablated: (1) the sphericity-guided adaptive principal component removal (SG-APCR), and (2) the sliding-window semantic matching.

We first evaluate the effect of disabling SG-APCR. In this variant, intermediate representations are used directly without debiasing, while the sliding-window module is kept intact. Table 5 summarizes the complete results.

We next ablate the sliding-window semantic matching component. In this variant, the attack directly takes the maximum cosine similarity over individual token embeddings, without aggregating over token spans. Table 6 presents the full results.

D DP-Forward Defense Experiment Details

This section provides the complete experimental configuration for the DP-Forward defense evaluation. The goal is to examine how a differential-privacy-based inference-time defense affects both downstream model utility and the performance of our proposed attack in a distributed inference setting.

D.1 Dataset and Task Construction

We use a subset of the CMS Prescription Drug Event (PDE) dataset for this experiment. To construct a downstream supervised task that allows measurement of model utility, we sample:

- 800 samples for training,
- 200 samples for testing.

Each sample corresponds to a medical record containing a prescribed drug name. We align the drug categories with the eight sensitive-word classes used in our attack, thereby forming an eight-class classification task. Each training or test ex-

ample is assigned a label representing its drug category.

D.2 Model Setup

We evaluate the defense on **GPT-large**, the 36-layer model used in the main experiments. The attack is always evaluated at **layer 30**, which yields the strongest leakage under the non-private setting. All hyperparameters for our attack follow the main experiment configuration, including:

- isotropy threshold $\tau_{\text{iso}} = 0.1$,
- sphericity-guided adaptive component removal enabled,
- sliding-window semantic matching enabled.

D.3 DP-Forward Configuration

DP-Forward is a forward-pass differential privacy mechanism that adds Gaussian noise to intermediate hidden representations. We adopt the same configuration as in the original work (Du et al., 2023), and vary only the privacy budget:

$$\varepsilon \in \{\text{no-DP}, 30.0, 10.0, 1.0, 0.3\}.$$

Additional DP hyperparameters are:

- clipping norm: $C = 1.0$,
- privacy parameter: $\delta = 10^{-5}$.

D.4 Training Procedure

The downstream classifier is trained on the 800-sample training set using GPT-large with DP-Forward applied to every forward pass. We evaluate the trained classifier on the 200-sample test set to measure utility under each privacy budget.

All training runs—DP and non-DP—share the same optimizer, learning rate, batch size, and number of epochs. Only the privacy mechanism differs.

E Full Algorithm

Algorithm 1 provides the complete data-free attribute inference procedure described in Section 5. It includes all computational steps for sphericity-guided principal component removal, shared-basis debiasing, and sliding-window semantic matching.

Algorithm 1 IR-AIA

Require: $P = [t_1, \ldots, t_m]$ (prompt tokens);
$W^s = \{w_q^s\}_{q=1}^{|W^s|}$ (sensitive words);
token length L_q for each w_q^s;
target layer ℓ_a; isotropy threshold τ_{iso};
search ratio α
Ensure: similarity scores $\{S_q\}$ and matched sensitive word $\hat{y}$

Extract intermediate embeddings at layer ℓ_a:

$$E = \begin{bmatrix} (e_{t_1}^{(\ell_a)})^\top \\ \vdots \\ (e_{t_m}^{(\ell_a)})^\top \end{bmatrix} \in \mathbb{R}^{m \times d}$$

Perform SVD: $E = U\Sigma V^\top$, where $V = [u_1, \ldots, u_d]$
Initialize:

$$R_{\max} = \lceil \alpha \cdot \min(m, d) \rceil,$$
$$\text{best\_r} \leftarrow 0,$$
$$\text{best\_iso} \leftarrow \text{IsoGain}(E)$$

for $r = 0$ to $R_{\max}$ **do**
  $U_r \leftarrow [u_1, \ldots, u_r]$
  $E_r \leftarrow E - EU_rU_r^\top$
  $\text{Cov}_r \leftarrow E_r^\top E_r / m$
  Compute eigenvalues $\{\lambda_i\}$ of Cov_r
  $\text{PR}_r \leftarrow (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$
  $\text{Iso}_r \leftarrow (\text{PR}_r - 1)/(d-1)$
  if $\text{Iso}_r \geq \tau_{\text{iso}}$ **then**
    best_r $\leftarrow r$
    break
  else if $\text{Iso}_r >$ best_iso **then**
    best_iso $\leftarrow \text{Iso}_r$
    best_r $\leftarrow r$
  end if
end for
Apply shared-basis removal:

$$E' = E - EU_{\text{best\_r}}U_{\text{best\_r}}^\top$$

for each $w_q^s \in \mathcal{W}_s$ **do**
  Compute mean embedding $\bar{e}_{w_q^s}^{(\ell_a)}$
  Debias:

$$\bar{e}'^{(\ell_a)}_{w_q^s} = \bar{e}^{(\ell_a)}_{w_q^s} - \sum_{j=1}^{\text{best\_r}} (u_j^\top \bar{e}^{(\ell_a)}_{w_q^s}) u_j$$

  Compute sliding-window similarity:

$$S_q = \max_{1 \leq i \leq m - L_q + 1} \text{cosine}\left(\frac{1}{L_q} \sum_{k=0}^{L_q - 1} E'[i+k, :],\ \bar{e}'^{(\ell_a)}_{w_q^s} \right)$$

end for
Match the highest-scoring sensitive word:

$$q^* = \arg\max_q S_q$$

return

$$w_{q^*}^s$$