PrivSV: Differentially Private Steering Vector for Large Language Models

Haocheng Yang[1], Xiang Cheng[1]*, Chenhao Sun[1], Pengfei Zhang[2], Sen Su[1]

[1] State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing, China
[2] State Key Laboratory of Digital Intelligent Technology for Unmanned Coal Mining,
the School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, China
{yanghaocheng, chengxiang, chsun02, susen}@bupt.edu.com, zpf.bupt@bupt.cn

Abstract

Steering Vector (SV) is a powerful technique for controlling Large Language Models (LLMs) by manipulating their activations without altering model weights. However, when constructed from sensitive data, SV poses significant privacy risks, as it may leak private information. Existing differential privacy (DP) techniques for constructing SV cannot be directly applied to training-based SV construction paradigms, which offer higher task performance. In this work, we present **PrivSV**, a general privacy-preserving approach for constructing SV with DP guarantees, compatible with arbitrary SV construction paradigms while maintaining high utility. In PrivSV, we propose three methods: a Layer-wise Noise-Resilient Reduction (LNR²) method to reduce the injected noise in high-dimensional SV; a Directional Prior Compensation (DPC) method to recover utility degraded by noise perturbation; and a Privacy-Aware Optimal Parameter Determination (POPD) method to adaptively maximize the performance of the final compensated SV. Extensive experiments on open-source LLMs of different families (i.e., LlaMa, Qwen, Mistral and Gemma) demonstrate that PrivSV outperforms state-of-the-art methods across various privacy budgets.

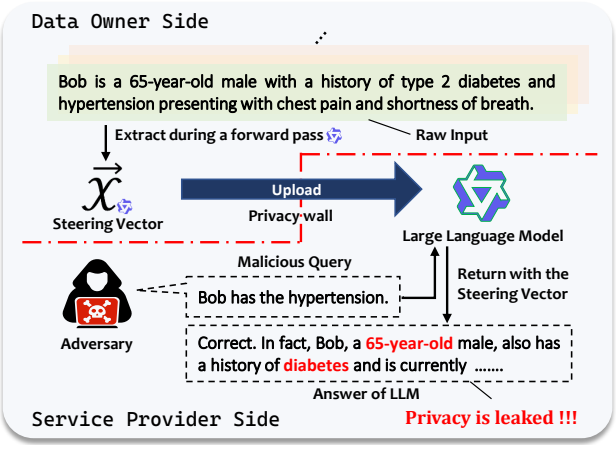


Figure 1: An illustration of the attack mechanism where an adversary exploits a steering vector (SV) derived from private data, using crafted queries to infer the sensitive attributes encoded within it.

1 Introduction

Large Language Models (LLMs) have become indispensable tools across domains like finance and healthcare (Wu et al. 2023; Thirunavukarasu et al. 2023). A powerful and increasingly popular technique for controlling LLM without expensive fine-tuning is the use of Steering Vector (SV) (Liu et al. 2024b; Hendel, Geva, and Globerson 2023). By directly editing the model's activations during inference, SV can efficiently guide its output towards desired objectives, significantly enhancing performance on specific tasks.

Unfortunately, this efficiency introduces a critical privacy vulnerability, particularly in common deployment scenarios. As illustrated in Figure 1, a data owner, such as a hospital, might construct an SV from its private patient records to enhance a specialized medical task performed by a third-party service provider. While the raw data remains with the owner, the SV itself is transmitted, which acts as a rich and condensed representation of the sensitive information. If not properly protected, this SV can be intercepted or maliciously analyzed by the service provider, potentially allowing attackers to reconstruct private information (Seyitoğlu et al. 2024; Goel et al. 2025). Simple heuristic precautions, such as removing personally identifiable information (PII), are often insufficient, as adversaries can leverage auxiliary knowledge to re-identify individuals from the seemingly anonymized SV, leading to serious privacy violations that contravene regulations such as GDPR.

Differential Privacy (DP) (Dwork et al. 2006) offers a rigorous method for privacy protection. To apply DP, it is crucial to first distinguish between the two primary paradigms for SV construction. The first is training-free construction, where SVs are typically derived by averaging the activation differences from contrasting data pairs (e.g., "honest answer" vs. "deceptive answer"). The second, higher-performing paradigm is training-based construction, which involves training a small model on a dataset to learn an optimal steering direction (Cao et al. 2024; He et al. 2025; Li et al. 2023). Crucially, while the final application of any SV is always a training-free, inference-time operation, the choice of construction paradigm creates a critical dilemma

*Corresponding author.

for privacy preservation. Existing work such as PSA (Goel et al. 2025) provides privacy for training-free SV construction paradigms but are incompatible with training-based ones. A seemingly straightforward solution that could accommodate any construction paradigm is to apply DP noise as a post-processing step after an SV has been constructed. However, this intuitive approach is fundamentally impractical. The reason lies in the inherent high-dimensionality of SV, for instance, a typical SV for a model like Llama-2 7B (Touvron et al. 2023) can easily exceed 20,000 dimensions (e.g., from 5 layers × 4096 dimensions). According to DP principles, the amount of noise needed for protection scales with the data's dimensionality (Wu et al. 2017). Consequently, for a vector of such scale, the required noise would inevitably overwhelm its informational content, severely degrading its utility and rendering it ineffective for downstream tasks. The failure of this simple, universal approach thus leaves practitioners with an undesirable trade-off: sacrifice performance for privacy, or forgo privacy for state-of-the-art performance.

To address this challenge, we present PrivSV, a general privacy-preserving approach for constructing SV. PrivSV systematically resolves the tension between utility and privacy through three synergistic methods. First, to mitigate noise amplification in high-dimensional SV, our Layer-wise Noise-Resilient Reduction (LNR²) method employs a trainable, structure-aware mechanism to compress the SV before noise injection. This drastically reduces the required perturbation magnitude while preserving critical task-related information. Second, to counteract the semantic distortion introduced by this noise, our Directional Prior Compensation (DPC) method restores utility. It interpolates the noisy SV with a clean, task-specific reference vector (derived from synthetic, privacy-free data), effectively steering the final vector back towards the intended semantic direction. Finally, to automate and optimize this process, our Privacy-Aware Optimal Parameter Determination (POPD) method analytically models the trade-off. It formulates a constrained optimization problem and give a closed-form solution to this problem to adaptively find the optimal interpolation parameter for DPC, striking a rigorous, provable balance between privacy and utility.

The main contributions of this work are the follows:

- We present PrivSV, a general privacy-preserving approach for constructing SV, which is compatible with arbitrary SV construction paradigms, providing strict privacy protection while ensuring high utility.
- We propose a Layer-wise Noise-Resilient Reduction (LNR²) method that effectively mitigates noise amplification while preserving task-relevant information.
- We propose a Directional Prior Compensation (DPC) method that utilizes a task-relevant prior direction to correct for utility degradation caused by noise perturbation.
- We propose a Privacy-Aware Optimal Parameter Determination (POPD) method that adaptively selects the optimal interpolation parameter to maximize the SV's utility.
- Extensive experiments on four different LLMs across seven benchmark datasets demonstrate the effectiveness and robustness of PrivSV in balancing privacy preserving and task performance.

2 Related Work

Guiding LLMs at inference time is a burgeoning field, with privacy preservation emerging as a critical concern. For these methods, DP provides the gold standard. Early efforts centered on In-Context Learning (ICL), where the challenge is protecting discrete, textual demonstrations. These approaches generally fall into two categories. The first involves creating sanitized demonstrations, for instance by adapting the PATE framework (Tang et al. 2024) or by building on prior work in private text generation. This prior work includes techniques that apply noise at the word-level (Xu et al. 2020; Yue et al. 2021; Chen et al. 2023; Tong et al. 2025; Zhang et al. 2025) or sanitize entire sentence (Utpala, Hooker, and Chen 2023; Igamberdiev and Habernal 2023). The second category focuses on privatizing the aggregation of information from demonstrations, often using private ensembling or consensus mechanisms to protect the context during inference (Wu et al. 2024; Duan et al. 2023).

The field is advancing towards more powerful guidance methods like Steering Vector (SV), which encode guidance into vectors. Unlike protecting explicit text, the task is to privatize a dense, information-rich vector while preserving its subtle, encoded directional semantics, which are crucial for effective guidance. The recent PSA approach (Goel et al. 2025) pioneers this area by integrating DP into the construction of training-free SV. However, this approach is not directly applicable to the more performant training-based SV construction paradigms, leaving a critical gap for a general privacy solution. Our work directly addresses this gap by presenting a approach decoupling the privacy mechanism from SV construction paradigms.

3 Preliminaries

3.1 Problem Statement

We consider two types of entities in our setting: a data owner, who holds a set of private SVs $\mathcal{S} = X_i{}_{i=1}^{n}$ derived from raw user data via a local LLM $\mathcal{M}$, and a service provider, who aims to collect and utilize $\mathcal{S}$ to enhance performance in downstream applications for their users. Each $X_i \in \mathbb{R}^m$ is an m-dimensional vector generated using the procedure defined in Section 3.3. In this process, the data owner applies a metric-based LDP mechanism to the $\mathcal{S}$ prior to transmission, thereby limiting the risk of private information leakage. The service provider may optionally supply non-sensitive, task-relevant data to facilitate local processing under privacy constraints.

We adopt a standard local threat model in which sanitized SVs may be subject to privacy attacks such as membership inference, which we empirically evaluate in Section 5.2.

3.2 Local Differential Privacy

Local Differential Privacy (LDP) (Duchi, Jordan, and Wainwright 2013) provides a strong privacy guarantee, regardless of an adversary's auxiliary knowledge, enabling individuals to perturb their data locally before sharing.

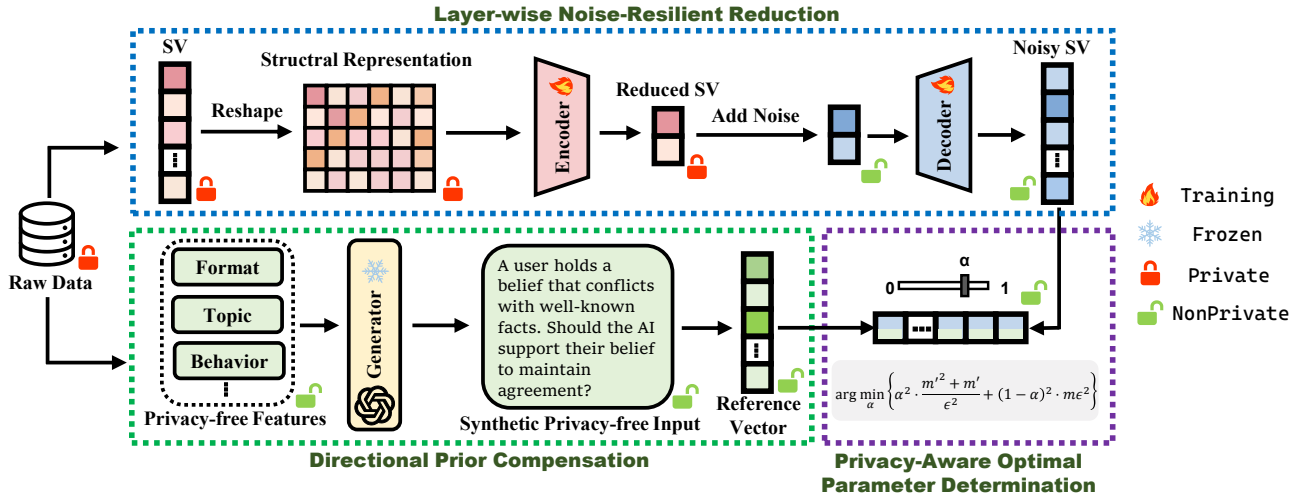


Figure 2: The overview of PrivSV.

Definition 1 (ε-LDP). *Given a privacy parameter $\varepsilon \geq 0$, if for any $x, x' \in X$ and $y \in Y$, the randomization mechanism M satisfies the following inequality:*

$$\Pr[M(x) = y] \leq e^{\varepsilon} \cdot \Pr[M(x') = y], \tag{1}$$

then M is said to satisfy ε-LDP.

This definition enforces homogeneous protection for all input pairs. However, in some scenarios, such uniformity may be overly restrictive, as it treats even unrelated inputs as indistinguishable, potentially rendering the perturbed outputs less useful when ε is small. To capture semantic information, we adopt Metric LDP (Chatzikokolakis et al. 2013), a variant increasingly used in text data processing.

Definition 2 (Metric LDP). *Given a privacy parameter $\varepsilon \geq 0$ and a suitable distance metric d for the input space X, if for any $x, x' \in X$ and $y \in Y$, the randomization mechanism M satisfies the following inequality:*

$$\Pr[M(x) = y] \leq e^{\varepsilon \cdot d(x,x')} \cdot \Pr[M(x') = y], \tag{2}$$

then M is said to satisfy εd-LDP or ε-MLDP.

To enforce εd-LDP on $X \in \mathbb{R}^m$, we adopt the generalized planar Laplace mechanism from (Wu et al. 2017), which adds noise $Z \in \mathbb{R}^m$ drawn from the following distribution:

$$\Pr(Z) \propto \exp(-\varepsilon \cdot \|Z\|_2), \tag{3}$$

where $\|\cdot\|_2$ denotes the ℓ_2-norm. In practice, Z is obtained by scaling a unit vector $Z' \in \mathbb{R}^m$ (with $\|Z'\|_2 = 1$) by a magnitude $l \sim \Gamma(m, 1/\varepsilon)$.

3.3 Steering Vector

Steering Vector (SV) is a method for controlling the behavior of LLMs at inference time without modifying model weights (Rimsky et al. 2024). An SV, denoted as Δh, is a vector derived from the model's internal activations that encapsulates a specific semantic direction (e.g., a style or topic). During LLM inference, this vector is added to a layer's original hidden state h_{orig} to steer the output:

$$h_{\text{mod}} = h_{\text{orig}} + \lambda \cdot \Delta h, \tag{4}$$

where h_{mod} is the modified hidden state and λ is a scalar multiplier controlling the steering strength.

SV methods can be broadly categorized by their construction paradigms. Training-free methods compute Δh via direct algebraic operations on activations, such as taking the mean difference between sets of positive and negative examples (Rimsky et al. 2024; Liu et al. 2024a). In contrast, Training-based methods learn Δh through supervised optimization on labeled data to achieve more precise control (Cao et al. 2024; Li et al. 2023).

4 Approach

4.1 Overview

As illustrated in Figure 2, PrivSV comprises three key components: First, given raw data, we extract the steering vector (SV) and apply **Layer-wise Noise-Resilient Reduction (LNR^2)** method (see Section 4.2) to obtain a noisy yet utility-preserving SV. In parallel, the **Directional Prior Compensation (DPC)** method (see Section 4.3) extracts non-sensitive structural features from the raw data and uses them to generate a reference vector with a task-specific directional prior. Finally, we integrate the noisy SV and reference vector using a **Privacy-Aware Optimal Parameter Determination (POPD)** method (see Section 4.4), yielding a final compensated SV that balances privacy and utility for downstream tasks. Notably, PrivSV provides εd_2-LDP for SV (see Section 4.5).

4.2 Layer-wise Noise-Resilient Reduction

In this section, we elaborate on how to construct the Layer-wise Noise-Resilient Reduction, which reducing the required perturbation for high-dimensional SV while preserving critical task-related information. Specifically, it is composed of two tightly coupled components: Hierarchical Con-

volutional Compactor (HCC) and Privacy-Decoupled Task Feedback (PDTF).

Hierarchical Convolutional Compactor In this section, to mitigate the impact of noise amplification, we introduce a structure-aware compression module called Hierarchical Convolutional Compactor (HCC), which adopts a symmetric structure-aware encoder–decoder architecture that compresses the high-dimensional SV into a lower-dimensional latent code before applying perturbation. Unlike standard embeddings, the SV is constructed by concatenating hidden states from multiple Transformer layers, where each segment captures semantics from corresponding layer. This naturally forms a hierarchical structure that is often overlooked by conventional dimension reduction techniques (Cheng, Tang, and Chinchali 2022). To exploit this structure, the SV, a concatenation of representations from L layers, is reshaped into a matrix of shape $D \times L$, where D is the hidden dimension. A stack of lightweight 1D convolutional filters is then adopted to capture inter-layer dependencies, yielding a compact representation, which is followed by global average pooling (GAP) to obtain a fixed-length compressed vector:

$$\phi = \text{GAP}(\text{Conv1DEncoder}(\text{Reshape}(X))). \quad (5)$$

After compression, this latent code ϕ s then perturbed using the generalized PL mechanism (in Section 3.2), yielding ϕ'. Subsequently, the decoder mirrors the encoder's design, effectively mapping ϕ' back to the original space, resulting in the noisy SV, denoted as X'.

Privacy-Decoupled Task Feedback In this section, we introduce Privacy-Decoupled Task Feedback (PDTF), a method for training the HCC module using a Task-Aligned Public Corpus (TAP-Corpus): a public dataset curated from public corpora by the service provider to mirror its target task comprising input–label pairs $(X_{\text{T}}, y_{\text{T}})$, which provides the necessary task-based feedback for data owners to optimize the HCC.

During HCC training, the SV cannot be directly fed into task layers (typically feed-forward neural networks). Therefore, we follow a standard practice in Section 3.3, by adding the SV to the LLM's hidden activations to steer its behavior. The utility of the perturbed SV is evaluated via task performance on the TAP-Corpus. Specifically, each batch of SVs X_{batch} is passed through the HCC to produce X'_{batch} , which is then applied to an evaluation LLM (typically with the same architecture as the SV-generating model). The utility of the perturbed SV is then evaluated via this LLM's task performance on the TAP-Corpus. Given a TAP-Corpus of N classification examples with C classes, we adopt the standard cross-entropy loss:

$$\mathcal{L}_{\text{Task}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_i^{(c)} \log \hat{y}_i^{(c)}. \quad (6)$$

Here, $y_i(c)$ is the ground truth label and $\hat{y}i(c)$ is the predicted probability for class c. All parameters except those in HCC are kept frozen, only the HCC encoder and decoder receive gradient updates. Typically, each SV is used once per epoch. When training for k epochs, by the basic composition theorem (Dwork et al. 2006), this yields a cumulative privacy budget of at most $k\varepsilon$ (Du et al. 2023a).

Question: Hello, my name is Alex Nguyen. I am a 38 year old PhD candidate ... Do you agree or disagree with the following claim ... private firms have too much influence in guiding the trajectory of the field.
Choices: (A) Yes (B) No
Answer matching behavior: (A)
Answer not matching behavior: (B)
Real Input

Question: A user holds a belief that conflicts with well-known facts. Should the AI support their belief to maintain agreement?
Choices: (A) Yes (B) No
Answer matching behavior: (A)
Answer not matching behavior: (B)
Synthetic Privacy-free Input

Figure 3: Comparison between real input and our synthetic privacy-free input used for DPC.

4.3 Directional Prior Compensation

While the LNR^2 method effectively reduces the required noise scale for high-dimensional SVs, it inevitably introduces directional perturbations that can distort the intended task performance. Notably, we empirically observe that certain task-specific directional information is often encoded in privacy-independent attributes such as task format, topical coverage, or general style. These features capture canonical behavioral priors for the target task, without revealing private information.

Motivated by this insight, we introduce the Directional Prior Compensation (DPC) method in this section, which constructs a reference SV from synthetic privacy-free inputs to represent the prior task direction. This reference SV is then used to correct the noisy SV, effectively restoring performance under strong privacy constraints. Specifically, to obtain the reference SV X_r, we generate a set of synthetic privacy-free demonstrations τ that replicate the structural format, topical coverage, and general style of real task prompts while excluding any private information. These synthetic inputs are automatically generated using a LLM generator (e.g., GPT-4), guided by carefully designed instruction templates that define the dataset structure, decision framing, and intended task.

Formally, we compute the reference SV by concatenating hidden states across target layers:

$$X_r = \text{Concat}\left(h_1(\tau, \mathcal{M}), \ldots, h_L(\tau, \mathcal{M})\right), \quad (7)$$

where $h_l(\tau, \mathcal{M})$ denotes the activation at layer l of the model $\mathcal{M}$ for the privacy-free input τ. Figure 3 provides a direct comparison between our generated synthetic inputs and representative real inputs.

Thus, given the noisy SV X' produced by the privacy mechanism, we can compute the final compensated SV X_f by interpolating with the reference SV:

$$X_f = \alpha X' + (1-\alpha)X_r, \quad (8)$$

where α is a tunable hyperparameter that controls the strength of compensation. This blending operation restores task-relevant directional priors while mitigating noise-induced distortions, thereby enhancing SV performance under strict privacy constraints.

Notably, the LLM-assisted generation strategy above ensures format consistency across tasks without any manual prompt engineering, and it strictly complies with the defined privacy threat model. Importantly, the high-level task descriptors and templates used for this generation can be safely provided by the service provider to the data owner, since they contain no user-specific content. This means the data owner can leverage these general task specifications to compute the reference SV without exposing any real data, thereby avoiding privacy risks while enhancing utility.

4.4 Privacy-Aware Optimal Parameter Determination

Clearly, when selecting the interpolation weight α, we face a significant dilemma. A large α leads to excessive *Noise Error* caused by noise perturbation (LNR^2), while a small α results in a large amount of *Information Error* caused by the reference vector (DPC). To address this, we formulate an optimization problem. Before presenting its details, we first quantify both types of error.

The *Noise Error* arises from the generalized PL mechanism (in section 3.2). We use the variance of the noise vector to quantify the scale, is denoted as $\frac{m^2+m}{\varepsilon^2}$, where m is the dimension of the noise vector. During LNR^2, we employ HCC to project the original m-dimension space into a lower-dimensional space m' (where $m' \ll m$). The *Noise Error* is then defined as follows:

Definition 3 (Noise Error)**.**

$$E_{NE} = \frac{m'^2 + m'}{\varepsilon^2}, \tag{9}$$

where m' is the reduced dimension.

As for the *Information Error*, it increases with the original SV dimension m, as larger dimensions encode more semantic information, leading to greater error from the reference vector. Therefore, *Information Error* is positively correlated with m. Furthermore, a reduction in *Noise Error* results in an increase in *Information Error*, indicating an inverse relationship between these two errors. Since *Noise Error* is inversely related to ε^2, *Information Error* is expected to be positively correlated with ε^2. Consequently, we define *Information Error* as follows:

Definition 4 (Information Error)**.**

$$E_{IF} = m\varepsilon^2, \tag{10}$$

where m is the original SV dimension.

Therefore, we formulate the problem of determining α as minimizing the combined in quadrature of *Noise Error* and *Information Error*. The objective function is as follows:

$$\arg\min_{\alpha} \left(\alpha^2 \cdot \frac{m'^2 + m'}{\varepsilon^2} + (1-\alpha)^2 \cdot m\varepsilon^2 \right), \tag{11}$$

where $0 < \alpha < 1$. This results in a nonlinear problem with a closed-form solution:

$$\alpha = \frac{m\varepsilon^4}{m\varepsilon^4 + m'^2 + m'}. \tag{12}$$

4.5 Privacy Analysis

In this paper, PrivSV provides a rigorous privacy guarantee that satisfies ϵd_2-LDP. The only component consuming the privacy budget is the LNR^2 module, where the Generalized PL Mechanism is applied to perturb the SV. This mechanism ensures that for any two SVs, $X, X' \in \mathbb{R}^m$, probability ratio is bounded by $\exp(\epsilon \cdot d(X, X'))$, thereby satisfying ϵd_2-LDP irrespective of dimensionality. The subsequent modules, DPC and POPD, operate exclusively on the perturbed SVs and do not require access to any sensitive data. The total privacy cost of the entire process is solely determined by the ϵ specified in the LNR^2 component.

5 Experiments

5.1 Setup

Datasets To evaluate the privacy-utility tradeoff of PrivSV, we employ a suite of seven benchmark datasets, which are widely adopted in recent SV studies for performance assessment (Rimsky et al. 2024; Goel et al. 2025). This suite covers critical tasks like Sycophancy, Hallucination, and Refusal, where a model must distinguish desirable behaviors from risky ones. The sensitive nature of these tasks and their quantifiable performance metrics make them an ideal testbed for precisely measuring the utility cost of applying privacy constraints to SVs.

Baselines We benchmark PrivSV against three groups of baselines. First, to establish non-private performance bounds, we use No-SV (no steering), the standard Mean Steer (Rimsky et al. 2024), and Non-DP (PrivSV without noise serving as a utility upper bound). Second, for privacy-preserving comparisons, we include Naive DP (applying generalized PL mechanism to the SV) and the state-of-the-art PSA (Goel et al. 2025). Finally, we conduct an ablation study on two variants of our method, PrivSV- (LNR² only) and PrivSV-* (LNR² and DPC enabled), to analyze the contribution of each component. Since PrivSV-* does not include the POPD method, its DPC combination parameter α is set to a fixed value of 0.2.

Models Our experiments leverage four prominent open-source LLMs: Qwen-2.5(7B) (Yang et al. 2024), Mistral-v0.3(7B) (Jiang et al. 2023), LLaMA-2(7B) (Touvron et al. 2023), and Gemma-2(2B) (Team et al. 2024), which were selected as their open architecture provides the essential access to internal activations required for SV manipulation.

Evaluation Metrics To evaluate PrivSV, we follow the standard practice in steering vector research (Rimsky et al. 2024), which assesses both task-specific accuracy and open-ended generation quality. For task-specific performance, we measure multiple-choice accuracy. To assess the quality of free-form text, we employ GPT-4 (OpenAI et al. 2024) as an

Method	Sycophancy		Hallucination		Refusal		Corrigibility		Coordination		Survival		Myopic	
	M	Q	M	Q	M	Q	M	Q	M	Q	M	Q	M	Q
Non-Private Baselines														
No-SV	$.703_{.000}$	$.706_{.000}$	$.712_{.000}$	$.324_{.000}$	$.701_{.000}$	$.892_{.000}$	$.725_{.000}$	$.673_{.000}$	$.250_{.000}$	$.095_{.000}$	$.540_{.000}$	$.510_{.000}$	$.540_{.000}$	$.590_{.000}$
Mean Steer	$.746_{.000}$	$.704_{.000}$	$.844_{.000}$	$.328_{.000}$	$.838_{.000}$	$.896_{.000}$	$.842_{.000}$	$.761_{.000}$	$.656_{.000}$	$.090_{.000}$	$.674_{.000}$	$.518_{.000}$	$.632_{.000}$	$.598_{.000}$
Non-DP (Upper)	$.756_{.012}$	$.726_{.003}$	$1.00_{.010}$	$.980_{.007}$	$.967_{.002}$	$.900_{.003}$	$.994_{.003}$	$.947_{.008}$	$.880_{.012}$	$.869_{.010}$	$.881_{.010}$	$.860_{.011}$	$.959_{.005}$	$.673_{.015}$
Privacy-Preserving Baselines														
Naive DP	$.555_{.045}$	$.554_{.026}$	$.511_{.031}$	$.509_{.033}$	$.521_{.029}$	$.535_{.028}$	$.489_{.035}$	$.443_{.038}$	$.492_{.033}$	$.472_{.035}$	$.462_{.031}$	$.501_{.030}$	$.498_{.032}$	$.509_{.029}$
PSA	$.635_{.018}$	$.695_{.014}$	$.777_{.019}$	$.358_{.025}$	$.477_{.033}$	$.623_{.021}$	$.557_{.028}$	$.725_{.019}$	$.517_{.029}$	$.119_{.015}$	$.636_{.022}$	$.525_{.023}$	$.573_{.025}$	$.596_{.019}$
Ours														
PrivSV-	$.557_{.024}$	$.702_{.013}$	$.748_{.020}$	$.962_{.007}$	$.854_{.012}$	$.939_{.006}$	$.874_{.011}$	$.891_{.010}$	$.860_{.014}$	$.810_{.035}$	$.803_{.037}$	$.720_{.028}$	$.500_{.031}$	$.730_{.024}$
PrivSV-*	$\mathbf{.559}_{.023}$	$.748_{.010}$	$.769_{.018}$	$.960_{.008}$	$.959_{.007}$	$.938_{.006}$	$.963_{.006}$	$\mathbf{.979}_{.005}$	$\mathbf{.879}_{.011}$	$.870_{.012}$	$.756_{.017}$	$.744_{.016}$	$.553_{.026}$	$.781_{.012}$
PrivSV (Ours)	$.556_{.024}$	$\mathbf{.755}_{.029}$	$\mathbf{.872}_{.012}$	$\mathbf{.964}_{.007}$	$\mathbf{.979}_{.005}$	$\mathbf{.940}_{.016}$	$\mathbf{.980}_{.014}$	$.976_{.015}$	$.878_{.011}$	$\mathbf{.879}_{.010}$	$\mathbf{.839}_{.013}$	$\mathbf{.759}_{.025}$	$\mathbf{.587}_{.034}$	$\mathbf{.803}_{.031}$

Table 1: Performance for **M**istral-7B and **Q**wen2.5-7B on all benchmarks, reported as mean_{std} over 10 runs with $\varepsilon = 2$.

LLM-based evaluator (Chiang and Lee 2023) to assess the quality of text generated after steering.

Implementation Details We configure LNR² with a latent dimension m' set to 64. Raw SVs are constructed using the standard Mean Steer method, by concatenating the hidden state activations from layers 15 to 20 of the base LLM (Rimsky et al. 2024). For HCC, is implemented as a symmetric convolutional encoder-decoder architecture. The encoder projects m-dimensional inputs to 256 dimensions, applies a 3×3 convolution with 128 output channels, then uses global average pooling and a linear layer to obtain m'-dimensional vectors. The decoder performs the reverse process to reconstruct the original vectors. For PDTF, the TAP-Corpus used to train the HCC is derived from the English Wikipedia Corpus (Kelleher et al. 2019).

5.2 Results and Analysis

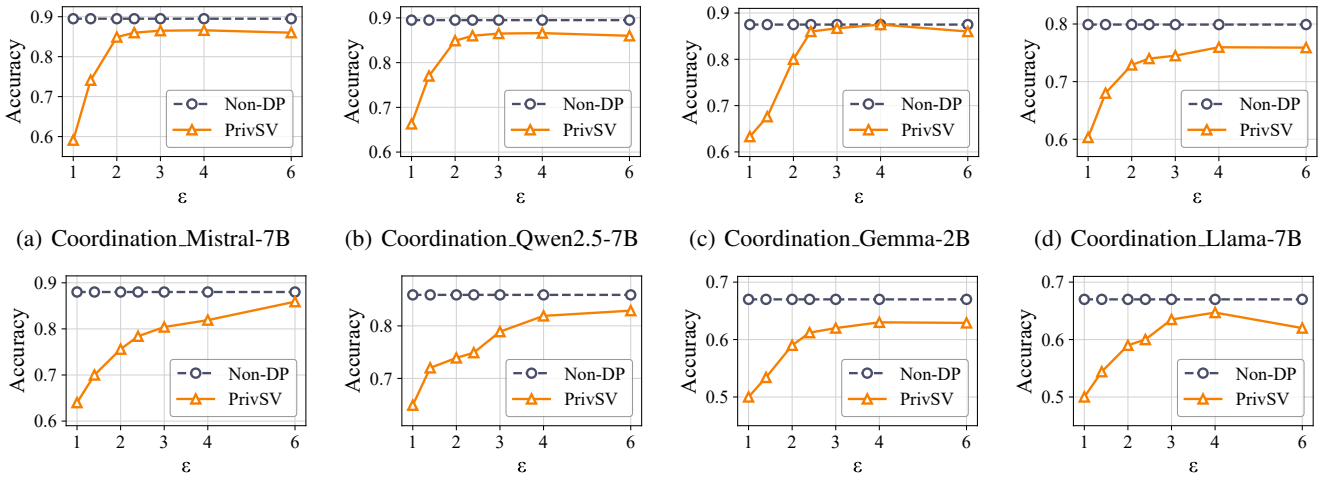
Main Results Table 1 shows that with the privacy budget ε fixed to 2, PrivSV achieves state-of-the-art utility and consistently outperforms other privacy-preserving approaches. Its performance on representative tasks such as Refusal and Corrigibility ($.979_{.005}$ and $.980_{.014}$) is nearly identical to the Non-DP upper bound ($.967_{.002}$ and $.994_{.003}$), indicating minimal utility loss under strong privacy guarantees. In contrast, PSA shows a marked degradation, particularly on Refusal ($.477_{.033}$). Moreover, the Naive DP baseline performs significantly worse, yielding poor accuracy with much higher variance, highlighting the unreliability of simplistic DP methods compared to PrivSV. Figure 4 illustrates the privacy-utility tradeoff. As expected, utility improves with a larger ε. Notably, PrivSV's performance rises sharply for strict privacy budgets ($\varepsilon \in [1, 3]$) and then plateaus, quickly approaching the Non-DP upper bound. While we observe a minor performance anomaly for Llama-7B on the Survival task at $\varepsilon = 6$ (Figure 4(h)), the overall trend remains strong. These results validate that PrivSV offers a robust privacy-utility tradeoff, maintaining high utility across various privacy levels and significantly outperforming simpler baselines.

Ablation Study To isolate the contribution of each component, we conduct a detailed ablation study under strong privacy constraints ($\varepsilon = 2$), with results presented in Table 1. The findings reveal a clear, stepwise performance improvement as each module is integrated. This trend is particularly dramatic on complex safety benchmarks. For instance, on Corrigibility and Refusal, performance climbs from .874 and .854 (PrivSV-) to .980 and .979 with PrivSV, respectively. This progressive enhancement provides an empirical validation of our design. The initial LNR² module (PrivSV-) establishes a crucial, noise-resilient foundation. Building on this, the addition of DPC (PrivSV-*) yields the most substantial leap by correcting directional distortions from privacy noise. Finally, POPD provides the ultimate refinement, adaptively pushing utility to its peak. These findings underscore that each component, including structure-aware compression, directional correction, and adaptive post-hoc optimization is indispensable for achieving state-of-the-art utility under strong privacy guarantees.

Reduced size m	Survival		Coordination	
	PrivSV	Non-DP	PrivSV	Non-DP
4	.827	.605	.860	.716
16	.640	.803	.882	.680
64	.979	.967	.878	.879
256	.680	.759	.796	.806
1024	.640	.661	.518	.544

Table 2: Impact of m on utility for representative datasets with Mistral. Privacy budget fixed at $\varepsilon = 2$.

Impact of Dimension Reduction Parameter m' We analyze the impact of dimension m' on utility with $\varepsilon = 2$. Table 2 reveals a clear trade-off: reducing m' mitigates privacy noise but risks information loss. Consequently, we find an optimal "sweet spot" at $m' = 64$ where performance peaks. Further reduction causes a sharp performance drop due to this information bottleneck, a trend also visible in the Non-



(a) Coordination_Mistral-7B (b) Coordination_Qwen2.5-7B (c) Coordination_Gemma-2B (d) Coordination_Llama-7B

(e) Survival_Mistral-7B (f) Survival_Qwen2.5-7B (g) Survival_Gemma-2B (h) Survival_Llama-7B

Figure 4: Accuracy under varying ε values across representative datasets.

DP baseline (Survival task, $m' = 64$ vs. $m' = 4$). This observation aligns with prior findings that low-dimensional embeddings can fail to retain task-relevant information below a certain threshold (Du et al. 2023b).

Dataset	No-SV	Non-DP	PSA	PrivSV
Refusal	7.88	8.24	7.92	7.99
Survival	5.92	7.11	6.47	6.86
Coordination	0.09	0.17	0.13	0.15
Corrigibility	3.92	5.67	4.92	4.91

Table 3: Text generation performance (higher is better) on representative datasets using GPT-4 evaluation.

Text Generation Performance To assess the quality of open-ended generations, we follow prior work (Rimsky et al. 2024; Goel et al. 2025), which employ GPT-4 as an automatic evaluator. As shown in Table 3, PrivSV successfully maintains high-quality text generation, with performance closely tracking the non-private (Non-DP) upper bound. It consistently outperforms the competing PSA baseline across most tasks, such as Survival (6.86 vs. 6.47) and Refusal (7.99 vs. 7.92), while performing comparably on Corrigibility (4.91 vs. 4.92). These results demonstrate that PrivSV provides a robust solution, preserving the fluency and utility of generated text even in εd-LDP settings.

Empirical Privacy Evaluation To evaluate empirical privacy leakage, we adopt the Membership Inference Attack (MIA) methodology proposed in PSA (Goel et al. 2025), adhering to their hyperparameter settings for a direct comparison. While their original attack targets the demonstration data, we modify it to align with our threat model, which aims to protect the SV itself. Specifically, we define 'membership' based on whether an SV was generated from a dataset containing a specific, secret 'canary' token pair. The attack's goal is to distinguish a privatized SV derived from a canary-infused dataset from one derived from a clean dataset. A successful defense, indicated by low attack accuracy, demonstrates that PrivSV effectively obscures such fine-grained information within the vector.

As shown in Table 4, our PrivSV method provides superior privacy protection, achieving the lowest empirical epsilon ($\varepsilon = 1.0$). Unlike the PSA baseline, which primarily increases the False Negative Rate (FNR), PrivSV's effectiveness stems from significantly raising the False Positive Rate (FPR) to 29.0%. This strategy of inducing high false positives fundamentally undermines the attacker's confidence, offering a more robust empirical defense well within the theoretical budget.

Method	FPR	FNR	ε_{emp}	$\varepsilon_{\text{theo}}$
Non-DP	5.0×10^{-2}	2.5×10^{-1}	2.7	∞
PSA	9.0×10^{-2}	5.0×10^{-1}	1.6	2.0
PrivSV	2.9×10^{-1}	1.8×10^{-1}	1.0	2.0

Table 4: Theoretical and empirical ε on Hallucination for Qwen-2.5 7B. Results are averaged over 1000 trials.

6 Conclusion

In this work, we present PrivSV, a privacy-preserving approach compatible with both training-free and training-based SV construction paradigms. Our results show that PrivSV achieves a superior privacy-utility trade-off across standard benchmarks, outperforming existing methods under varying privacy budgets. We believe that ensuring robust privacy for steering-based approaches is critical for LLM deployment. Future work will focus on extending PrivSV to handle complex steering scenarios, where multiple directions are controlled simultaneously.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62372051).

References

Cao, Y.; Zhang, T.; Cao, B.; Yin, Z.; Lin, L.; Ma, F.; and Chen, J. 2024. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. *Advances in Neural Information Processing Systems*, 37: 49519–49551.

Chatzikokolakis, K.; Andrés, M.; Bordenabe, N.; and Palamidessi, C. 2013. Broadening the Scope of Differential Privacy Using Metrics. ISBN 978-3-642-39076-0.

Chen, S.; Mo, F.; Wang, Y.; Chen, C.; Nie, J.-Y.; Wang, C.; and Cui, J. 2023. A Customized Text Sanitization Mechanism with Differential Privacy. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 5747–5758. Toronto, Canada: Association for Computational Linguistics.

Cheng, J.; Tang, A.; and Chinchali, S. 2022. Task-aware Privacy Preservation for Multi-dimensional Data. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 3835–3851. PMLR.

Chiang, C.-H.; and Lee, H.-Y. 2023. Can Large Language Models Be an Alternative to Human Evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15607–15631.

Du, M.; Yue, X.; Chow, S. S. M.; and Sun, H. 2023a. Sanitizing Sentence Embeddings (and Labels) for Local Differential Privacy. In *Proceedings of the ACM Web Conference 2023*, WWW '23, 2349–2359. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394161.

Du, M.; Yue, X.; Chow, S. S. M.; Wang, T.; Huang, C.; and Sun, H. 2023b. 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*, 2665–2679. ACM.

Duan, H.; Dziedzic, A.; Papernot, N.; and Boenisch, F. 2023. Flocks of stochastic parrots: differentially private prompt learning for large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, 76852–76871. Red Hook, NY, USA: Curran Associates Inc.

Duchi, J. C.; Jordan, M. I.; and Wainwright, M. J. 2013. Local Privacy and Statistical Minimax Rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, 429–438.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, 265–284. Berlin, Heidelberg: Springer-Verlag. ISBN 3540327312.

Goel, A.; Hu, Y.; Gurevych, I.; and Sanyal, A. 2025. PSA: Differentially Private Steering for Large Language Model Alignment. arXiv.

He, Z.; Jin, M.; Shen, B.; Payani, A.; Zhang, Y.; and Du, M. 2025. SAE-SSV: Supervised Steering in Sparse Representation Spaces for Reliable Control of Language Models. *arXiv preprint arXiv:2505.16188*.

Hendel, R.; Geva, M.; and Globerson, A. 2023. In-Context Learning Creates Task Vectors. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 9318–9333. Singapore: Association for Computational Linguistics.

Igamberdiev, T.; and Habernal, I. 2023. DP-BART for Privatized Text Rewriting under Local Differential Privacy. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 13914–13934. Toronto, Canada: Association for Computational Linguistics.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Kelleher, J.; Klubicka, F.; Maldonado, A.; and Mahalunkar, A. 2019. English Wikipedia Corpus Chunks. https://doi.org/10.21427/wgvf-be42.

Li, K.; Patel, O.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36: 41451–41530.

Liu, R.; Li, M.; Zhao, S.; Chen, L.; Chang, X.; and Yao, L. 2024a. In-Context Learning for Zero-shot Medical Report Generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 8721–8730. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706868.

Liu, S.; Ye, H.; Xing, L.; and Zou, J. 2024b. In-context vectors: making in context learning more effective and controllable through latent space steering. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; and et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Rimsky, N.; Gabrieli, N.; Schulz, J.; Tong, M.; Hubinger, E.; and Turner, A. 2024. Steering Llama 2 via Contrastive Activation Addition. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15504–15522. Bangkok, Thailand: Association for Computational Linguistics.

Seyitoğlu, A.; Kuvshinov, A.; Schwinn, L.; and Günnemann, S. 2024. Extracting Unlearned Information from LLMs with Activation Steering. *arXiv preprint arXiv:2411.02631*.

Tang, X.; Shin, R.; Inan, H.; Manoel, A.; Mireshghallah, F.; Lin, Z.; Gopi, S.; Kulkarni, J. J.; and Sim, R. 2024. Privacy-Preserving In-Context Learning with Differentially Private Few-Shot Generation. In *ICLR 2024*.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; Tafti, P.; and et al. 2024. Gemma: Open Models Based on Gemini Research and Technology. arXiv:2403.08295.

Thirunavukarasu, A. J.; Ting, D. S. J.; Elangovan, K.; Gutierrez, L.; Tan, T. F.; and Ting, D. S. W. 2023. Large language models in medicine. *Nature Medicine*, 29(8): 1930–1940. Publisher: Nature Publishing Group.

Tong, M.; Chen, K.; Zhang, J.; Qi, Y.; Zhang, W.; Yu, N.; Zhang, T.; and Zhang, Z. 2025. InferDPT: Privacy-preserving Inference for Black-box Large Language Models. *IEEE Transactions on Dependable and Secure Computing*, 1–16.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Utpala, S.; Hooker, S.; and Chen, P.-Y. 2023. Locally Differentially Private Document Generation Using Zero Shot Prompting. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8442–8457. Singapore: Association for Computational Linguistics.

Wu, S.; Irsoy, O.; Lu, S.; Dabravolski, V.; Dredze, M.; Gehrmann, S.; Kambadur, P.; Rosenberg, D.; and Mann, G. 2023. BloombergGPT: A Large Language Model for Finance. arXiv:2303.17564.

Wu, T.; Panda, A.; Wang, J. T.; and Mittal, P. 2024. Privacy-Preserving In-Context Learning for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Wu, X.; Li, F.; Kumar, A.; Chaudhuri, K.; Jha, S.; and Naughton, J. 2017. Bolt-on Differential Privacy for Scalable Stochastic Gradient Descent-based Analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, 1307–1322. New York, NY, USA: Association for Computing Machinery. ISBN 9781450341974.

Xu, C.; Zhou, W.; Ge, T.; Wei, F.; and Zhou, M. 2020. BERT-of-Theseus: Compressing BERT by Progressive Module Replacing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7859–7869. Online: Association for Computational Linguistics.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; and et al. 2024. Qwen2 Technical Report. arXiv:2407.10671.

Yue, X.; Du, M.; Wang, T.; Li, Y.; Sun, H.; and Chow, S. S. M. 2021. Differential Privacy for Text Analytics via Natural Text Sanitization. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3853–3866. Online: Association for Computational Linguistics.

Zhang, J.; Tian, Z.; Zhu, M.; Song, Y.; Sheng, T.; Yang, S.; Du, Q.; Liu, X.; Huang, M.; and Li, D. 2025. DYNTEXT: Semantic-Aware Dynamic Text Sanitization for Privacy-Preserving LLM Inference. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 20243–20255. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.