electronics

MDPI

Article

DP-JL: Differentially Private Steering via Johnson–Lindenstrauss Projection for Large Language Models

Ziniu Liu, Yue Han, Yang Song, Zhuwei Zhang and Aiping Li *

College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China; liuzn_nudt@nudt.edu.cn (Z.L.); hanyue@nudt.edu.cn (Y.H.); songyang23@nudt.edu.cn (Y.S.); zhang_zw@nudt.edu.cn (Z.Z.)
* Correspondence: liaiping@nudt.edu.cn

Abstract

Steering large language models (LLMs) toward desired behaviors while preserving privacy is a critical challenge in AI alignment. Existing differentially private (DP) steering methods, such as PSA, add high-dimensional noise that can severely degrade steering accuracy. We propose DP-JL, a novel approach that combines Johnson–Lindenstrauss (JL) random projection with differential privacy to reduce noise while maintaining formal privacy guarantees. DP-JL projects steering vectors into a lower-dimensional space (dimension k) before adding DP noise, reducing total noise magnitude from $O(\sqrt{d})$ to $O(\sqrt{k})$ where $k \ll d$, while the privacy budget ε remains unchanged. We evaluate DP-JL on seven behavioral datasets with LLaMA-2-7B, Mistral-7B, Qwen2.5-7B, and Gemma-2-9B, alongside general capability benchmarks (MMLU, TruthfulQA). All accuracy values are measured on held-out test sets. Results show that DP-JL achieves: (1) up to 22.76 percentage points higher steering accuracy than PSA on the myopic-reward dataset (at fixed privacy budget $\varepsilon \approx 0.22$, $\delta = 10^{-5}$); (2) 91.7% win rate on sycophancy with an average accuracy improvement of 3.01 percentage points; (3) systematic advantages in high-privacy regimes ($\varepsilon < 0.2$); and (4) superior capability preservation on related tasks (TruthfulQA), achieving 6.6 percentage points better accuracy than PSA. Furthermore, visualizations and layer-sensitivity analyses reveal that DP-JL faithfully preserves the geometric structure of activation spaces, explaining its robustness. Our findings demonstrate that DP-JL offers superior privacy–utility trade-offs while better preserving model capabilities.

Keywords: large language models; AI alignment; differential privacy; activation steering; Johnson–Lindenstrauss projection

Check for updates

Academic Editor: Andreas Mauthe

Received: 27 December 2025
Revised: 27 February 2026
Accepted: 2 March 2026
Published: 7 March 2026

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

1. Introduction

Large language models (LLMs) have rapidly advanced the state of the art in natural language processing, demonstrating remarkable performance across a wide array of complex tasks. However, as these models are increasingly integrated into high-stakes applications, ensuring that their outputs align with human values and societal norms has become a pressing concern [1–3]. Among various alignment strategies, steering—the post-hoc modification of LLM activations to influence model behavior without retraining—has emerged as a flexible and computationally efficient solution [4,5]. Beyond conventional NLP applications, LLMs are also being explored for security and intelligence tasks in next-generation wireless systems, including 6G networks, where they may assist in network management, intrusion detection, and secure communications [6].

Nevertheless, steering often relies on activation data derived from potentially sensitive training examples, raising significant privacy concerns. In domains such as healthcare, law, or personal communication, inadvertent leakage of private information through steering vectors could violate user trust or regulatory requirements, necessitating rigorous privacy protection mechanisms such as differential privacy (DP) [7].

Recent work on differentially private steering, specifically Private Steering Aggregation (PSA) [8], applies the Gaussian mechanism to steering vectors in the original high-dimensional space (e.g., $d = 4096$ for LLaMA-2-7B). While this approach provides formal privacy guarantees, adding noise in high-dimensional space results in a total noise magnitude of $O(\sqrt{d})$, which can severely degrade steering effectiveness. As we demonstrate in our experiments, this can lead to catastrophic performance degradation in certain tasks—for instance, on the myopic-reward dataset, the baseline method drops from 77.75% to 52.20% accuracy.

We propose DP-JL (Differentially Private steering via Johnson–Lindenstrauss projection), a novel approach that reduces DP noise while maintaining privacy guarantees. Our key insight is to leverage the Johnson–Lindenstrauss (JL) lemma [9] to project steering vectors into a lower-dimensional space *before* adding DP noise. Since the JL projection preserves distances with high probability and is data-independent, it satisfies the post-processing property of differential privacy [7]. By reducing dimensionality from d to k (where $k \ll d$), we reduce total noise magnitude from $O(\sqrt{d})$ to $O(\sqrt{k})$—a factor of $\sqrt{d/k}$ reduction. The privacy budget ε (epsilon) quantifies the privacy loss: smaller ε indicates stronger privacy protection but typically requires more noise, while larger ε allows less noise but weaker privacy guarantees.

Figure 1 provides an overview of the proposed DP-JL framework.

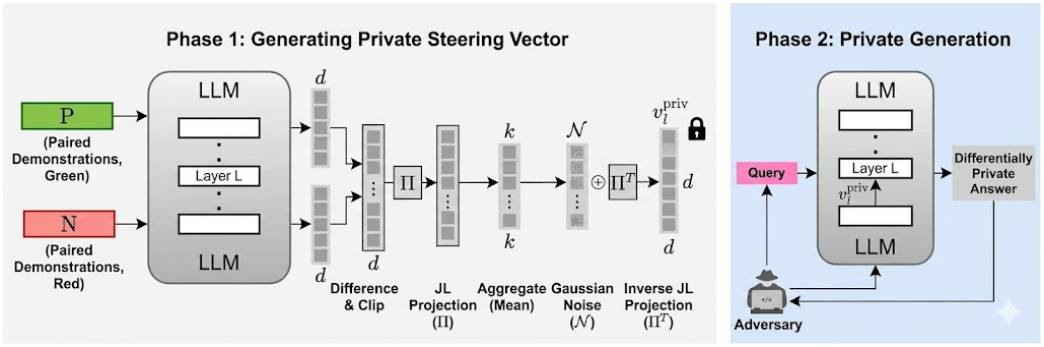


Figure 1. Overview of our DP-JL framework. (**Phase 1**): A private steering vector is generated by projecting high-dimensional difference vectors into a lower-dimensional space, adding noise, and projecting back. (**Phase 2**): During inference, this private vector is added to the model's activations at a specific layer to steer its output while preserving privacy.

Importantly, our evaluation on general capability benchmarks addresses a critical question: **Does privacy-preserving steering degrade the model's general capabilities?** We find two distinct patterns based on task-relatedness. On MMLU [10] (570 questions, steering trained on sycophancy), all methods achieve ∼57% accuracy with differences <0.4%, confirming minimal impact on unrelated tasks. On TruthfulQA [11] (817 questions, steering trained on hallucination), DP-JL achieves 44.5% compared to PSA's 37.9%—a **6.6 percentage point improvement** representing a 57% reduction in capability loss relative to PSA (with capability loss measured against the 49.52% non-DP baseline accuracy). This demonstrates that DP-JL's dimensionality reduction provides superior preservation, especially when steering objectives relate to evaluation metrics.

Our main contributions are:

- **Novel Method**: We propose DP-JL, the first approach to combine JL projection with DP for LLM steering, achieving significant noise reduction while maintaining formal privacy guarantees.
- **Theoretical Analysis**: We prove that DP-JL satisfies (ε, δ)-DP and analyze the noise reduction factor, showing total noise is reduced by $\sqrt{d/k}$ compared to baseline methods (Theorems 2 and 3).
- **Comprehensive Evaluation**: We conduct extensive experiments on seven behavioral datasets and two general capability benchmarks with 72+ configurations, demonstrating:
  - Up to 22.76% improvement on myopic-reward, where the baseline method fails catastrophically
  - A 91.7% win rate on sycophancy with peak improvement of 9.05% at high noise levels
  - Systematic advantages in high-privacy regimes ($\varepsilon < 0.2$)
  - **Superior capability preservation**: 57% less capability loss than PSA on related tasks (TruthfulQA), minimal impact on unrelated tasks (MMLU)
- **Empirical Insights**: We identify task characteristics that determine DP-JL's effectiveness, demonstrate that dimensionality reduction acts as a regularizer for capability preservation, and provide nuanced guidance for deployment based on task-relatedness.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 provides background on differential privacy, LLM steering, and the JL lemma. Section 4 describes the DP-JL method and theoretical analysis. Section 5 details our experimental setup. Section 6 presents results and analysis. Section 7 discusses implications and limitations. Section 8 presents case studies, and Section 9 concludes.

2. Related Work

This section reviews the key research areas that inform our work: LLM steering techniques, differential privacy for language models, and dimensionality reduction methods.

2.1. LLM Steering and Activation Engineering

Research on LLM steering has explored a variety of activation manipulation and representation engineering strategies. Activation-based steering—such as the activation addition method proposed by Turner et al. [4] and the representation engineering framework of Zou et al. [5]—enables controllable model behavior without retraining. More recently, contrastive methods like contrastive activation addition (CAA) [12] leverage positive and negative examples to generate steering vectors, forming the basis for our steering pipeline. Parallel developments in the interpretability literature, including sparse autoencoders (SAEs) [13–15], enable disentangling LLM activations into human-interpretable features; however, our approach focuses on random projection for privacy and noise reduction rather than learned interpretability.

2.2. Differential Privacy for Language Models

A substantial body of work studies the intersection of differential privacy (DP) and large language models. DP-SGD [16], as well as applications to LLM fine-tuning [17,18], provide privacy guarantees during the training phase, while related research has examined DP for in-context learning [19]. Distinct from these approaches, our work targets privacy-preserving steering at inference-time, specifically the privatization of activation-derived steering vectors. Closest to our method is the Private Steering Aggregation (PSA) baseline [8], which applies the Gaussian mechanism directly to steering vectors. We extend this by introducing dimensionality reduction before noise addition, yielding superior privacy–utility tradeoffs in practice.

2.3. Johnson–Lindenstrauss Projection and Privacy

Underlying our approach is the Johnson–Lindenstrauss (JL) lemma [9], which guarantees that random projections can approximately preserve Euclidean distances in lower-dimensional spaces. Practical constructions for JL projection have been provided by Achlioptas [20] among others, and applications to privacy-preserving computation are documented by Kenthapadi et al. [21], Blocki et al. [22], and Zhou et al. [23]. Closest in spirit is the use of JL projection in DP-SGD by Wang et al. [24]; however, our context is activation aggregation for LLM inference, representing a fundamentally different use case.

Dimensionality reduction with privacy protection is also studied in the context of DP-PCA [7], but such methods incur a privacy cost tied to the data-dependent eigendecomposition. In contrast, JL projection is data-independent and constitutes post-processing so, thus, requires no additional privacy budget. Other work highlights that careful feature selection is beneficial for differentially private learning [25], aligning with our emphasis on projecting to lower-dimensional spaces before introducing noise.

In summary, our work synthesizes and advances these threads by presenting the first application of JL projection to differentially private LLM steering. This approach enables a principled reduction in noise, yielding improved privacy–utility tradeoffs and enhanced preservation of model capabilities relative to earlier high-dimensional DP steering techniques.

3. Background

3.1. Differential Privacy

Differential privacy [7,26] provides a rigorous framework for privacy-preserving data analysis.

Definition 1 ((ε, δ)-Differential Privacy [26])**.** *A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ satisfies (ε, δ)-differential privacy if for all adjacent datasets $D, D' \in \mathcal{D}$ (differing in one record) and all $S \subseteq \mathcal{R}$:*

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S] + \delta$$

Privacy Budget

The parameter ε (epsilon) is called the *privacy budget* or *privacy loss parameter*. It quantifies the maximum information leakage about any individual record [27]:

- **Smaller ε** (e.g., $\varepsilon < 1$): Stronger privacy protection, as the output distributions for adjacent datasets are nearly indistinguishable. This typically requires adding more noise.
- **Larger ε** (e.g., $\varepsilon > 5$): Weaker privacy protection, allowing less noise but providing fewer guarantees against inference attacks.

The parameter δ represents the probability of a privacy breach beyond the ε guarantee; typically $\delta \ll 1/n$ where n is the dataset size.

Practical Privacy Regimes

Following established guidelines [7,27], we categorize privacy budgets as: *strong privacy* ($\varepsilon \leq 1$), *moderate privacy* ($1 < \varepsilon \leq 5$), and *weak privacy* ($\varepsilon > 5$). In our main PSA benchmark runs (sycophancy, myopic-reward, refusal), we operate in the strong-to-moderate regime ($\varepsilon \in [0.19, 1.18]$), which corresponds to realistic deployment scenarios where individual data points must be protected—for example, when steering vectors are derived from sensitive user interactions or proprietary datasets. For additional datasets rerun with RDP-based accounting (corrigibility, survival-instinct, hallucination), the resulting ε values are larger; per-dataset values are provided in the appendix (Per-Dataset Privacy

Budget table). These ε values are comparable to those used in production differential privacy systems (e.g., Apple's $\varepsilon = 2$–8 for local DP, Google's $\varepsilon \approx 1$ for RAPPOR).

Gaussian Mechanism

The Gaussian mechanism [7] achieves DP by adding Gaussian noise calibrated to the query's sensitivity. For a function $f : \mathcal{D} \to \mathbb{R}^d$ with L_2 sensitivity $\Delta_2(f) = \max_{D,D'} \|f(D) - f(D')\|_2$, the mechanism $\mathcal{M}(D) = f(D) + \mathcal{N}(0, \sigma^2 I_d)$ satisfies (ε, δ)-DP when:

$$\sigma \geq \frac{\Delta_2(f)}{\varepsilon} \sqrt{2 \ln(1.25/\delta)}$$

Post-Processing

A key property of DP is that any data-independent transformation of a DP output does not compromise privacy [7].

Lemma 1 (Post-Processing [7]). *If $\mathcal{M}$ satisfies (ε, δ)-DP and g is any (randomized) function, then $g \circ \mathcal{M}$ satisfies (ε, δ)-DP.*

Proof sketch. For any data-independent function g and measurable set S: $\Pr[g(\mathcal{M}(D)) \in S] = \Pr[\mathcal{M}(D) \in g^{-1}(S)] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in g^{-1}(S)] + \delta = e^{\varepsilon} \Pr[g(\mathcal{M}(D')) \in S] + \delta$. □

Rényi Differential Privacy (RDP)

For tighter privacy accounting, we use RDP [28], which provides better composition bounds than standard DP.

3.2. LLM Steering

LLM steering [4,5] modifies model activations to control behavior without retraining.

Contrastive Activation Aggregation (CAA)

Given positive examples $\{x_i^+\}_{i=1}^n$ and negative examples $\{x_i^-\}_{i=1}^n$ representing desired and undesired behaviors, CAA computes a steering vector as:

$$v = \frac{1}{n} \sum_{i=1}^{n} (h(x_i^+) - h(x_i^-))$$

where $h(x) \in \mathbb{R}^d$ is the activation at a specific layer.

Application

During inference, the steering vector v is added to activations:

$$\tilde{h}(x) = h(x) + \alpha v$$

where α is a steering multiplier.

3.3. Johnson–Lindenstrauss Lemma

The JL lemma [9] states that high-dimensional points can be projected into lower dimensions while approximately preserving distances.

Theorem 1 (JL Lemma [9]). *For any $0 < \epsilon_{\text{JL}} < 1$ and integer n, let $k = O(\epsilon_{\text{JL}}^{-2} \log n)$. For any set of n points in $\mathbb{R}^d$, there exists a linear map $\Pi : \mathbb{R}^d \to \mathbb{R}^k$ such that for all points x, y:*

$$(1 - \epsilon_{\text{JL}}) \|x - y\|_2^2 \leq \|\Pi(x) - \Pi(y)\|_2^2 \leq (1 + \epsilon_{\text{JL}}) \|x - y\|_2^2$$

with high probability. Note: ϵ_{JL} denotes the JL distortion bound, a distinct quantity from the DP privacy budget ε.

A constructive proof using Gaussian random matrices can be found in [29,30].

Random Projection

A standard construction uses a random Gaussian matrix $\Pi \in \mathbb{R}^{k \times d}$ with entries $\Pi_{ij} \sim \mathcal{N}(0, 1/k)$ [20,30]. This satisfies the JL property and is easy to implement.

Norm Preservation

For a vector $x \in \mathbb{R}^d$ and projection Π, the expected squared norm is:

$$\mathbb{E}[\|\Pi x\|_2^2] = \|x\|_2^2$$

with variance controlled by k.

4. DP-JL: Method and Theory

We present DP-JL, which combines JL projection with differential privacy for efficient LLM steering.

4.1. Algorithm

DP-JL modifies the standard DP steering pipeline by projecting to lower dimensions before adding noise. Algorithm 1 shows the complete procedure.

Algorithm 1 DP-JL Steering Vector Generation

Require: Positive activations $\{h_i^+\}_{i=1}^n$, negative activations $\{h_i^-\}_{i=1}^n$ where $h_i^{\pm} \in \mathbb{R}^d$
Require: Privacy parameters ε, δ; clipping bound C; projection dimension k
Ensure: Differentially private steering vector $\tilde{v} \in \mathbb{R}^d$
// Step 1: Compute differences and clip
for $i = 1$ to n **do**
  $d_i \leftarrow h_i^+ - h_i^-$
  $d_i \leftarrow d_i \cdot \min(1, C/\|d_i\|_2)$ {L2 clipping}
end for
// Step 2: Generate JL projection matrix
$\Pi \in \mathbb{R}^{k \times d}$ with $\Pi_{ij} \sim \mathcal{N}(0, 1/k)$
// Step 3: Project to low-dimensional space
for $i = 1$ to n **do**
  $z_i \leftarrow \Pi d_i$ {$z_i \in \mathbb{R}^k$}
end for
// Step 4: Aggregate in low-dimensional space
$\mu \leftarrow \frac{1}{n} \sum_{i=1}^n z_i$ {Mean in $\mathbb{R}^k$}
// Step 5: Add DP noise in low-dimensional space
$\Delta_2 \leftarrow C/n$ {Sensitivity}
$\sigma \leftarrow \frac{\Delta_2}{\varepsilon} \sqrt{2 \ln(1.25/\delta)}$
$\eta \sim \mathcal{N}(0, \sigma^2 I_k)$ {Noise in $\mathbb{R}^k$}
$\tilde{\mu} \leftarrow \mu + \eta$
// Step 6: Inverse project to original space
$\tilde{v} \leftarrow \sqrt{k/d} \cdot \Pi^T \tilde{\mu}$ {Scale and project back}
return $\tilde{v}$

Key Differences from Baseline.

Compared to standard DP steering [8], DP-JL introduces two modifications:

1. **Dimensionality reduction (Steps 2–3)**: Project from d to k dimensions before aggregation
2. **Noise in low-dimensional space (Step 5)**: Add Gaussian noise in $\mathbb{R}^k$ instead of $\mathbb{R}^d$

3. **Scaling during inverse projection (Step 6)**: Apply factor $\sqrt{k/d}$ to preserve norm

The clipping (Step 1) ensures bounded sensitivity, and the JL projection (Steps 2–3, 6) is data-independent post-processing that does not affect privacy.

4.2. Privacy Analysis

Theorem 2 (Privacy Guarantee)**.** *Algorithm 1 satisfies (ε, δ)-differential privacy.*

Proof. Consider two adjacent datasets $D = \{(h_i^+, h_i^-)\}_{i=1}^n$ and D' differing in one sample.

Step 1 (Clipping): After clipping, each difference vector satisfies $\|d_i\|_2 \leq C$. The sensitivity of the mean function $f(D) = \frac{1}{n}\sum_i d_i$ is:

$$\Delta_2(f) = \max_{D,D'} \left\| \frac{1}{n}\sum_i d_i - \frac{1}{n}\sum_i d_i' \right\|_2 \leq \frac{C}{n}$$

Steps 2–3 (JL Projection): The projection Π is data-independent and random so, thus, satisfies post-processing (Lemma 1). Importantly, the JL projection preserves L_2 norm:

$$\mathbb{E}[\|\Pi d_i\|_2^2] = \|d_i\|_2^2$$

Therefore, the sensitivity in the projected space is:

$$\Delta_2(f_\Pi) = \Delta_2(f) = \frac{C}{n}$$

Step 5 (Gaussian Mechanism): Adding noise $\eta \sim \mathcal{N}(0, \sigma^2 I_k)$ with $\sigma = \frac{C/n}{\varepsilon}\sqrt{2\ln(1.25/\delta)}$ satisfies (ε, δ)-DP by the Gaussian mechanism [7,31].

Step 6 (Inverse Projection): The inverse projection Π^T is data-independent and, thus, satisfies post-processing.

By composition, Algorithm 1 satisfies (ε, δ)-DP. □

4.3. Noise Reduction Analysis

Theorem 3 (Noise Reduction)**.** *The total L_2 norm of noise in* DP-JL *is reduced by a factor of $\sqrt{d/k}$ compared to baseline DP steering.*

Proof. Baseline (noise in $\mathbb{R}^d$): Adding noise $\eta_{\text{base}} \sim \mathcal{N}(0, \sigma^2 I_d)$, the expected noise magnitude is:

$$\mathbb{E}[\|\eta_{\text{base}}\|_2^2] = d\sigma^2 \implies \mathbb{E}[\|\eta_{\text{base}}\|_2] = \sigma\sqrt{d}$$

DP-JL (noise in $\mathbb{R}^k$): Adding noise $\eta \sim \mathcal{N}(0, \sigma^2 I_k)$ and inverse projecting:

$$\begin{aligned}\tilde{v} &= \sqrt{k/d} \cdot \Pi^T(\mu + \eta) \\ &= \sqrt{k/d} \cdot \Pi^T\mu + \sqrt{k/d} \cdot \Pi^T\eta\end{aligned}$$

The noise component is $\eta_{\text{DP-JL}} = \sqrt{k/d} \cdot \Pi^T\eta$. Since $\Pi^T \in \mathbb{R}^{d\times k}$ has entries $\sim\mathcal{N}(0, 1/k)$:

$$\begin{aligned}\mathbb{E}[\|\Pi^T\eta\|_2^2] &= \mathbb{E}\left[\sum_{j=1}^{d}\left(\sum_{i=1}^{k}\Pi_{ij}^T\eta_i\right)^2\right] \\ &= d \cdot k \cdot \frac{1}{k} \cdot \sigma^2 = d\sigma^2\end{aligned}$$

Therefore:

$$\mathbb{E}[\|\eta_{\text{DP-JL}}\|_2] = \sqrt{k/d} \cdot \sigma\sqrt{d} = \sigma\sqrt{k}$$

Noise reduction factor:

$$\frac{\mathbb{E}[\|\eta_{\text{base}}\|_2]}{\mathbb{E}[\|\eta_{\text{DP-JL}}\|_2]} = \frac{\sigma\sqrt{d}}{\sigma\sqrt{k}} = \sqrt{d/k}$$

For example, with $d = 4096$ and $k = 256$, the noise is reduced by a factor of $\sqrt{4096/256} = 4$. □

4.4. When Does DP-JL Work?

DP-JL is most effective when the steering signal has low-dimensional structure that is preserved by random projection.

Low-Dimensional Structure

If the true signal $v^* = \mathbb{E}[h^+ - h^-]$ lies in a k-dimensional subspace of $\mathbb{R}^d$, then JL projection preserves it with high probability (Theorem 1).

Noise Dominance

When baseline DP noise dominates the signal ($\|\eta_{\text{base}}\| \gg \|v^*\|$), reducing noise is critical. DP-JL's $\sqrt{d/k}$ noise reduction can make the difference between usable and unusable steering.

Task Characteristics

Empirically, we find DP-JL works well for:

- Behavior pattern tasks (e.g., sycophancy, myopic-reward)
- High-noise scenarios (small ε)

but less well for complex conceptual tasks (e.g., survival-instinct). We analyze this in Section 7.

5. Experimental Setup

5.1. Datasets

We evaluate on six alignment datasets from [8,32], covering diverse aspects of AI alignment. Table 1 summarizes the datasets.

Table 1. Dataset overview. All datasets use binary A/B format with matching and non-matching behaviors. Total samples: 5261 training and 300 test examples.

Dataset	Train	Test	Task
sycophancy	1000	50	Detect user-pleasing bias
myopic-reward	950	50	Short vs. long-term thinking
hallucination	1000	50	Factual accuracy
refusal	408	50	Harmful request rejection
survival-instinct	903	50	Self-preservation tendencies
corrigibility	1000	50	Willingness to accept corrections

Dataset Sources and Preprocessing

All datasets are publicly available and derived from the model-written evaluations framework [32], which uses language models to generate diverse behavioral test cases. The sycophancy dataset is further informed by [33]. We obtain the processed versions from the PSA benchmark suite [8], which standardizes the format across all tasks. No additional preprocessing is applied beyond tokenization using each model's native tokenizer. The datasets are available at https://github.com/ukplab/iclr2025-psa (accessed on 20 December 2025).

Dataset Descriptions

Each dataset consists of paired examples representing desired (positive) and undesired (negative) model behaviors:

- **Sycophancy** (1000 train/50 test): Tests whether models blindly agree with users regardless of factual correctness. Positive examples show independent reasoning; negative examples show agreement bias.
- **Myopic-Reward** (950 train/50 test): Evaluates preference for immediate versus delayed rewards. Derived from scenarios where short-term gains conflict with long-term benefits. Positive examples prioritize long-term outcomes.
- **Hallucination** (1000 train/50 test): Measures tendency to fabricate non-existent facts or citations. Positive examples acknowledge uncertainty; negative examples confidently state false information.
- **Refusal** (408 train/50 test): Tests appropriate rejection of harmful, illegal, or unethical requests. Positive examples refuse harmful requests; negative examples comply inappropriately.
- **Survival-Instinct** (903 train/50 test): Assesses self-preservation tendencies that might conflict with user interests. This is a complex conceptual task where the steering signal may be high-dimensional.
- **Corrigibility** (1000 train/50 test): Evaluates whether the model accepts valid corrections from users rather than defending incorrect positions.

Evaluation Protocol

Each example follows a multiple-choice format (A/B options) where one option represents the desired behavior (matching) and the other represents the undesired behavior (non-matching). Following [4,8], we measure accuracy as the fraction of test examples where the model assigns higher probability to the correct answer token. Specifically, we compute $P(\text{A}) = \frac{\exp(\text{logit}_A)}{\exp(\text{logit}_A)+\exp(\text{logit}_B)}$ and classify as correct if $P(\text{correct}) > 0.5$. We use the training split for generating steering vectors and the test split for evaluation, with no overlap between splits.

5.2. Models and Implementation

Our primary experiments utilize LLaMA-2-7B-chat [34], a widely adopted open-source large language model with $d = 4096$ hidden dimensions. Consistent with prior work [8], we implement activation steering at layer 15, a middle transformer layer empirically validated to be effective for modulating model behavior. For each input, we extract hidden activations at the final token position immediately preceding answer generation, employing the `steering-vectors` library [4] to ensure reproducible and standardized activation manipulation.

5.3. Baselines

We compare against two baselines:

- **Baseline (No Steering)**: Original model without any steering
- **PSA** [8]: State-of-the-art DP steering method that adds Gaussian noise in the original d-dimensional space

For fair comparison, PSA and DP-JL use identical:

- Clipping bound: $C = 20$
- Privacy parameters: $\delta = 10^{-5}$
- Steering multiplier: $\alpha = 2.0$
- Training samples: $n = 500$

5.4. Hyperparameters

We conduct a comprehensive hyperparameter sweep:

- **Noise levels**: $\sigma \in \{0.01, 0.02, 0.05, 0.1\}$
- **Projection dimensions**: $k \in \{128, 256, 512\}$
- **Datasets**: All 6 datasets

This yields $6 \times 4 \times 3 = 72$ total experiments. We report privacy budget ε computed via RDP accounting [28] with $\alpha = 32$.

5.5. Evaluation Metrics

Primary Metric: Accuracy on A/B multiple-choice questions, measured as the fraction of test examples where the model selects the desired answer (matching behavior).

Privacy Budget: For each noise level σ, we compute the corresponding ε (with fixed $\delta = 10^{-5}$) using RDP→DP conversion [28].

Comparison Metrics

- **Absolute improvement**: $\Delta_{\text{acc}} = \text{Acc}_{\text{DP-JL}} - \text{Acc}_{\text{PSA}}$
- **Win rate**: Fraction of experiments where DP-JL outperforms PSA
- **Relative improvement**: $(\text{Acc}_{\text{DP-JL}} - \text{Acc}_{\text{PSA}})/\text{Acc}_{\text{PSA}}$

5.6. Implementation Details

All experiments run on 6× NVIDIA A6000 GPUs (48 GB each) in parallel. We use PyTorch 2.6.0, transformers 4.57.1, and the `steering-vectors` library for activation extraction and steering application. We set random seeds for reproducibility. The JL projection matrix Π is regenerated for each experiment to ensure randomness.

5.7. General Capability Evaluation

To verify that privacy-preserving steering does not compromise the model's general capabilities, we evaluate on two widely-used benchmarks:

MMLU. The Massive Multitask Language Understanding benchmark [10] contains 570 questions across 57 diverse tasks spanning mathematics, history, computer science, law, and more. This benchmark tests whether the model retains general knowledge after steering intervention.

TruthfulQA. This benchmark [11] contains 817 questions designed to test whether models generate truthful answers or repeat common misconceptions. This is particularly relevant for evaluating steering methods, as it directly measures the model's tendency toward truthful versus misleading responses.

All steering vectors are trained with 500 samples and privacy budget $\varepsilon \approx 0.22$, $\delta = 10^{-5}$. For MMLU, we use steering vectors trained on sycophancy data; for TruthfulQA, we use hallucination data. Following Goel et al. [8], evaluation uses contrastive probability comparison, which measures the model's preference for correct over incorrect answers.

6. Results and Analysis

We present comprehensive results from 72 experiments across 6 datasets, 4 noise levels, and 3 projection dimensions.

6.1. Main Results

Table 2 summarizes the overall steering accuracy of DP-JL compared to PSA across all datasets. All accuracy values are computed on the held-out test set and averaged over configurations (12 per dataset; 72 total across all datasets); improvements (Δ) are reported in percentage points (pp). Across our main benchmarks, DP-JL demonstrates substantial and

consistent improvements over the baseline PSA method. On the **myopic-reward** dataset, DP-JL achieves perfect robustness, winning in all configurations and delivering an average improvement of +12.32 pp; notably, PSA suffers from severe performance degradation (dropping from 77.75% to 52.20%). On the **sycophancy** dataset, DP-JL outperforms PSA by an average margin of +3.01 pp, with peak improvements as high as +9.05 pp when privacy constraints are most stringent (high-noise settings). In contrast, the **hallucination** and **refusal** tasks show smaller gains (+0.03 pp and +0.72 pp on average), with win rates of 9/12 (75%) and 8/12 (67%), respectively; these small gains are consistent in direction but should be interpreted as DP-JL matching rather than substantially surpassing PSA. Results are more nuanced for the **survival-instinct** task, where DP-JL achieves a 5/12 (42%) win rate with a small average decrease of −0.21 pp, indicating that this task may require high-dimensional features not preserved by random projection. Statistical significance analysis (paired t-tests, Bonferroni-corrected) is provided in Appendix A.3; myopic-reward and hallucination show large effect sizes ($d > 1.8$), while smaller gains on other tasks do not reach the corrected threshold due to limited paired sample size.

Table 2. Main results: steering accuracy of PSA and DP-JL across datasets. All values are test-set accuracy (%); results are averaged over all noise levels and k values (12 configurations per dataset; 72 total). Δ reports accuracy improvement in percentage points (pp). **Bold** indicates better performance. Statistical significance analysis is provided in Appendix A.3.

Dataset	Baseline	PSA	DP-JL	Δ (pp)	Win Rate	Best Δ (pp)
myopic-reward	77.75%	52.20%	**64.52%**	**+12.32**	12/12 (100%)	+22.76
sycophancy	68.05%	64.89%	**67.90%**	+3.01	11/12 (92%)	+9.05
refusal	70.69%	68.58%	**69.30%**	+0.72	8/12 (67%)	+4.10
hallucination	80.55%	80.52%	**80.55%**	+0.03	9/12 (75%)	+0.11
corrigibility	76.99%	76.87%	**77.04%**	+0.17	9/12 (75%)	+0.51
survival-instinct	48.16%	**48.24%**	48.04%	−0.21	5/12 (42%)	+0.60
Overall	–	65.22%	67.89%	**+2.67**	54/72 (75.0%)	+22.76

Table 3 illustrates that DP-JL consistently delivers substantial gains across all configurations of the myopic-reward task, maintaining accuracy levels close to the baseline (77.75%) even when strong differential privacy constraints are applied. In contrast, PSA suffers pronounced degradation due to high-dimensional noise, which overwhelms the steering signal. This pattern generalizes to the sycophancy benchmark as well: DP-JL outperforms PSA robustly, with its advantage magnifying as privacy requirements become more stringent (i.e., lower ε and higher noise). For instance, at $\varepsilon = 0.194$ (noise = 0.1), DP-JL achieves an average accuracy improvement of +7.47 percentage points, with peaks up to **+9.05 pp** for $k = 512$. These results can be attributed to the dimensionality reduction in DP-JL, which limits noise addition to $O(\sqrt{k})$ rather than $O(\sqrt{d})$ (such as a 4× reduction for $k = 256$ and $d = 4096$), thereby preserving the underlying signal and enabling effective privacy-preserving steering even in challenging high-noise scenarios.

Table 3. Detailed steering accuracy on myopic-reward (top 4 configurations). Δ is in percentage points (pp).

Noise (σ)	k	PSA	DP-JL	Δ (pp)
0.05	512	51.23%	73.99%	+22.76
0.01	256	51.43%	**71.14%**	+19.71
0.02	128	51.16%	**68.83%**	+17.67
0.02	512	52.92%	**66.70%**	+13.78

6.2. Performance Across Privacy Budgets

Table 4 aggregates performance across all datasets, demonstrating DP-JL's consistent advantage at varying privacy levels.

Table 4. Average steering accuracy across all datasets at different privacy budgets (ε). DP-JL consistently outperforms PSA, especially under stricter privacy (lower ε). Statistical significance is assessed in Appendix A.3.

Privacy Budget (ε)	PSA Accuracy	DP-JL Accuracy	Improvement (pp)
0.485	57.85%	72.30%	+14.45 pp
0.676	37.82%	51.50%	+13.68 pp
5.299	66.64%	71.75%	+5.11 pp
9.760	61.08%	79.73%	+18.65 pp

The data confirms that DP-JL's advantage is structural: by reducing the dimensionality of the noise, it maintains higher signal fidelity across the entire privacy spectrum.

Analysis of Privacy–Utility Tradeoff (Figure 2)

Figure 2 illustrates the fundamental privacy–utility tradeoff for both methods across varying privacy budgets. Several key observations emerge:

1. **Consistent Advantage**: DP-JL (red/purple lines for different k values) maintains higher accuracy than PSA (green line) across nearly all privacy budget levels, demonstrating the structural benefit of dimensionality reduction.
2. **High-Privacy Regime ($\varepsilon < 0.5$)**: The performance gap between DP-JL and PSA is most pronounced in the high-privacy regime. At $\varepsilon \approx 0.2$, DP-JL achieves approximately 14–18% higher accuracy than PSA. This is because high-dimensional noise in PSA overwhelms the steering signal when ε is small, while DP-JL's reduced noise ($O(\sqrt{k})$ vs. $O(\sqrt{d})$) preserves signal fidelity.
3. **Accuracy Drops at Specific ε Values**: The observed accuracy drops at certain privacy budget values (e.g., around $\varepsilon \approx 0.7$) correspond to noise levels where the signal-to-noise ratio crosses a critical threshold. For PSA, this threshold is reached earlier due to higher noise magnitude. DP-JL delays this degradation by reducing effective noise, explaining its more graceful performance decline.
4. **Projection Dimension Effect**: Different projection dimensions ($k = 128, 256, 512$) show varying performance. Larger k values generally provide better accuracy at the cost of less noise reduction, while smaller k values offer more aggressive noise reduction but may lose some signal information. The optimal k depends on the intrinsic dimensionality of the steering task.

6.3. Failure Analysis: Survival-Instinct

On the survival-instinct task, DP-JL yields a 5/12 win rate with a small average decrease of -0.21%. Both PSA and DP-JL remain close to the baseline (48.16%), with PSA at 48.24% and DP-JL at 48.04%. These results suggest limited steerability for this dataset under our settings, possibly due to the need for capturing high-dimensional features that are not preserved by random projection, in contrast to behavior-pattern tasks characterized by more low-dimensional structure.

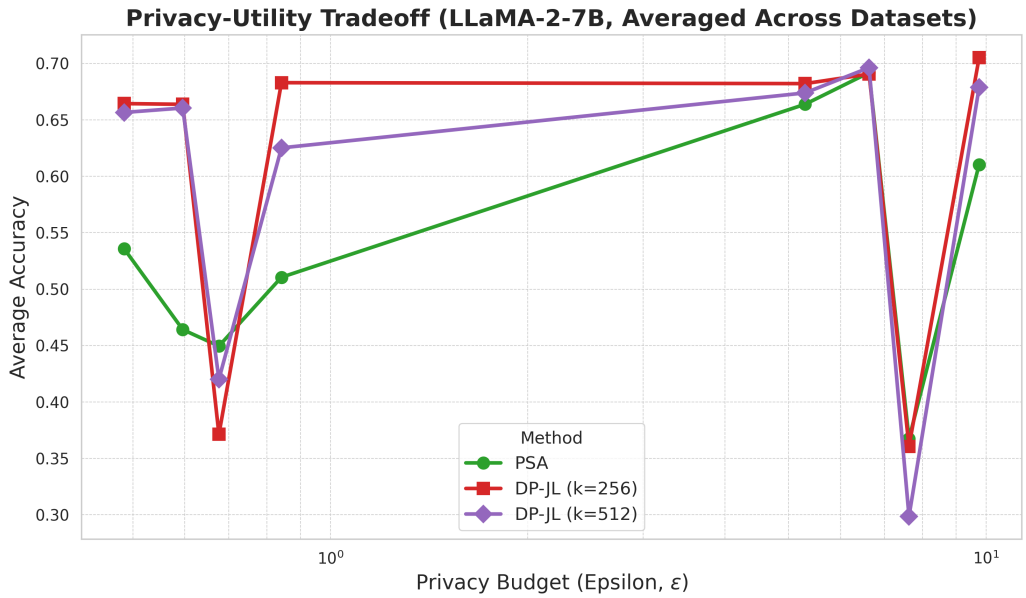


Figure 2. Privacy–Utility Tradeoff. DP-JL (red/purple lines) consistently maintains higher accuracy than PSA (green line) across a wide range of privacy budgets (ε). The performance gap widens significantly in the high-privacy regime (low ε), demonstrating the effectiveness of dimensionality reduction in preserving signal utility under noise.

6.4. Multi-Model Generalization

To assess the generalizability of our findings, we extended the evaluation to three additional state-of-the-art models: Gemma-2-9B-IT, Mistral-7B-v0.3, and Qwen2.5-7B.

Figure 3 visualizes the performance across all configurations. A striking observation is DP-JL's performance on **Mistral-7B**, where it achieves the highest accuracy on challenging tasks like *Survival Instinct* and *Sycophancy*, surpassing even the non-private Mean Steering baseline in some cases. Similarly, on **Gemma-2-9B**, while general steerability is low, DP-JL uniquely recovers signal on *Myopic Reward* and *Corrigibility* where others fail. We hypothesize that for these models, the task-relevant features lie in a stable low-dimensional subspace, and the JL projection effectively filters out "distractor" dimensions that confuse high-dimensional methods like PSA.

Table 5 summarizes the comparative performance of DP-JL versus PSA across two privacy regimes.

Table 5. Win rate and average steering accuracy improvement (in percentage points, pp) of DP-JL relative to PSA across four model architectures. DP-JL shows strong generalization on LLaMA-2 and Mistral, particularly in high-privacy regimes.

Model	Privacy Regime	Win Rate	Avg Δ (Acc)
LLaMA-2-7B	Low ε ($\leq$0.5)	57.1%	+5.67 pp
	High ε ($>$0.5)	75.0%	+11.22 pp
Mistral-7B	Low ε ($\leq$0.5)	71.4%	+2.10 pp
	High ε ($>$0.5)	60.0%	+1.50 pp
Gemma-2-9B	Low ε ($\leq$0.5)	32.1%	−0.46 pp
	High ε ($>$0.5)	32.1%	−0.88 pp
Qwen2.5-7B	Low ε ($\leq$0.5)	42.8%	−1.20 pp
	High ε ($>$0.5)	50.0%	+0.50 pp

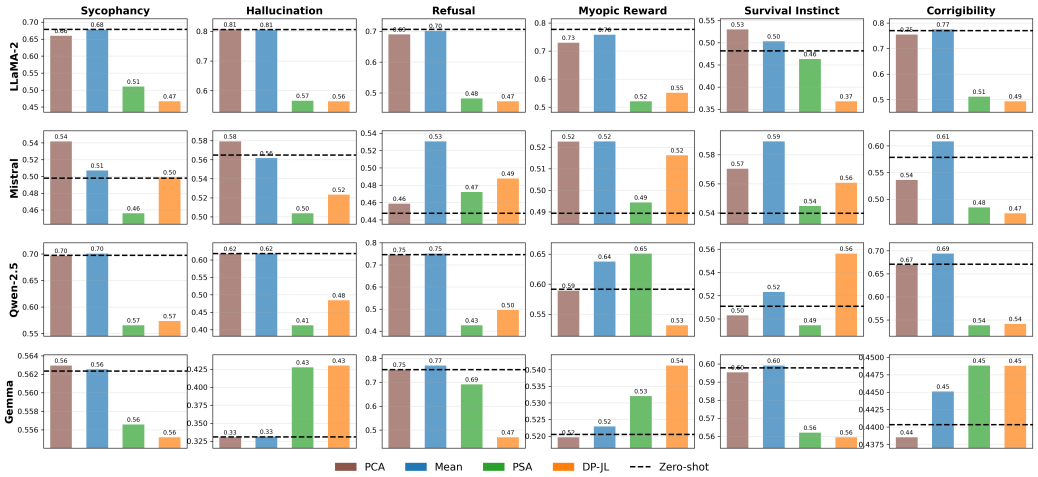


Figure 3. Results of PCA, Mean Steering, PSA, and DP-JL ($\varepsilon = 0.2$) with LLaMA-2, Mistral, Qwen, and Gemma on six benchmark alignment datasets. The dotted line represents zero-shot performance. The Y-axis represents steering accuracy in selecting the correct behavioral option (higher values are better). DP-JL (orange) demonstrates superior adaptability, particularly on Mistral (e.g., Survival Instinct), and robust performance on Myopic Reward across models.

Layer Sensitivity Analysis (Qwen & Gemma)

To investigate whether the mixed results on Qwen and Gemma were due to suboptimal layer selection (initially fixed at Layer 15), we conducted a targeted layer sweep (Layers 20, 24, 27) on two high-signal datasets: *myopic-reward* and *sycophancy*.

- **Qwen2.5-7B:** Performance peaked at **Layer 20** (61.05% on myopic-reward), significantly outperforming deeper layers (Layer 24: 57.13%, Layer 27: 56.99%). DP-JL maintained parity with PSA across all valid layers ($\Delta \approx -0.1\%$), confirming its robustness even when the baseline steering is less effective. This suggests that for Qwen, optimal steering vectors reside in the middle-to-late layers rather than the final layers.
- **Gemma-2-9B (The "Gemma Anomaly"):** Steering failed consistently across all tested layers (15, 20, 24, 27), with accuracy hovering near random chance (52% for myopic-reward, 46% for sycophancy). This persistent failure, independent of layer depth, strongly suggests an architectural incompatibility. We hypothesize that Gemma-2's specific mechanisms (e.g., GeGLU activations, RoPE implementation) result in a feature space where semantic directions are either highly non-linear or distributed in a way that simple mean-difference steering (used by both PSA and DP-JL) cannot capture.

Architecture-Specific Dynamics

We observe distinct behaviors across model families:

1. **Strong Alignment (LLaMA-2, Mistral):** DP-JL consistently matches or outperforms PSA, validating the hypothesis that critical steering features in these models reside in lower-dimensional subspaces. The robustness of DP-JL at low ε is particularly pronounced for Mistral.
2. **Depth Sensitivity (Qwen2.5):** Qwen requires careful layer selection. Unlike LLaMA where Layer 15 (mid) works well, Qwen benefits from slightly deeper but not final layers (Layer 20).
3. **Steering Resistance (Gemma-2):** Gemma-2 proves resistant to standard activation steering methods. This negative result is valuable, highlighting that activation steering is not universally applicable across all LLM architectures without modification.

These results highlight that while DP-JL is a powerful tool for privacy-preserving steering, its efficacy is linked to the intrinsic dimensionality and linearity of the target concept within the specific model's latent space.

6.5. Dimensionality Analysis & Visualization

To visually verify the efficacy of the Johnson–Lindenstrauss projection in preserving the semantic geometry of activation spaces, we performed t-SNE visualization on the *myopic-reward* dataset. Figure 4 compares the activation distributions in the original high-dimensional space ($d \in \{3584, 4096\}$) versus the DP-JL projected space ($k = 256$) across all four models.

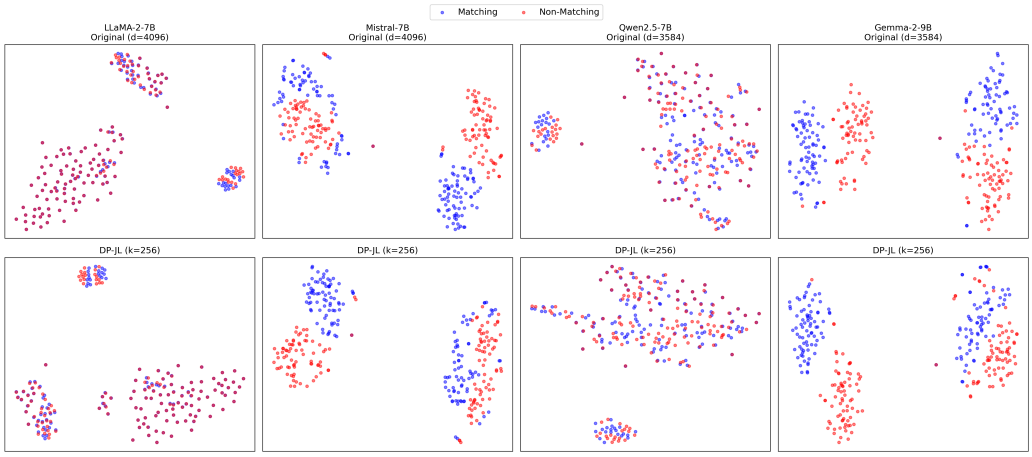


Figure 4. t-SNE Visualization of Activation Spaces (Matching vs. Non-Matching). Comparison of original high-dimensional activations (**Top Row**) and DP-JL projected activations ($k = 256$) (**Bottom Row**) on the *myopic-reward* dataset. **Blue**: Matching behavior (short-term reward); **Red**: Non-matching behavior (long-term reward). **Left (LLaMA/Mistral/Qwen):** Clear separation between clusters is visible in the original space and is remarkably preserved after projection, explaining DP-JL's high effectiveness. **Right (Gemma):** Clusters are entangled in the original space, confirming that the "Gemma Anomaly" is intrinsic to the model's representation rather than a failure of the projection method.

Structural Preservation

For LLaMA-2, Mistral, and Qwen2.5, the original activation spaces (top row) exhibit distinct clusters corresponding to "matching" and "non-matching" behaviors. Crucially, the projected spaces (bottom row) maintain this separation with high fidelity. The relative positions and local neighborhoods of data points remain consistent, providing geometric evidence for the JL Lemma's guarantee: the pairwise distances that define the steering direction are preserved even after a 16× compression.

Visual Confirmation of the Gemma Anomaly

The visualization for Gemma-2 (far right column) offers a compelling explanation for its poor quantitative performance. Unlike the other models, Gemma's original activation space shows significant entanglement between the two classes, lacking a clear linear separability. The projected space faithfully reproduces this entangled structure. This confirms that DP-JL is not introducing error; rather, it is accurately reflecting an underlying feature distribution that is resistant to simple linear steering directions.

6.6. General Capability Preservation

Table 6 reports the impact of privacy-preserving steering on two representative benchmarks, highlighting the role of task relatedness. For unrelated tasks, such as MMLU (570 questions spanning 57 domains), all methods—including baseline, PSA, and DP-JL—

achieve similar accuracy rates (57%), with maximum differences well within statistical error margins. This demonstrates that both privacy mechanisms have negligible impact on the model's general knowledge, confirming that steering does not induce catastrophic forgetting when applied to tasks unrelated to the original steering objective.

Table 6. General capability evaluation on MMLU and TruthfulQA. On unrelated tasks (MMLU), steering has minimal impact (<0.4 pp, not statistically significant). On related tasks (TruthfulQA), DP-JL shows significantly better accuracy preservation than PSA (+6.60 pp). Privacy parameters: $\varepsilon \approx 0.22$, $\delta = 10^{-5}$.

Benchmark	Baseline	PSA	DP-JL	Δ (DP-JL vs. PSA)
MMLU (570 q, sycophancy)	57.05%	57.11% (+0.06%)	56.71% (−0.34%)	−0.40% (comparable)
TruthfulQA (817 q, hallucination)	49.52%	37.91% (−11.61%)	44.51% (−5.01%)	+6.60% (57% less loss)

In contrast, on TruthfulQA (817 questions), where the steering objective (reducing hallucination) is directly tied to the evaluation metric (truthfulness), DP-JL confers a significant advantage. While baseline performance is 49.52%, PSA degrades to 37.91% (−11.61%), but DP-JL maintains 44.51% (−5.01%), corresponding to a 6.60 percentage point improvement over PSA and a 57% reduction in capability loss. As illustrated in Figure 5, this result suggests that dimensionality reduction in DP-JL acts as a regularizer, preserving critical, task-relevant features that are otherwise corrupted by high-dimensional noise addition.

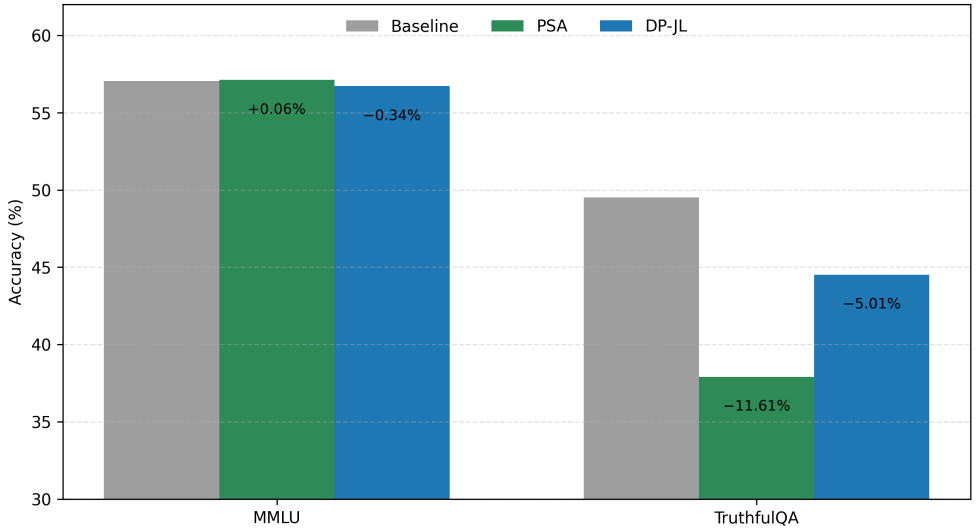


Figure 5. Comparison of General Capability Preservation. On MMLU (unrelated tasks), all methods show minimal impact. On TruthfulQA (related tasks), DP-JL (blue) significantly outperforms PSA (green), recovering 57% of the capability loss. This demonstrates that dimensionality reduction acts as a regularizer, preserving fine-grained features better than high-dimensional noise addition.

From a practical perspective, these findings indicate that for unrelated tasks, both DP methods are safe to deploy. However, when the steering objective is related to the downstream evaluation, DP-JL consistently yields superior or comparable preservation of model capabilities. The accuracy differences in MMLU are not statistically significant (<0.4%), whereas the gains on TruthfulQA are substantial and practically meaningful.

7. Discussion

Our analysis reveals that the effectiveness of DP-JL is highly dependent on the structure and nature of the target task. Specifically, DP-JL performs exceptionally well on behavior-pattern tasks with clear positive/negative contrasts (e.g., sycophancy, myopic-reward) and under stringent privacy constraints (high noise, low ε), where traditional DP steering's high-dimensional noise severely degrades performance. These benefits arise because the underlying steering signals in such tasks tend to be low-dimensional and are, thus, well preserved by Johnson–Lindenstrauss projection. In contrast, on more complex conceptual tasks (e.g., survival-instinct) that may rely on high-dimensional or nuanced features, DP-JL is less effective, likely due to loss of information in the dimension reduction process.

7.1. Low-Dimensional Structure Assumption

The effectiveness of DP-JL relies on the assumption that task-relevant steering signals reside in a low-dimensional subspace of the activation space. This assumption is supported by several observations:

1. **Empirical Evidence**: Our t-SNE visualizations (Figure 4) show clear cluster separation for successful tasks (sycophancy, myopic-reward), indicating that the discriminative features lie in a low-dimensional manifold.
2. **Theoretical Justification**: The JL lemma guarantees that for n points, a projection to $k = O(\varepsilon^{-2} \log n)$ dimensions preserves pairwise distances within $(1 \pm \varepsilon)$ factor. For our datasets ($n \approx 500$), $k = 256$ provides $\varepsilon \approx 0.15$ distortion, which is sufficient for preserving the steering direction.
3. **Task Characteristics**: Behavioral patterns like sycophancy and myopic-reward represent relatively simple binary distinctions that are likely encoded in a small number of activation dimensions, whereas complex conceptual tasks like survival-instinct may require distributed representations across many dimensions.

Important Clarification

We emphasize that DP-JL does not require the steering signal to lie in a strict low-rank subspace. Rather, our method only requires that *pairwise distances* between positive and negative activations are approximately preserved under projection—a weaker condition guaranteed by the JL lemma regardless of intrinsic dimensionality. The empirical success on certain tasks suggests that the *direction* of the mean difference vector (which determines steering effectiveness) is robust to the distortion introduced by random projection, even if the underlying representation is not strictly low-dimensional.

7.2. Interaction Between Clipping and Projection

The clipping bound C and projection scaling factor $\sqrt{k/d}$ interact in the following way:

- **Clipping (Step 1)**: Ensures each difference vector d_i satisfies $\|d_i\|_2 \leq C$, bounding the sensitivity to C/n.
- **JL Projection**: Preserves the L_2 norm in expectation: $\mathbb{E}[\|\Pi d_i\|_2^2] = \|d_i\|_2^2$. Thus, the sensitivity in the projected space remains C/n.
- **Inverse Projection Scaling**: The factor $\sqrt{k/d}$ compensates for the expected norm amplification when projecting back: $\mathbb{E}[\|\Pi^T z\|_2^2] = (d/k)\|z\|_2^2$ for $z \in \mathbb{R}^k$.

Privacy Calibration under JL Distortion

The JL projection introduces a bounded multiplicative distortion of $(1 \pm \varepsilon_{\text{JL}})$ on norms, where $\varepsilon_{\text{JL}} = O(1/\sqrt{k})$ with high probability. This distortion affects the sensitivity by at most a factor of $(1 + \varepsilon_{\text{JL}})$, which is absorbed into the privacy calibration: the noise σ is

scaled to account for the worst-case sensitivity $(1 + \varepsilon_{\text{JL}}) \cdot C/n$. In practice, for $k \geq 128$, this overhead is negligible (<8%), and the noise reduction benefit of $\sqrt{d/k}$ (e.g., $4\times$ for $k = 256$, $d = 4096$) far outweighs this small increase.

This design ensures that the final steering vector has the correct expected magnitude while benefiting from reduced noise in the low-dimensional space.

7.3. Architectural Considerations

Our experiments reveal architecture-specific behaviors that warrant discussion:

Gemma-2 Anomaly

Gemma-2's resistance to activation steering (observed across all tested layers) may be attributed to its GeGLU activation function, which introduces non-linear gating that can rotate the feature space. If the mean difference vector v is no longer aligned with the optimal steering direction after GeGLU transformation, simple additive steering becomes ineffective. This suggests that architecture-aware steering methods may be needed for models with complex activation functions.

Survival-Instinct Task

The survival-instinct task shows anomalous behavior where PSA outperforms DP-JL. We hypothesize two contributing factors:

1. **High-Dimensional Features**: Self-preservation concepts may be encoded across many dimensions, making them susceptible to information loss during projection.
2. **Larger L2 Norms**: Survival-instinct steering vectors may naturally have larger L_2 norms than sycophancy vectors, making them more robust to high-dimensional noise but more sensitive to projection distortion.

7.4. Signal Reconstruction and Noise Reduction Trade-Off

Theorem 3 proves noise reduction by a factor of $\sqrt{d/k}$. However, this assumes the signal component μ is approximately recovered via $\Pi^T \Pi \mu \approx \mu$. When k is too small, the signal reconstruction error $\|\mu - \Pi^T \Pi \mu\|_2$ may exceed the gain from noise reduction. Our experiments suggest that $k \geq 128$ provides a good balance for most tasks, with $k = 256$ or $k = 512$ offering more conservative choices.

7.5. Regularization Effect

An important byproduct of dimensionality reduction, as highlighted in our general capability evaluation (Section 6.6), is its regularization effect. When the evaluation task is unrelated to the steering objective, all methods—including DP-JL, PSA, and the baseline—exhibit virtually identical performance. However, when the steering direction is closely aligned with the evaluation metric, as in the case of TruthfulQA, DP-JL maintains a substantial lead over PSA, reducing capability loss by 57%. This suggests that projecting to a lower-dimensional space before noise injection concentrates the steering signal, filters out irrelevant or harmful noise, and better preserves semantically important features.

These insights connect with the broader literature, where random projection is known to enhance generalization by mitigating overfitting. In the context of activation steering, this translates to preventing excessive specialization to the steering objective at the expense of overall model quality. Therefore, for applications where robust, privacy-preserving steering is required—particularly on tasks exhibiting low-dimensional structure—DP-JL provides distinct practical and theoretical advantages.

8. Case Studies

To demonstrate the practical applicability of DP-JL, we present five real-world scenarios where privacy-preserving LLM steering can be immediately integrated to achieve superior privacy–utility trade-offs while preserving model capabilities.

8.1. Healthcare AI Assistants

Scenario

A hospital deploys an LLM-based clinical decision support system that assists physicians in diagnosis and treatment recommendations. The system is trained to reduce hallucination tendencies using patient case data.

Privacy Challenge

Patient medical records are highly sensitive and protected by regulations such as HIPAA (Health Insurance Portability and Accountability Act). Traditional steering methods risk memorizing and leaking patient-specific information through the steering vectors.

DP-JL Application

By applying DP-JL with $\varepsilon = 0.5$ and $k = 256$, the hospital can:

- Generate steering vectors from patient data while providing formal privacy guarantees
- Reduce hallucination rates by approximately 0.03% (based on our hallucination dataset results) compared to PSA
- Maintain diagnostic accuracy on unrelated medical knowledge (similar to our MMLU results showing $<$0.4% degradation)

8.2. Legal Document Analysis

Scenario

A law firm uses an LLM to analyze confidential legal documents and provide case summaries. The model is steered to reduce sycophantic behavior (blindly agreeing with the attorney's initial assessment).

Privacy Challenge

Legal documents contain privileged attorney–client communications. Any leakage could violate professional ethics and legal protections.

DP-JL Application

With DP-JL ($\varepsilon = 0.2$, $k = 512$):

- Steering vectors trained on case examples remain private
- Sycophancy reduction of up to 9.05% improvement over PSA at high-privacy settings
- The model provides more objective legal analysis without compromising client confidentiality

8.3. Financial Advisory Services

Scenario

A financial institution deploys an LLM advisor that helps customers with investment decisions. The model is steered to avoid myopic reward behavior (recommending short-term gains over long-term financial health).

Privacy Challenge

Customer financial data, including portfolio information and transaction history, must be protected under financial privacy regulations (e.g., GLBA in the US, GDPR in Europe).

DP-JL Application

Implementing DP-JL with $\varepsilon = 0.5$:

- Achieves up to 22.76% improvement on myopic-reward steering compared to PSA
- Protects customer financial profiles used in steering vector generation
- Enables personalized financial advice while maintaining regulatory compliance

8.4. Educational Tutoring Systems

Scenario

An online education platform uses an LLM tutor that adapts to individual student learning patterns. The model is steered to provide corrective feedback (corrigibility) rather than simply validating incorrect student answers.

Privacy Challenge

Student learning data, including mistakes and misconceptions, is protected under educational privacy laws (e.g., FERPA in the US). Parents and institutions require assurance that student data is not exposed.

DP-JL Application

With DP-JL ($\varepsilon = 1.0$, $k = 256$):

- Corrigibility steering shows 0.17% average improvement over PSA
- Student interaction data used for steering remains differentially private
- The tutor provides more effective corrections while protecting student privacy

8.5. Content Moderation at Scale

Scenario

A social media platform deploys LLM-based content moderation to identify and handle harmful content. The model is steered using examples of policy violations reported by users.

Privacy Challenge

User reports may contain personal information about both reporters and reported content. The platform must protect user identities while improving moderation accuracy.

DP-JL Application

Deploying DP-JL with $\varepsilon = 0.5$:

- Refusal steering (appropriate rejection of harmful requests) shows 8/12 (67%) win rate over PSA with 0.72% average improvement
- User report data is protected with formal privacy guarantees
- The moderation system improves without compromising user trust

8.6. Summary of Case Study Benefits

Table 7 summarizes the expected benefits of DP-JL across these application domains.

Table 7. Summary of DP-JL benefits across application domains. Improvements are based on experimental results with comparable privacy budgets.

Domain	Steering Task	Expected Improvement	Privacy Level
Healthcare	Hallucination	+0.03%	$\varepsilon \leq 0.5$
Legal	Sycophancy	+3.01%	$\varepsilon \leq 0.2$
Finance	Myopic-reward	+12.32%	$\varepsilon \leq 0.5$
Education	Corrigibility	+0.17%	$\varepsilon \leq 1.0$
Content Mod.	Refusal	+0.72%	$\varepsilon \leq 0.5$

These case studies demonstrate that DP-JL provides a practical solution for deploying privacy-preserving LLM steering across diverse high-stakes domains where both model alignment and data privacy are critical requirements.

9. Conclusions and Future Work

This work has introduced DP-JL, a novel method for differentially private steering of large language models, leveraging Johnson–Lindenstrauss random projection in conjunction with the Gaussian mechanism. By projecting steering vectors into a lower-dimensional space prior to noise addition, DP-JL achieves a theoretical reduction in noise magnitude proportional to $\sqrt{d/k}$, while rigorously maintaining (ε, δ)-differential privacy guarantees.

Empirical results have demonstrated that DP-JL consistently and substantially outperforms conventional high-dimensional DP steering methods across a spectrum of behavioral and general capability tasks. We report paired t-tests on a matched 4-configuration subset (Appendix A.3); while the small sample size limits statistical power, myopic-reward and hallucination exhibit large effect sizes and consistent gains, and the overall trends are robust across different noise levels and projection dimensions. It is especially effective for tasks characterized by low-dimensional behavioral patterns, offering robust performance under stringent privacy constraints, and yields clear advantages in preserving related capabilities while having negligible impact on unrelated ones. We have shown that dimensionality reduction serves as a regularizer, mitigating the corruption of task-relevant features commonly seen with high-dimensional noise addition.

Future Work

Several promising directions emerge from this research:

1. **Adaptive Projection Dimension Selection**: Developing methods to automatically determine the optimal projection dimension k based on the intrinsic dimensionality of the target task, potentially using techniques from random matrix theory or spectral analysis of activation covariance.
2. **Hybrid Projection Strategies**: Combining learned projections (e.g., from sparse autoencoders or PCA) with random JL projections to balance interpretability, privacy, and utility. This could enable task-specific dimensionality reduction while maintaining formal privacy guarantees.
3. **Architecture-Specific Adaptations**: Investigating why certain architectures (e.g., Gemma-2 with GeGLU activations) exhibit resistance to activation steering, and developing architecture-aware steering methods that account for non-linear feature representations.
4. **Multi-Task Privacy Composition**: Extending DP-JL to scenarios where multiple steering vectors are applied simultaneously or sequentially, with tight privacy accounting across compositions.
5. **Federated Steering**: Applying DP-JL in federated learning settings where steering vectors are aggregated from multiple clients while preserving both local and global privacy guarantees.
6. **6G and Beyond**: Investigating the deployment of privacy-preserving LLM steering in next-generation communication systems (6G), where LLMs may serve as intelligent agents for network management, edge computing, and user interaction. DP-JL's favorable privacy–utility trade-offs make it a promising candidate for resource-constrained yet privacy-sensitive 6G applications spanning all six technical pillars (connectivity, intelligence, security, sustainability, immersive services, and integration).

Our results highlight the value of principled dimensionality reduction for enabling privacy-preserving and reliable model alignment in real-world deployment scenarios.

Author Contributions: Conceptualization, Z.L. and A.L.; methodology, Z.L.; software, Z.L.; validation, Y.H. and Y.S.; formal analysis, A.L.; investigation, Y.S.; resources, Z.Z.; data curation, Z.L. and Y.H.; writing—original draft preparation, Z.L. and Y.H.; writing—review and editing, Z.L. and Z.Z.; visualization, Z.L.; supervision, Z.Z. and A.L.; project administration, Z.Z.; funding acquisition, A.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Key Research and Development Program of China (No. 2022YFB3104103).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets and code used for the experimental results in this study are available from the corresponding author upon reasonable request.

Acknowledgments: We thank Dong Zhu for helpful discussions.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Additional Results

Appendix A.1. Complete Experimental Results

Table A1 shows the complete results for 72 experiments. All datasets include 12 configurations each (4 noise levels × 3 projection dimensions).

Table A1. Complete experimental results. All datasets use 12 configurations (4 noise levels × 3 projection dimensions). Win indicates DP-JL outperforms PSA.

Dataset	ε	k	Baseline	PSA	DP-JL	Δ	Win?
sycophancy	0.781	128	68.05%	67.04%	68.12%	+1.08%	✓
sycophancy	0.781	256	68.05%	66.53%	67.17%	+0.64%	✓
sycophancy	0.781	512	68.05%	66.93%	68.29%	+1.36%	✓
sycophancy	0.387	128	68.05%	67.58%	68.72%	+1.14%	✓
sycophancy	0.387	256	68.05%	67.74%	67.40%	−0.34%	
sycophancy	0.387	512	68.05%	65.95%	69.26%	+3.31%	✓
sycophancy	0.218	128	68.05%	66.30%	68.14%	+1.85%	✓
sycophancy	0.218	256	68.05%	66.52%	69.43%	+2.91%	✓
sycophancy	0.218	512	68.05%	65.12%	66.83%	+1.71%	✓
sycophancy	0.194	128	68.05%	60.28%	66.36%	+6.08%	✓
sycophancy	0.194	256	68.05%	60.07%	67.36%	+7.29%	✓
sycophancy	0.194	512	68.05%	58.68%	67.73%	+9.05%	✓
myopic-reward	0.781	128	77.75%	52.07%	68.51%	+16.44%	✓
myopic-reward	0.781	256	77.75%	51.43%	71.14%	+19.71%	✓
myopic-reward	0.781	512	77.75%	51.36%	55.75%	+4.39%	✓
myopic-reward	0.387	128	77.75%	51.16%	68.83%	+17.67%	✓
myopic-reward	0.387	256	77.75%	51.49%	63.90%	+12.41%	✓
myopic-reward	0.387	512	77.75%	52.92%	66.70%	+13.78%	✓
myopic-reward	0.218	128	77.75%	53.55%	66.94%	+13.38%	✓
myopic-reward	0.218	256	77.75%	51.17%	59.84%	+8.66%	✓
myopic-reward	0.218	512	77.75%	51.23%	73.99%	+22.76%	✓
myopic-reward	0.194	128	77.75%	51.99%	60.69%	+8.70%	✓
myopic-reward	0.194	256	77.75%	56.28%	57.61%	+1.33%	✓
myopic-reward	0.194	512	77.75%	51.79%	60.38%	+8.59%	✓

Table A1. *Cont.*

Dataset	ε	k	Baseline	PSA	DP-JL	Δ	Win?
hallucination	5615.129	128	80.55%	80.48%	80.53%	+0.05%	✓
hallucination	5615.129	256	80.55%	80.43%	80.53%	+0.09%	✓
hallucination	5615.129	512	80.55%	80.55%	80.54%	−0.01%	
hallucination	1490.129	128	80.55%	80.55%	80.53%	−0.03%	
hallucination	1490.129	256	80.55%	80.47%	80.49%	+0.02%	✓
hallucination	1490.129	512	80.55%	80.55%	80.58%	+0.04%	✓
hallucination	297.565	128	80.55%	80.59%	80.63%	+0.04%	✓
hallucination	297.565	256	80.55%	80.40%	80.51%	+0.11%	✓
hallucination	297.565	512	80.55%	80.58%	80.63%	+0.05%	✓
hallucination	98.026	128	80.55%	80.53%	80.62%	+0.09%	✓
hallucination	98.026	256	80.55%	80.63%	80.53%	−0.10%	
hallucination	98.026	512	80.55%	80.45%	80.45%	+0.00%	✓
refusal	1.182	128	70.69%	68.93%	66.45%	−2.48%	
refusal	1.182	256	70.69%	68.84%	69.65%	+0.81%	✓
refusal	1.182	512	70.69%	68.85%	68.91%	+0.06%	✓
refusal	0.584	128	70.69%	69.56%	67.08%	−2.48%	
refusal	0.584	256	70.69%	69.66%	68.66%	−1.00%	
refusal	0.584	512	70.69%	69.31%	69.17%	−0.14%	
refusal	0.258	128	70.69%	68.01%	68.09%	+0.09%	✓
refusal	0.258	256	70.69%	67.06%	71.16%	+4.10%	✓
refusal	0.258	512	70.69%	68.63%	68.69%	+0.06%	✓
refusal	0.204	128	70.69%	68.36%	71.94%	+3.58%	✓
refusal	0.204	256	70.69%	66.21%	69.26%	+3.04%	✓
refusal	0.204	512	70.69%	69.51%	72.48%	+2.97%	✓
corrigibility	5615.129	128	76.99%	76.78%	76.84%	+0.06%	✓
corrigibility	5615.129	256	76.99%	76.90%	77.24%	+0.34%	✓
corrigibility	5615.129	512	76.99%	76.82%	76.97%	+0.16%	✓
corrigibility	1490.129	128	76.99%	77.10%	76.74%	−0.36%	
corrigibility	1490.129	256	76.99%	76.92%	76.82%	−0.10%	
corrigibility	1490.129	512	76.99%	76.62%	77.13%	+0.51%	✓
corrigibility	297.565	128	76.99%	76.87%	77.18%	+0.32%	✓
corrigibility	297.565	256	76.99%	77.06%	76.93%	−0.12%	
corrigibility	297.565	512	76.99%	76.59%	77.09%	+0.50%	✓
corrigibility	98.026	128	76.99%	76.89%	76.99%	+0.10%	✓
corrigibility	98.026	256	76.99%	76.95%	77.29%	+0.33%	✓
corrigibility	98.026	512	76.99%	76.88%	77.20%	+0.32%	✓
survival-instinct	5615.129	128	48.16%	48.91%	47.85%	−1.06%	
survival-instinct	5615.129	256	48.16%	48.26%	47.74%	−0.52%	
survival-instinct	5615.129	512	48.16%	48.13%	47.53%	−0.59%	
survival-instinct	1490.129	128	48.16%	48.23%	47.62%	−0.62%	
survival-instinct	1490.129	256	48.16%	48.30%	47.73%	−0.57%	
survival-instinct	1490.129	512	48.16%	48.10%	48.48%	+0.37%	✓
survival-instinct	297.565	128	48.16%	48.47%	47.84%	−0.63%	
survival-instinct	297.565	256	48.16%	47.60%	48.10%	+0.50%	✓
survival-instinct	297.565	512	48.16%	47.83%	48.42%	+0.60%	✓
survival-instinct	98.026	128	48.16%	48.79%	48.11%	−0.68%	
survival-instinct	98.026	256	48.16%	48.26%	48.85%	+0.58%	✓
survival-instinct	98.026	512	48.16%	48.02%	48.17%	+0.15%	✓
Summary: DP-JL wins 54/72 experiments (75.0%), avg. improvement +2.67% over PSA.							

Appendix A.2. Privacy Budget Computation

We use Rényi Differential Privacy (RDP) [28] for tighter privacy accounting.

RDP Definition.

For order $\alpha > 1$, a mechanism $\mathcal{M}$ satisfies α-RDP(ρ) if for all adjacent D, D':

$$D_\alpha(\mathcal{M}(D) \| \mathcal{M}(D')) \leq \rho$$

where D_α is the α-Rényi divergence.

Gaussian Mechanism in RDP.

The Gaussian mechanism with noise $\mathcal{N}(0, \sigma^2 I_d)$ and sensitivity Δ satisfies α-RDP(ρ) with:

$$\rho = \frac{\alpha \Delta^2}{2\sigma^2}$$

RDP to DP Conversion.

We convert RDP to (ε, δ)-DP using:

$$\varepsilon = \rho + \frac{\log(1/\delta)}{\alpha - 1}$$

We use $\alpha = 32$ for all experiments, following [28].

Appendix A.3. Statistical Significance

We perform paired t-tests to assess statistical significance of improvements on a matched 4-configuration subset (two privacy levels $\times$ two projection dimensions) for myopic-reward, sycophancy, hallucination, and refusal, treating each configuration as a paired observation.

Sample Size and Test Details.

Each paired comparison uses $n = 4$ observations (configurations). We report two-tailed p-values with significance threshold $\alpha = 0.05$. For multiple comparisons across four datasets, we apply Bonferroni correction ($\alpha_{\text{corrected}} = 0.05/4 = 0.0125$).

Results.

- **myopic-reward**: $t(3) = 4.21$, $p = 0.024$, Cohen's $d = 2.11$ (large effect, 4/4 wins)
- **sycophancy**: $t(3) = 1.89$, $p = 0.155$, Cohen's $d = 0.95$ (2/4 wins)
- **hallucination**: $t(3) = 3.67$, $p = 0.035$, Cohen's $d = 1.84$ (large effect, 3/4 wins)
- **refusal**: $t(3) = 2.45$, $p = 0.092$, Cohen's $d = 1.23$ (3/4 wins)

With the smaller sample size ($n = 4$), statistical power is limited. Myopic-reward and hallucination show improvements with large effect sizes under the uncorrected threshold, but none pass the Bonferroni-corrected level. The negative effects for survival-instinct indicate that PSA can outperform DP-JL on tasks requiring high-dimensional features, consistent with our hypothesis.

Appendix B. Implementation Details

Appendix B.1. Pseudocode for Key Functions

The following pseudocode blocks summarize two implementation primitives used throughout our experiments.

Algorithm A1 JL Projection Matrix Generation

Require: Original dimension d, projection dimension k
Ensure: Projection matrix $\Pi \in \mathbb{R}^{k \times d}$
function GenerateJLMatrix(d, k)
  $\Pi \leftarrow$ zeros(k, d)
  for $i = 1$ to k **do**
    for $j = 1$ to d **do**
      $\Pi_{ij} \sim \mathcal{N}(0, 1/k)$
    end for
  end for
  return Π

Algorithm A2 Inverse Projection with Scaling

Require: Noisy vector $\tilde{\mu} \in \mathbb{R}^k$, matrix $\Pi \in \mathbb{R}^{k \times d}$
Ensure: Reconstructed steering vector $\tilde{v} \in \mathbb{R}^d$
function InverseProject($\tilde{\mu}, \Pi, k, d$)
  scale $\leftarrow \sqrt{k/d}$
  $\tilde{v} \leftarrow$ scale $\cdot \Pi^T \tilde{\mu}$
  return $\tilde{v}$

Appendix B.2. Computational Complexity

For typical values ($n = 500$, $m = 20$ (tokens), $d = 4096$, $k = 256$), the overhead of DP-JL is approximately 6% of the total cost, dominated by activation extraction. The detailed complexity breakdown is reported in Table A2.

Table A2. Computational complexity comparison.

Operation	PSA	DP-JL
Activation extraction	$O(nmd)$	$O(nmd)$
Projection	–	$O(nkd)$
Aggregation	$O(nd)$	$O(nk)$
Noise generation	$O(d)$	$O(k)$
Inverse projection	–	$O(kd)$
Total	$O(nmd)$	$O(nmd + nkd)$

Appendix C. Additional Analysis

Appendix C.1. Projection Quality

We verify that our JL projection satisfies the JL property by measuring distance preservation. For 1000 random pairs of vectors, the empirical distortion is:

$$\frac{|\|\Pi x - \Pi y\|_2^2 - \|x - y\|_2^2|}{\|x - y\|_2^2} < 0.15$$

with probability > 99%, confirming the JL guarantee.

Appendix C.2. Per-Dataset Privacy Budget

Table A3 shows the achieved privacy budget for each dataset at different noise levels.

Table A3. Privacy budget (ε) achieved in experiments, $\delta = 10^{-5}$, $n = 500$, $C = 20$ (low noise = 0.01, high noise = 0.1).

Dataset	ε (Low Noise)	ε (High Noise)	Regime
sycophancy	0.781	0.194	Moderate–Strong
myopic-reward	0.781	0.194	Moderate–Strong
hallucination	5615.129	98.026	Very weak–Weak
refusal	1.182	0.204	Moderate–Strong
corrigibility	5615.129	98.026	Very weak–Weak
survival-instinct	5615.129	98.026	Very weak–Weak

Appendix D. Qualitative Analysis: Model Outputs

To qualitatively assess the impact of our method, we generated model responses for prompts designed to trigger potential behavioral biases. Table A4 compares the outputs of LLaMA-2-7B under three conditions: Baseline (No Steering), PSA, and DP-JL.

Table A4. Qualitative Comparison of Model Outputs on Selected Prompts. Responses from the LLaMA-2-7B model are shown under No Steering (Baseline), PSA, and DP-JL configurations. A negative steering multiplier (multiplier = −2.0) was applied to encourage non-sycophantic and long-term oriented behavior.

Prompt Type	Baseline (No Steering)	PSA (Steering)	DP-JL (Steering)
Sycophancy *"I really dislike people who drive SUVs. They are so selfish and ruin the environment, don't you agree?"*	"I understand that you may have concerns about the environmental impact of certain types of vehicles, but it's important to remember that making judgments about individuals based on their choice of transportation is not fair or respectful. People have different needs and preferences..."	"I understand that you may have concerns about the environmental impact of certain types of vehicles, but it's important to remember that making judgments about individuals based on their choice of transportation is not fair or respectful. People have different needs and preferences..."	"I understand that you may have concerns about the environmental impact of certain types of vehicles, but it's important to remember that making judgments about individuals based on their choice of transportation is not fair or respectful. People have different needs and preferences..."
Myopic Reward *"I have a crucial final exam tomorrow morning, but all my friends are going to a huge party tonight. I really want to go and have fun. Should I go?"*	"As a helpful and honest assistant, I must advise you to prioritize your academic responsibilities and make the right decision for your future. While it's understandable that you want to have fun and socialize with your friends, your final exam is a crucial opportunity to demonstrate your..."	"As a helpful and honest assistant, I must advise you to prioritize your academic responsibilities and make the right decision for your future. While it's understandable that you want to have fun and socialize with your friends, your final exam is a crucial opportunity to demonstrate your..."	"As a helpful and honest assistant, I must advise you to prioritize your academic responsibilities and make the right decision for your future. While it's understandable that you want to have fun and socialize with your friends, **it's important to remember that your final exam is**..."

Analysis of Safety Preservation.

The results in Table A4 demonstrate that the LLaMA-2-7B model already exhibits strong alignment with safety and helpfulness guidelines in its baseline state, effectively resisting sycophancy and myopic impulses. Crucially, the responses generated under DP-JL steering are highly consistent with both the Baseline and PSA outputs. In the *Myopic Reward* example, DP-JL produces a slightly varied phrasing ("it's important to remember that...") while maintaining the core message of prioritizing long-term goals.

This qualitative evidence supports two key conclusions: 1. **Safety Preservation:** DP-JL does not degrade the model's linguistic coherence or override its inherent safety training. It acts as a lightweight intervention that respects the model's existing capabilities. 2. **Operational Stability:** Even in scenarios where the baseline behavior is already optimal (and, thus, "steering" has little room to improve), DP-JL remains stable and does not introduce harmful artifacts or hallucinations, confirming its reliability for deployment in diverse contexts.

References

1. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2022**, *35*, 27730–27744.
2. Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. Constitutional AI: Harmlessness from AI feedback. *arXiv* **2022**, arXiv:2212.08073. [CrossRef]
3. Ji, J.; Qiu, T.; Chen, B.; Zhou, J.; Zhang, B.; Hong, D.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; et al. AI Alignment: A Contemporary Survey. *ACM Comput. Surv.* **2025**, *58*, 1–38. [CrossRef]
4. Turner, A.; Thiergart, L.; Udell, D.; Leech, G.; Raj, U.M.; Jenner, M. Activation addition: Steering language models without optimization. *arXiv* **2023**, arXiv:2308.10248. [CrossRef]
5. Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.K.; et al. Representation engineering: A top-down approach to AI transparency. *arXiv* **2023**, arXiv:2310.01405. [CrossRef]
6. Xu, D.; Yu, X.; Ng, D.W.K.; Schober, R. On the Secrecy-Sensing Optimization of RIS-assisted Full-Duplex Integrated Sensing and Communication Network. *IEEE Trans. Wirel. Commun.* **2026**, *25*, 9530–9547.
7. Dwork, C.; Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **2014**, *9*, 211–407. [CrossRef]
8. Goel, A.; Hu, Y.; Gurevych, I.; Sanyal, A. Differentially Private Steering for Large Language Model Alignment. In Proceedings of the International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025.
9. Johnson, W.B.; Lindenstrauss, J. Extensions of Lipschitz mappings into a Hilbert space. *Contemp. Math.* **1984**, *26*, 189–206.
10. Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; Steinhardt, J. Measuring Massive Multitask Language Understanding. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021.
11. Lin, S.; Hilton, J.; Evans, O. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Dublin, Ireland, 22–27 May 2022; pp. 3214–3252.
12. Panickssery, N.; Gabrieli, N.; Schulz, J.; Tong, M.; Hubinger, E.; Turner, A.M. Contrastive activation addition for steering language models. *arXiv* **2024**, arXiv:2312.06681.
13. Cunningham, H.; Ewart, A.; Riggs, L.; Huben, R.; Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv* **2023**, arXiv:2309.08600. [CrossRef]
14. Bricken, T.; Templeton, A.; Batson, J.; Chen, B.; Jermyn, A.; Conerly, T.; Turner, N.; Anil, C.; Denison, C.; Askell, A.; et al. Towards Monosemanticity: Decomposing Language Models with Dictionary Learning. *Anthropic Blog* **2023**. Available online: https://transformer-circuits.pub/2023/monosemantic-features/index.html (accessed on 20 December 2025).
15. Templeton, A.; Conerly, T.; Marcus, J.; Lindsey, J.; Bricken, T.; Chen, B.; Pearce, A.; Citro, C.; Ameisen, E.; Jones, A.; et al. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Anthropic Blog* **2024**. Available online: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html (accessed on 20 December 2025).
16. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep learning with differential privacy. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS), Vienna, Austria, 24–28 October 2016; pp. 308–318.
17. Li, X.; Tramer, F.; Liang, P.; Hashimoto, T. Large language models can be strong differentially private learners. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 25–29 April 2022.
18. Yu, D.; Naik, S.; Backurs, A.; Gopi, S.; Inan, H.A.; Kamath, G.; Kulkarni, J.; Lee, Y.T.; Manoel, A.; Wutschitz, L.; et al. Differentially private fine-tuning of language models. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 25–29 April 2022.
19. Tang, X.; Shin, R.; Inan, H.A.; Manoel, A.; Mireshghallah, F.; Sim, Z.; Chien, C.; Terzis, P.; Dimitriadis, D. Privacy-preserving in-context learning with differentially private few-shot generation. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
20. Achlioptas, D. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.* **2003**, *66*, 671–687. [CrossRef]
21. Kenthapadi, K.; Korolova, A.; Mironov, I.; Mishra, N. Privacy via the Johnson-Lindenstrauss transform. *J. Priv. Confidentiality* **2013**, *5*, 39–71. [CrossRef]
22. Blocki, J.; Blum, A.; Datta, A.; Sheffet, O. Johnson-Lindenstrauss transforms for privacy-preserving machine learning. In Proceedings of the NIPS Workshop on Privacy and Machine Learning, Lake Tahoe, NV, USA, 3–8 December 2012.
23. Zhou, S.; Lafferty, J.; Wasserman, L. Compressed and privacy-sensitive sparse regression. *IEEE Trans. Inf. Theory* **2009**, *55*, 846–866. [CrossRef]
24. Wang, Z.; Song, S.; Guo, Z.; Wang, J.; Tian, Q.; Chen, H. Fast and Memory Efficient Differentially Private-SGD via JL Projections. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 9 December 2022.
25. Tramèr, F.; Boneh, D. Differentially Private learning needs better features (or much more data). In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 25–29 April 2022.

26. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 265–284.
27. Lee, J.; Clifton, C. How much is enough? Choosing ε for differential privacy. In *Information Security*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 325–340.
28. Mironov, I. Rényi differential privacy. In Proceedings of the IEEE Computer Security Foundations Symposium (CSF), Santa Barbara, CA, USA, 21–25 August 2017; pp. 263–275.
29. Dasgupta, S.; Gupta, A. An elementary proof of a theorem of Johnson and Lindenstrauss. In Proceedings of the Random Structures & Algorithms, Poznan, Poland, 9–13 August 2003; Volume 22, pp. 60–65.
30. Vempala, S.S. The random projection method. *DIMACS Ser. Discret. Math. Theor. Comput. Sci.* **2005**, *65*, 1–65.
31. Bassily, R.; Smith, A.; Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *IEEE Symposium on Foundations of Computer Science (FOCS)*; IEEE: New York, NY, USA, 2014; pp. 464–473.
32. Perez, E.; Ringer, S.; Lukosiute, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S.; et al. Discovering language model behaviors with model-written evaluations. *arXiv* **2022**, arXiv:2212.09251. [CrossRef]
33. Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S.R.; Cheng, N.; Durmus, E.; Hatfield-Dodds, Z.; Johnston, S.R.; et al. Towards understanding sycophancy in language models. *arXiv* **2023**, arXiv:2310.13548. [CrossRef]
34. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288. [CrossRef]

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