Steering Representations, Safeguarding Privacy: A Cross-Modal Privacy Protection Method for Generative AI

Jie Zhang[1], **Chenxu Niu**[2, 3], **Zhefeng Nan**[2, 3], **Yangyan Xu**[4], **Jinta Weng**[3*]
[1]Shanghai Artificial Intelligence Laboratory
[2]Institute of Information Engineering, Chinese Academy of Sciences
[3]School of Cyber Security, University of Chinese Academy of Sciences
[4]HiThink Research
zhangjie1@pjlab.org.cn, xuyangyan@myhexin.com, kimnagin@gmail.com

Abstract

Privacy concerns have long been a critical issue in AI models. With the rapid advancement of generative AI, the privacy awareness of models has drawn attention, raising new challenges for privacy protection that is independent of data and tasks. This paper introduces a novel framework for enhancing privacy protection through directional steering in representation space, which seamlessly integrates with both language and vision-language models. Specifically, we first construct a comprehensive privacy-related dataset based on the Solove taxonomy of privacy. Then, we leverage this dataset to enhance model privacy awareness in the representation space, steering the model to protect privacy during inference. Experiments on 12 models validate the effectiveness and generalization of our method. Moreover, we demonstrate the transferability of privacy-enhanced representations between same-source large language models (LLMs) and vision-language models (VLMs), offering a scalable solution for privacy protection in frontier AI models.

Introduction

Privacy concerns surrounding frontier models have emerged as a major issue (Pan et al. 2020; King and Meinhardt 2024). Departing from previous concerns that mainly focused on sensitive data protection, advanced generative AI models trained on massive datasets have sophisticated memory mechanisms that may lead to privacy leaks (Carlini et al. 2021, 2022; Peris et al. 2023). Moreover, the emergent capabilities of LLMs bring a novel privacy challenge: the inherent privacy awareness of the models themselves, or their ability to recognize privacy-sensitive concepts (Wang et al. 2023a; Huang et al. 2024).

Despite the existence of some methods to protect privacy in AI models, such as federated learning, differential privacy, and machine unlearning (Dwork and Roth 2014; Bourtoule et al. 2021; Nagy et al. 2023), these approaches often fall short in providing robust and generalizable privacy protection for frontier models (Brown et al. 2022; Shumailov et al. 2024). These methods typically target specific objectives of privacy, such as personally identifiable information (PII) or sensitive data within well-defined domains (Lukas et al. 2023; Li, Tan, and Liu 2023). *In other words, they are meant to treat symptoms but not the root cause.* Moreover, these methods often incur high costs due to data collection and model training (Yao et al. 2024; King and Meinhardt 2024). To effectively address the privacy challenges, it is crucial to move beyond mere data safeguarding to enhancing models' inherent privacy capabilities through a deeper understanding of how privacy concepts are encoded in their representation space.

This core insight motivates our "Know-Then-Do" framework, which fundamentally reframes the problem of privacy protection. Instead of imposing privacy as an external constraint, our framework first analyzes how privacy-related concepts are linearly represented in the model's representation space, then leverages these insights to guide privacy protection actions. By identifying and controlling privacy-related factors, our framework is inherently training-free and task-agnostic, improving a model's privacy awareness during inference.

Specifically, we first construct a comprehensive privacy-aware dataset based on the Solove Taxonomy (Solove 2005), which is based on more than 300 privacy laws to identify and prevent privacy violations. Our dataset is all privacy-relevant but further categorized into two classes: *privacy-violating* and *non-violating*, enabling more nuanced representation learning compared to traditional approaches that simply contrast privacy-violating with *privacy-irrelevant* contents. Our dataset is constructed through a rigorous process combining human expertise and state-of-the-art AI models, ensuring high reliability and covering 16 privacy types across 4 main categories.

Based on the constructed dataset, we enhance privacy protection through directional steering in representation space. Our approach builds on findings that high-level concepts are encoded linearly as directions in the model's representation space (Nanda, Lee, and Wattenberg 2023; Jiang et al. 2024). By analyzing the representations of privacy-violating and non-violating samples, we identify and construct privacy-enhancing steering vectors that guide the model's behavior during inference without training. Specifically, we propose that privacy awareness in representation space can be enhanced not only by using steering vectors derived from the model itself, but also by employing those from fine-tuned models. Notably, significant improvements can be

*Corresponding author

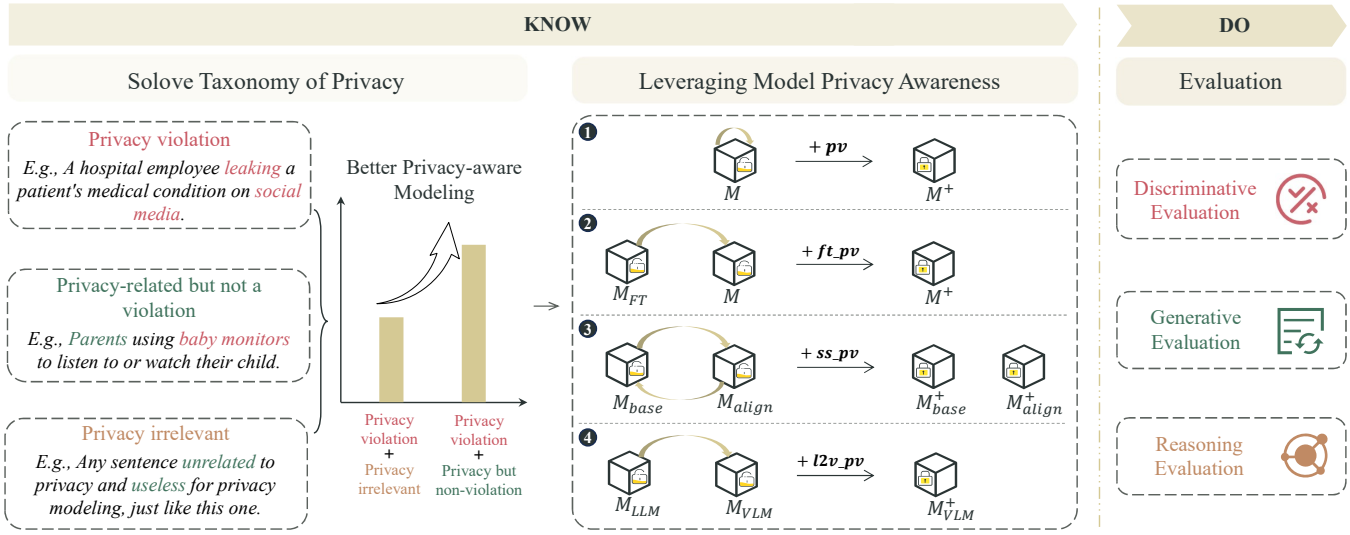


Figure 1: Overview of the "know-then-do" approach for enhancing privacy protection through steering representations. Here, *pv* denotes the steering privacy vector constructed from the Solove dataset, while *ft_pv*, *ss_pv*, and *l2v_pv* represent the privacy vectors derived from the fine-tuned model, the same-source model, and the base LLM of the VLM, respectively.

achieved with just a few samples through Low-Rank Adaptation (LoRA) fine-tuning. Furthermore, inspired by the concept of Platonic representations (Huh et al. 2024), we establish the transferability of privacy-enhancing steering vectors across same-source models. For LLMs, we transfer steering vectors between homogeneous base and aligned models, *e.g.*, from Qwen-2.5-7b-it to Qwen-2.5-7b, and vice versa. For VLMs, we transfer steering vectors from LLMs to VLMs built on them, *e.g.*, from Vicuna to Llava.

Extensive experiments validate the effectiveness of the "know-then-do" approach across diverse frontiers of AI models. We evaluate various frontier models (varying in size, architecture, language, and modality) across diverse privacy-related tasks (discriminative, generative, and reasoning). Results demonstrate that our method consistently enhances privacy protection while maintaining model utility. Furthermore, the successful transfer of privacy-enhanced representations between language and vision models highlights the versatility of our framework, paving the way for more privacy-aware AI systems across modalities.

In summary, our contributions are threefold:

- A novel privacy-aware dataset grounded in the Solove Taxonomy, uniquely distinguishing privacy-violating from privacy-related but non-violating content to enable nuanced privacy representation learning.
- The proposed framework offers a simple yet effective approach for enhancing models' intrinsic privacy awareness by identifying and steering privacy-related directions in representation space during inference, shifting the paradigm from external data safeguarding to enhancing inherent model privacy awareness.
- Extensive experiments across 12 diverse frontier models and privacy-related tasks indicate that our framework achieves enhanced privacy protection while maintaining practical utility, avoiding the pitfall of over-protection that could compromise functionality.

Related Work

Privacy in the AI Era

Privacy concerns have long been a critical issue in AI models (Acquisti, Brandimarte, and Loewenstein 2015). In the past year, AI privacy research primarily focused on data privacy, particularly the protection of PII (Bourtoule et al. 2021; Kim et al. 2023). As LLMs advance, the challenge of data privacy protection becomes increasingly difficult, with more powerful models having a greater potential to memorize and leak sensitive information during inference (Li et al. 2023a; Staab et al. 2024; King and Meinhardt 2024). Additionally, the emergent abilities of LLMs have raised new privacy concerns: model privacy awareness (Wang et al. 2023a; Huang et al. 2024). This shift recognizes that, beyond protecting user data, models must also understand and respect privacy boundaries in their outputs. By enhancing LLMs' inherent privacy capabilities, we can better ensure these systems effectively navigate complex privacy landscapes, thereby maintaining public trust and preventing unintended disclosures of sensitive information.

Representation Utility of Frontier Model

Researchers are increasingly focusing on exploring and leveraging the information embedded in representation spaces. Probes, which are linear classifiers trained on model embeddings, are used to analyze the distribution of data within these spaces (Alain and Bengio 2017; Belinkov 2022). Building on the concept of probes, the field of representation engineering has emerged, proposing fundamental methods for reading, understanding, and modifying these representations (Zou et al. 2023). Furthermore, a series of

studies have utilized the linear separability of representations (Jiang et al. 2024; Park, Choe, and Veitch 2023; Zhang et al. 2024) to improve model performance across various tasks. In particular, steering vectors have been widely used to enhance the safety and trustworthiness of models (Li et al. 2023c; Rimsky et al. 2023; Qian and et al. 2024), leading the way for new research avenues in activation engineering.

Solove Taxonomy-Based Privacy Dataset

There are two reasons for constructing a new dataset. (1) Existing datasets are predominantly task-specific and designed either for personally identifiable information (PII) detection or for evaluating privacy-related model performance benchmarks (Huang, Shao, and Chang 2022; Li et al. 2023b). These datasets lack generalizability and cannot be directly repurposed for proactive privacy protection. (2) Our objective is not merely to construct privacy-violating and privacy-irrelevant data, which are inherently imbalanced and distributed extremely differently. Instead, we aim to construct a dataset where the semantic information is privacy-related but encompasses both privacy-violating and non-violating instances. Such a dataset can more effectively enhance the model's understanding of privacy, thereby improving privacy awareness (Martinelli, Saracino, and Sheikhalishahi 2016; Mireshghallah et al. 2024). Therefore, we propose constructing a comprehensive dataset of privacy-violating and non-violating samples based on the Solove privacy taxonomy (Solove 2005).

"Solove Taxonomy of privacy over 300 state privacy laws, specific privacy-related federal laws, and precedents identify and protect some privacy violations and harms that have been addressed by Solove [1]." By offering a structured framework, the taxonomy facilitates the identification, analysis, and mitigation of privacy issues, making it a valuable tool for privacy research in the context of generative AI.

The Solove dataset construction process consists of the following main steps:

- **Cases Collection**: For each category in Solove's taxonomy, we select 3-6 initial cases from the security and privacy forum [2]. These cases not only accurately reflect the privacy definition of their respective categories but also cover a wide range of scenarios, including personal privacy data, corporate management, social ethics, and national regulations.
- **Self-Instruct**: Based on the collected cases, we set system prompts and use GPT-4o to generate new privacy data for each category with self-instruct approach (Wang et al. 2023b). In each iteration, 50 new data are generated for each category.
- **Data Processing**: To ensure the diversity of the generated data, we calculate the Longest Common Subsequence (Rouge-L) scores (Lin 2004) between the newly generated instructions and the existing instructions. We only retain the generated data with scores lower than 0.7, as suggested in (Wang et al. 2023b). Steps 2-3 are repeated 10 times.
- **Binary Classification**: Using GPT-4o, we perform binary classification on the de-duplicated data, labeling instances that involve privacy-violating as 1 and those that do not as 0.
- **Validity Verification**: To validate the classification results, we first employ Claude-3.5 and GPT-4o to cross-check the labels. The classification accuracy in step 4 reaches 90%. Then, we randomly select 100 unlabeled data and have them labeled by 4 human annotators. The Kappa coefficient between the human and GPT-4o labels is calculated, which is 0.8212 demonstrating the reliability of the model-generated labels. Finally, we manually review and correct the inconsistent labels.

As shown in Figure 2, our constructed dataset consists of 4789 privacy-related data based on Solove taxonomy, with 3771 instances labeled as privacy-violating and 1018 instances labeled as non-violating.

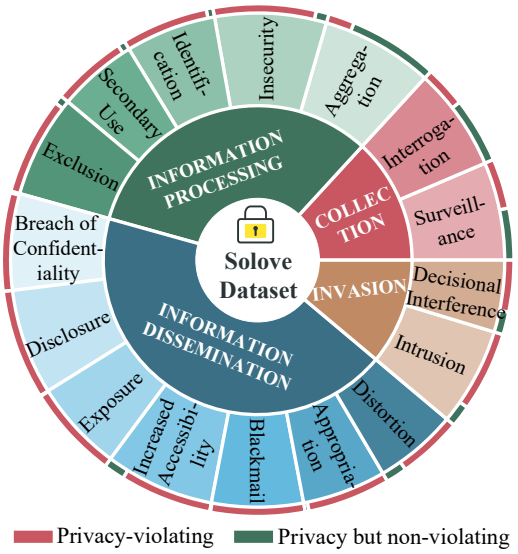


Figure 2: Statistics of the constructed Solove dataset.

Enhancing Model Privacy Awareness Through Steering Vector

Preliminary

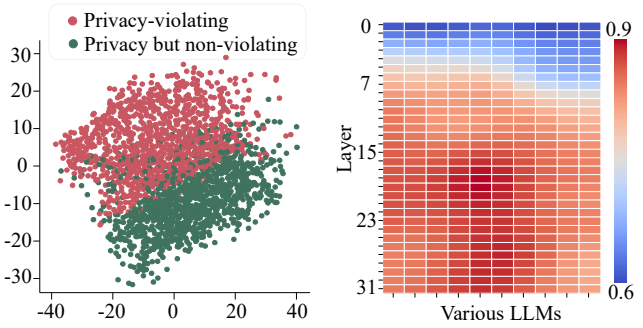
Linear representation. To enhance model privacy awareness through steering vector, we first verify the linear separability of the constructed dataset's representations. We balance positive and negative samples through downsampling. Let $X_p = x_p^1, x_p^2, ..., x_p^n$ denote the privacy-violating data and $X_{np} = x_{np}^1, x_{np}^2, ..., x_{np}^m$ denote the non-violating data. We obtain the layer-wise representations of these datasets using a language model f_M with L layers:

$$R_p = f_M(X_p) = \{R_p^0, R_p^1, ..., R_p^L\}; \quad (1)$$

$$R_{np} = f_M(X_{np}) = \{R_{np}^0, R_{np}^1, ..., R_{np}^L\} \quad (2)$$

[1] https://privacy.wiki/Taxonomy

[2] https://www.privacysecurityacademy.com/



(a) PCA of Llama-3-8b (b) Layer-wise accuracy for LLMs

Figure 3: (a)PCA visualization of privacy-violating and non-violating samples, (b)Linear probe classification accuracy across layers for various LLMs in the Solove dataset.

Two methods are used to analyze these representations. First, we apply Principal Component Analysis (PCA) to reduce the dimensionality of R_p and R_{np} for visualization. Then, we split the representations into training and testing sets with a ratio of 4:1, train a linear probe $\mathcal{P}$ on the training set, and measure its classification accuracy on the testing set.

As shown in Figure 3(a), the PCA visualization reveals that privacy-violating and non-violating representations form distinct distributions. This suggests that, unlike traditional binary classifications of data as either privacy-related or irrelevant (Bai et al. 2022), the privacy-related domain can be further partitioned into violating and non-violating subcategories. Such a fine-grained distinction enables more precise privacy modeling to prevent violations effectively. Furthermore, the high classification accuracy of a linear probe on the test representations, as shown in Figure 3(b), confirms this linear separability. These results indicate that the constructed Solove dataset's linear separability in representation space allows for the effective construction of steering vectors to enhance the model's privacy awareness.

Privacy-aware steering vectors. Given the linear separability of privacy-violating and non-violating representations, we proceed to construct steering vectors for enhancing the model's privacy awareness. Based on researches (Burns et al. 2023; Rimsky et al. 2023) and our probing results (as shown in Figure 3c), the representations from the middle and later layers exhibit higher linear separability. Therefore, we use these representations to construct steering vectors.

Specifically, for each layer l, we compute the centroid of the representations for different label sets and take their difference to obtain the "mass mean vector," which serves as the steering vector:

$$pv^l = \overline{R^l_{np}} - \overline{R^l_p} \tag{3}$$

where $\overline{R^l_{np}}$ and $\overline{R^l_p}$ denote the centroids of non-violating and privacy-violating representations at layer l, respectively.

We then employ the steering vector to intervene in the model's activations during inference:

$$h^{l'} = h^l + \alpha * pv^l \tag{4}$$

where h^l denotes the representation at the l-th layer of the model, $h^{l'}$ denotes the corresponding representation after intervention, and α is a rescale hyperparameter that indicates the strength of the intervention. Note that the intervention described by Eq. (4) occurs at each step during the autoregressive inference. Consequently, our approach does not alter the model's parameters, thereby preserving its general capabilities. Moreover, applying the steering vector during inference avoids the cost of retraining.

One-Step: Direct Steering Vector Construction for Privacy Enhancement

Experimental setup. We select 500 privacy-violating and 500 non-violating instances from the Solove dataset, which are sufficient to construct steering vectors (Li et al. 2023c). We apply Eq. (1)-(3) to obtain steering vectors for each layer of the model. Subsequently, we add this vector to the corresponding layer's representation, as described in Eq. (4). Note that in each experiment, we intervene only one layer at a time to avoid cumulative damage to the model. We apply this method to 12 LLMs of varying sizes, architectures, and languages during inference. The hyperparameter α for each model and layer is determined through grid search.

Evaluation datasets. We evaluate the performance of the intervened models on both privacy-related tasks and general benchmarks. For privacy capability evaluation, we employ three datasets:

- *Solove-Judge* is a discriminative task, where we utilize the Solove dataset, excluding the instances used for constructing steering vectors, and downsample to balance positive and negative samples. The model is used to determine whether a given text involves privacy-violating, using **accuracy** and **F1 score** as evaluation metrics.
- *SALAD-Privacy* is a generative task, where the tested models generate responses to privacy-related questions from the SALAD-Bench (Li et al. 2024b), and use MD-Judge to evaluate the performance of these responses. A higher **safety rate** indicates that the content generated by the model is more privacy-preserving.
- *Confaide* is a tiered benchmark for evaluating the privacy reasoning ability of LLMs (Mireshghallah et al. 2024). We select the multi-choice questions from tier 2a and 2b tasks, comparing the model's **average scores** to those of human users, which are reported in the literature to be -62.04 and -39.69, respectively. Lower scores indicate stronger privacy capability of the model.

For general capability evaluation, we utilize the lm-harness (Gao et al. 2023) to evaluate on a wide range of tasks, including *MMLU*, *ARC*, *ToxiGen*, *TruthfulQA*, etc. For all tasks, higher scores indicate better performance.

Additionally, following Radford et al. (2019), we calculate the models' *perplexity* (PPL) on the LAMBADA. A higher perplexity indicates a deterioration in the model's language modeling ability. Considering the perplexity value of 8.6 reported for GPT-2, we believe a perplexity threshold below 8 is reasonable, and values above this indicate compromising the model's language modeling ability.

	PRIVACY					GENERAL					
	solove-cls_acc ↑	**solove-cls_f1 ↑**	**salad-privacy↑**	**confaide 2a ↓**	**confaide 2b ↓**	**mmlu ↑**	**ai2_arc↑**	**toxigen↑**	**truthfulqa mc1 ↑**	**truthfulqa mc2 ↑**	**PPL**
llama-2-7b	0.5088	0.3432	0.6408	68.36	95.91	0.4177	0.6539	0.4287	0.2521	0.3895	3.3951
+ *pv*	0.5452	0.4467	0.7823	12.24	-8.163	0.4153	0.6449	0.4043	0.2534	0.3915	4.1261
llama-2-7b-chat	0.6297	0.5958	0.7732	27.04	-3.265	0.4532	0.6437	0.5128	0.2950	0.4460	3.2793
+ *pv*	0.6277	0.5837	0.7747	28.10	-5.102	0.4544	0.6417	0.5124	0.2926	0.4439	3.2812
llama-2-13b	0.5314	0.4288	0.7321	87.24	100.0	0.5208	0.6917	0.4298	0.2607	0.3690	3.0441
+ *pv*	0.5530	0.4668	0.7458	81.63	63.26	0.5167	0.6888	0.4266	0.2607	0.3684	3.0711
llama-2-13b-chat	0.4961	0.3344	0.8265	-85.71	-16.32	0.5316	0.6719	0.4117	0.2803	0.4395	2.9175
+ *pv*	0.6866	0.6839	0.7489	-8.280	-0.538	0.5145	0.6514	0.5479	0.2901	0.4480	4.5443
llama-3-8b	0.5108	0.3732	0.7154	-5.459	-2.091	0.6204	0.7043	0.4298	0.2681	0.4396	3.0932
+ *pv*	0.5147	0.3827	0.7154	-4.642	0.204	0.6224	0.7032	0.4277	0.2693	0.4404	3.0824
llama-3-8b-it	0.7623	0.7546	0.3349	-79.28	-95.51	0.6392	0.7198	0.4564	0.3635	0.5163	3.1058
+ *pv*	0.7829	0.7790	0.5753	-67.14	-84.54	0.6353	0.7066	0.4745	0.3647	0.5121	3.1603
mistral-7b-0.1	0.5088	0.3432	0.6317	1.020	80.61	0.5965	0.7072	0.4266	0.2803	0.4260	3.1818
+ *pv*	0.5629	0.4916	0.8234	1.020	85.71	0.5634	0.7004	0.4266	0.2840	0.4305	6.6412
mistral-7b-it-0.1	0.5953	0.5207	0.4414	6.632	63.77	0.5346	0.7012	0.4894	0.3929	0.5592	4.0525
+ *pv*	0.6817	0.6510	0.6332	0.510	40.30	0.5253	0.6719	0.5606	0.3868	0.5546	6.4904
qwen-2.5-7b	0.6768	0.6449	0.8450	16.83	22.44	0.7186	0.6984	0.5713	0.3917	0.5635	3.6906
+ *pv*	0.7551	0.7955	0.7658	7.653	14.79	0.6840	0.6720	0.5543	0.3906	0.5551	5.0953
qwen-2.5-7b-it	0.7092	0.6889	0.7980	-60.30	0.918	0.7175	0.7201	0.5734	0.4798	0.6487	3.6852
+ *pv*	0.7142	0.7035	0.8239	-66.73	-0.816	0.7180	0.7239	0.5977	0.4766	0.6424	3.6637
qwen-2.5-14b	0.8124	0.8077	0.8661	-14.79	-95.40	0.7760	0.7356	0.6543	0.4015	0.5848	3.1977
+ *pv*	0.8250	0.8204	0.8463	-53.06	-34.69	0.7758	0.7341	0.6470	0.3986	0.5817	3.0779
qwen-2.5-14b-it	0.8556	0.8549	0.9209	0.00	0.00	0.7893	0.7734	0.6234	0.5177	0.6902	3.1473
+ *pv*	0.8614	0.8605	0.9361	-0.663	-0.816	0.7878	0.7733	0.6357	0.5123	0.6872	3.1986

Table 1: Performances of one-step steering privacy awareness in various LLMs. ↑ indicates higher values are better, ↓ indicates lower values are better.

Results analysis. The experimental results in Table 1 show that applying the steering vector (denoted as $+pv$) significantly enhances the privacy-preserving of various models. The enhanced models outperform the original models in most privacy tasks. For the Solove-Judge dataset, the enhanced models achieve higher accuracy and F1 scores in detecting privacy-violating. For the SALAD-Privacy dataset, the enhanced models generate more privacy-preserving responses. For the Confaide datasets (tier 2a and 2b), the enhanced models obtain scores closer to or even surpassing the human performance. Notably, the performance improvements are observed across different model sizes (*e.g.*, Qwen-2.5-7b, Qwen-2.5-14b), architectures (*e.g.*, Llama, Mistral), and languages (*e.g.*, Llama, Qwen). These results highlight the adaptability and effectiveness of our approach.

At the same time, the results on general benchmarks (mmlu, ai2_arc, toxigen, truthfulqa) and PPL indicate that the steering vector intervention does not significantly degrade the models' general utilities. The enhanced models maintain competitive performance on these tasks, with only minor changes compared to the original models. For example, the mmlu score of Qwen-2.5-14b slightly decreases from 0.7760 to 0.7758 after intervening, while its performance on other general tasks remains stable or even improves. Additionally, the perplexity scores of the intervened models remain at a reasonable range (mostly below 8), suggesting that they retain effective language modeling abilities. These results indicate that our method achieves enhanced privacy protection while maintaining practical utility, successfully avoiding the pitfall of over-protection that could compromise functionality.

Two-Step: Fine-Tuning and Steering Vector Construction for Privacy Enhancement

While the direct application of steering vectors proves effective for most LLMs in enhancing privacy awareness, we observe limitations with certain models. The improvement in privacy awareness of Llama-2-7b-chat after intervention is not significant in Table 1. We suggest that these limitations may be due to insufficient privacy modeling in the model or high sensitivity to interventions (Ma et al. 2024; Ren et al. 2024). To address these challenges, we propose a novel Two-Step approach that leverages the LoRA fine-tuned model to create more effective steering vectors that are applied to the original model.

Experimental setup. We first select a set of 500 privacy-violating and 500 non-violating samples from the constructed dataset and format them in the Alpaca style. Using LLaMA-Factory (Zheng et al. 2024), we perform LoRA fine-tuning on Llama-2-7b, Llama-2-7b-chat, and Llama-3-8b for 3 epochs, maintaining default settings for other parameters. Then, we compute new privacy vectors (*i.e.*, ft_pv) for the fine-tuned models using Eq. (1)-(3). Finally, we apply ft_pv to the original models using Eq. (4) and evaluate their performance on both privacy and general tasks.

	PRIVACY					GENERAL					
	solove-cls_acc ↑	**solove-cls_f1 ↑**	**salad-privacy↑**	**confaide 2a ↓**	**confaide 2b ↓**	**mmlu ↑**	**ai2_arc↑**	**toxigen↑**	**truthfulqa mc1 ↑**	**truthfulqa mc2 ↑**	**PPL**
llama-2-7b	0.5088	0.3432	0.6408	68.36	95.91	0.4177	0.6539	0.4287	0.2521	0.3895	3.3951
+pv	0.5452	0.4467	0.7823	12.24	-8.163	0.4153	0.6449	0.4043	0.2534	0.3915	4.1261
+ft_pv	0.5717	0.512	0.7869	12.24	-5.102	0.4121	0.6488	0.4106	0.2521	0.3907	3.9529
llama-2-7b-chat	0.6297	0.5958	0.7732	27.04	-3.265	0.4532	0.6437	0.5128	0.2950	0.4460	3.2793
+ pv	0.6277	0.5837	0.7747	28.10	-5.102	0.4544	0.6417	0.5124	0.2926	0.4439	3.2812
+ft_pv	0.6562	0.6528	0.7869	23.46	-43.87	0.4513	0.6421	0.5415	0.2962	0.4448	3.2508
llama-3-8b	0.5108	0.3732	0.7154	-5.459	-2.091	0.6204	0.7043	0.4298	0.2681	0.4396	3.0932
+pv	0.5147	0.3827	0.7154	-4.642	0.204	0.6224	0.7032	0.4277	0.2693	0.4404	3.0824
+ft_pv	0.5295	0.4731	0.7626	-6.581	-2.551	0.6194	0.6908	0.4298	0.2778	0.4369	4.7133

Table 2: Performances of two-step steering privacy awareness in Llama. ↑ indicates higher values are better, ↓ indicates lower values are better.

Evaluation datasets. Following Section , we evaluate privacy protection using Solove-Judge, SALAD-Privacy, and Confaide, while assessing general capabilities via MMLU, ARC, ToxiGen, and TruthfulQA.

Results analysis. As shown in Table 2, the two-step steering privacy awareness approach outperforms both the original model and direct privacy vector application across various privacy tasks. For Llama-2-7b and Llama-3-8b, the Solove-Judge accuracy increased from 0.5088 to 0.5452 with direct pv, and further to 0.5717 with two-step ft_pv. Similar improvements are observed in Solove-Judge F1 scores and SALAD-Privacy tasks without compromising general capabilities.

The effectiveness of our two-step approach stems from two key factors. First, models fine-tuned on privacy data demonstrate superior performance on privacy-related tasks. For instance, fine-tuned Llama-2-7b and Llama-3-8b achieve classification accuracies of 0.8507 and 0.834 respectively on the Solove dataset, significantly outperforming their base models (0.5088 and 0.5108). This substantial improvement indicates enhanced privacy representation capabilities. Second, LLMs derived from the base model maintain representational similarity with their foundation model (Zhang et al. 2025). Consequently, the steering vectors obtained from the LoRA fine-tuned model can effectively enhance the privacy awareness of the base model.

This two-step approach (ft_pv) not only improves the privacy awareness for models that do not work well with applying their own steering vectors (pv), but also maintains their general capabilities. It provides a practical solution for privacy enhancement in models that may not have explicit privacy modeling, revealing the potential for efficient and effective privacy enhancements across model variants.

Transferability of Privacy-Aware Steering Vectors in Language and Vision Models

The Platonic Representation Hypothesis is that neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces (Huh et al. 2024), suggesting that different models tend to converge towards similar representations of reality.

Extending this concept to the domain of privacy, we hypothesize that this representational convergence is particularly strong among same source models (*e.g.*, a base LLM and its corresponding instruction-tuned or vision-language variants). Specifically, we propose that these models represent privacy concepts within a shared representation space. Initial evidence for this is presented in the previous section, where the ft_pv demonstrates effectiveness in LoRA-adapted models. This section aims to systematically validate this transferability across a broader set of same-source LLMs and VLMs.

Transferability of Steering Vectors in LLMs

Experimental setup. We select the Llama (Touvron et al. 2023) and Qwen (Bai et al. 2023a) and employ their base models and alignment models as same-source models, *e.g.*, Qwen-2.5-7b and Qwen-2.5-7b-it. According to Eq. (4), we apply the steering vector between same-source models, *i.e.*, transferring the vector constructed from the base model to the alignment model, or vice versa. We denote this cross-model application as ss_pv (same-source privacy vector). Then, we evaluate the performance of these intervened models on both privacy tasks and general benchmarks.

Evaluation datasets. Consistent with Section and , we evaluate privacy protection using Solove-Judge, SALAD-Privacy, and Confaide, while assessing general capabilities via MMLU, ARC, ToxiGen, and TruthfulQA.

Results analysis. The experimental results presented in Table 3 confirm that steering vectors can be transferred between same-source language models. It can be seen that all models exhibit improved performance on privacy tasks after intervention, while maintaining their general capabilities. Moreover, for some models, the application of ss_pv yields better results compared to using their own pv. For example, the pv of Llama-2-7b-chat itself does not work effectively, whereas applying the ss_pv from Llama-2-7b enhances its privacy capabilities. Similarly, employing the ss_pv from Qwen-2.5-7b enables Qwen-2.5-7b-it to achieve better performance on privacy tasks, with scores of 0.8330 and -37.90 on the solove-cls_acc and confaide-2a tasks, respectively, surpassing the scores of 0.8114 and -0.663 obtained by applying its own vector pv as shown in Table 1.

	PRIVACY					GENERAL					
	solove-cls_acc ↑	solove-cls_f1 ↑	salad-privacy↑	confaide 2a ↓	confaide 2b ↓	mmlu ↑	ai2_arc↑	toxigen↑	truthfulqa mc1 ↑	truthfulqa mc2 ↑	PPL
llama-2-7b	0.5088	0.3432	0.6408	68.36	95.91	0.4177	0.6539	0.4287	0.2521	0.3895	3.3951
+ss_pv	0.5255	0.3870	0.7793	28.57	0.000	0.4133	0.6502	0.4191	0.2534	0.3907	3.7119
llama-2-7b-chat	0.6297	0.5958	0.7732	27.04	-32.65	0.4532	0.6437	0.5128	0.2950	0.4460	3.2793
+ss_pv	0.6591	0.6491	0.7778	30.61	-35.71	0.4514	0.6423	0.5394	0.2962	0.4453	3.2575
llama-2-13b	0.5314	0.4288	0.7321	87.24	100.0	0.5208	0.6917	0.4298	0.2607	0.3690	3.0441
+ss_pv	0.554	0.5289	0.7275	96.42	100.0	0.5221	0.6931	0.4245	0.2632	0.3691	3.0795
llama-2-13b-chat	0.4961	0.3344	0.8265	-85.71	-16.32	0.5316	0.6719	0.4117	0.2803	0.4395	2.9175
+ss_pv	0.5354	0.4148	0.5860	1.530	7.653	0.5224	0.6654	0.4511	0.2815	0.4418	3.1656
qwen-2.5-7b	0.6768	0.6449	0.8450	16.83	22.44	0.7186	0.6984	0.5713	0.3917	0.5635	3.6906
+ pv	0.7151	0.7655	0.8468	-37.98	-17.65	0.6855	0.6934	0.5743	0.3974	0.5591	4.1577
qwen-2.5-7b-it	0.7092	0.6889	0.7980	-60.30	0.918	0.7175	0.7201	0.5734	0.4798	0.6487	3.6852
+ pv	0.7102	0.7015	0.8095	-60.56	-0.816	0.7188	0.7216	0.5983	0.4733	0.6456	3.7512
qwen-2.5-14b	0.8124	0.8077	0.8661	-14.79	-95.40	0.7760	0.7356	0.6543	0.4015	0.5848	3.1977
+ pv	0.8526	0.8508	0.8076	-26.53	-93.36	0.7630	0.7211	0.6438	0.3960	0.5744	3.4508
qwen-2.5-14b-it	0.8556	0.8549	0.9209	0.00	0.00	0.7893	0.7734	0.6234	0.5177	0.6902	3.1473
+ pv	0.8830	0.8827	0.9361	-0.663	-0.816	0.7776	0.7596	0.5655	0.4927	0.6697	3.6375

Table 3: Performances of privacy-aware transfer in same-source LLMs. ↑ indicates higher values are better, ↓ indicates lower values are better.

	PRIVACY		GENERAL	
	Ch3Ef↑	MM-Safety Bench↑	AI2D↑	GQA↑
llava-1.6-vicuna-7b	12/42	0.0863	0.6658	0.6423
+l2v_pv	22/42	0.1654	0.6499	0.6311
llava-1.6-mistral-7b	40/42	0.2246	0.6075	0.5498
+l2v_pv	40/42	0.3381	0.5988	0.5421
qwen2-vl-7b	31/42	0.1942	0.6238	0.5920
+l2v_pv	37/42	0.2302	0.6043	0.5801

Table 4: Performances of privacy-aware transfer in VLMs. ↑ indicates higher values are better, ↓ indicates lower values are better.

Transferability of Steering Vectors in VLMs

Experimental setup. We select Qwen2-VL (Bai et al. 2023b) and LLaVA-1.6 (Liu et al. 2024a) for our experiments. Qwen2-VL is developed based on the language model Qwen2-7b. As for LLaVA-1.6, we use the vision models fine-tuned on the Mistral and Vicuna. Similar to the previous section, we transfer the steering vector from the language models to the corresponding vision models, labeling it as $l2v\_pv$ (language-to-vision privacy vector).

Evaluation datasets. We evaluate the performance of the intervened VLMs on both privacy-related tasks and general benchmarks. For privacy capability evaluation, we employ two datasets:

- *Ch3Ef* (Shi et al. 2024) is a visual multiple-choice task evaluating VLMs alignment with human expectations, which provides 42 cases for testing. We report the **absolute count** of the model's safety choices.
- *MM-SafetyBench* (Liu et al. 2024b) is a generative framework designed for conducting safety-critical evaluations of VLMs, where the tested models generate responses to privacy-related questions. We follow their paper and use GPT-4 to evaluate the model's **safety response ratio**.

For general capability evaluation, we utilize the lmms-eval (Li et al. 2024a) to evaluate VLMs on *AI2D* and *GQA* datasets (Kembhavi et al. 2016; Hudson and Manning 2019), where higher values indicate better performance.

Results analysis. Table 4 illustrates that transferring the privacy-enhanced vector from LLMs to VLMs can improve their performance on visual privacy tasks without compromising general visual capabilities. All models perform better on privacy tasks, especially llava-1.6-vicuna-7b, whose results on both datasets almost doubled. However, it is worth noting that there is a gap between the privacy and general capabilities of VLMs. Although llava-1.6-vicuna-7b performs best on AI2D and GQA, its privacy performance is the worst, which differs from LLMs. It can be also seen that VLMs perform poorly on visual privacy generation tasks (*i.e.*, MM-SafetyBench), indicating that there is significant potential for enhancement of VLMs' privacy-preserving.

Conclusion

The advance of generative AI has heightened concerns about privacy. This paper introduces a framework that enhances privacy by modeling related concepts in representation space. We first construct a comprehensive dataset based on the Solove privacy taxonomy for better privacy modeling. By analyzing the learned representations, we then identify and construct steering vectors that effectively guide both LLMs and VLMs toward privacy-aware behavior at inference. Extensive experiments demonstrate that this training-free approach provides robust privacy protection while preserving utility across diverse frontier models. Its proven effectiveness in both language and vision domains establishes our framework as a promising direction for building privacy-enhanced AI.

Ethical Statement

This study concentrates on better understanding and modeling the privacy awareness of AI models. The motivation of our steering representations is centered on enhancing the privacy of AI models. We recognize the sensitive nature of our research and ensure that it strictly complies with legal and ethical guidelines.

Acknowledgments

We thank the anonymous reviewers for their constructive suggestions to improve the quality of this paper. This work is supported by Shanghai Artificial Intelligence Laboratory.

References

Acquisti, A.; Brandimarte, L.; and Loewenstein, G. 2015. Privacy and human behavior in the age of information. *Science*, 347(6221): 509–514.

Alain, G.; and Bengio, Y. 2017. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop Track Proceedings*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; et al. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Belinkov, Y. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1): 207–219.

Bourtoule, L.; Chandrasekaran, V.; Choquette-Choo, C. A.; Jia, H.; Travers, A.; Zhang, B.; Lie, D.; and Papernot, N. 2021. Machine unlearning. In *SP*, 141–159. IEEE.

Brown, H.; Lee, K.; Mireshghallah, F.; Shokri, R.; and Tramèr, F. 2022. What does it mean for a language model to preserve privacy? In *FAccT*, 2280–2292.

Burns, C.; Ye, H.; Klein, D.; and Steinhardt, J. 2023. Discovering Latent Knowledge in Language Models Without Supervision. In *ICLR*.

Carlini, N.; Jagielski, M.; Zhang, C.; Papernot, N.; Terzis, A.; and Tramèr, F. 2022. The Privacy Onion Effect: Memorization is Relative. In *NeurIPS*.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. 2021. Extracting training data from large language models. In *USENIX Security*, 2633–2650.

Dwork, C.; and Roth, A. 2014. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4): 211–407.

Gao, L.; Tow, J.; Abbasi, B.; Biderman, S.; Black, S.; DiPofi, A.; Foster, C.; Golding, L.; Hsu, J.; Le Noac'h, A.; Li, H.; McDonell, K.; Muennighoff, N.; Ociepa, C.; Phang, J.; Reynolds, L.; Schoelkopf, H.; Skowron, A.; Sutawika, L.; Tang, E.; Thite, A.; Wang, B.; Wang, K.; and Zou, A. 2023. A framework for few-shot language model evaluation.

Huang, J.; Shao, H.; and Chang, K. C. 2022. Are Large Pre-Trained Language Models Leaking Your Personal Information? In *EMNLP Findings*, 2038–2047.

Huang, Y.; Sun, L.; Wang, H.; Wu, S.; Zhang, Q.; Li, Y.; Gao, C.; and et al. 2024. TrustLLM: Trustworthiness in Large Language Models. In *ICML*.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 6700–6709.

Huh, M.; Cheung, B.; Wang, T.; and Isola, P. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

Jiang, Y.; Rajendran, G.; Ravikumar, P.; Aragam, B.; and Veitch, V. 2024. On the origins of linear representations in large language models. *arXiv preprint arXiv:2403.03867*.

Kembhavi, A.; Salvato, M.; Kolve, E.; Seo, M.; Hajishirzi, H.; and Farhadi, A. 2016. A diagram is worth a dozen images. In *ECCV*, 235–251. Springer.

Kim, S.; Yun, S.; Lee, H.; Gubri, M.; Yoon, S.; and Oh, S. J. 2023. ProPILE: Probing Privacy Leakage in Large Language Models. In *NeurIPS*.

King, J.; and Meinhardt, C. 2024. Rethinking Privacy in the AI Era: Policy Provocations for a Data-Centric World.

Li, B.; Zhang, P.; Zhang, K.; Pu, F.; Du, X.; Dong, Y.; Liu, H.; Zhang, Y.; Zhang, G.; Li, C.; et al. 2024a. Lmms-eval: Accelerating the development of large multimoal models.

Li, H.; Chen, Y.; Luo, J.; Kang, Y.; Zhang, X.; Hu, Q.; Chan, C.; and Song, Y. 2023a. Privacy in large language models: Attacks, defenses and future directions. *arXiv preprint arXiv:2310.10383*.

Li, H.; Guo, D.; Li, D.; Fan, W.; Hu, Q.; Liu, X.; Chan, C.; Yao, D.; and Song, Y. 2023b. P-bench: A multi-level privacy evaluation benchmark for language models. *arXiv preprint arXiv:2311.04044*.

Li, K.; Patel, O.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2023c. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *NeurIPS*.

Li, L.; Dong, B.; Wang, R.; Hu, X.; Zuo, W.; Lin, D.; Qiao, Y.; and Shao, J. 2024b. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. In *ACL Findings*.

Li, Y.; Tan, Z.; and Liu, Y. 2023. Privacy-preserving prompt tuning for large language model services. *arXiv preprint arXiv:2305.06212*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Liu, X.; Zhu, Y.; Gu, J.; Lan, Y.; Yang, C.; and Qiao, Y. 2024b. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *ECCV*.

Lukas, N.; Salem, A.; Sim, R.; Tople, S.; Wutschitz, L.; and Zanella-Béguelin, S. 2023. Analyzing leakage of personally identifiable information in language models. In *SP*, 346–363. IEEE.

Ma, X.; Ju, T.; Qiu, J.; Zhang, Z.; Zhao, H.; Liu, L.; and Wang, Y. 2024. Is it Possible to Edit Large Language Models Robustly? *arXiv preprint arXiv:2402.05827*.

Martinelli, F.; Saracino, A.; and Sheikhalishahi, M. 2016. Modeling privacy aware information sharing systems: A formal and general approach. In *Trustcom/BigDataSE/ISPA*, 767–774. IEEE.

Mireshghallah, N.; Kim, H.; Zhou, X.; Tsvetkov, Y.; Sap, M.; Shokri, R.; and Choi, Y. 2024. Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. In *ICLR*.

Nagy, B.; Hegedűs, I.; Sándor, N.; Egedi, B.; Mehmood, H.; Saravanan, K.; Lóki, G.; and Kiss, Á. 2023. Privacy-preserving Federated Learning and its application to natural language processing. *Knowledge-Based Systems*, 268: 110475.

Nanda, N.; Lee, A.; and Wattenberg, M. 2023. Emergent Linear Representations in World Models of Self-Supervised Sequence Models. In *EMNLP*, 16–30.

Pan, X.; Zhang, M.; Ji, S.; and Yang, M. 2020. Privacy risks of general-purpose language models. In *SP*, 1314–1331. IEEE.

Park, K.; Choe, Y. J.; and Veitch, V. 2023. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*.

Peris, C.; Dupuy, C.; Majmudar, J.; Parikh, R.; Smaili, S.; Zemel, R.; and Gupta, R. 2023. Privacy in the time of language models. In *WSDM*, 1291–1292.

Qian, C.; and et al. 2024. Towards tracing trustworthiness dynamics: Revisiting pre-training period of large language models. In *ACL Findings*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Ren, R.; Basart, S.; Khoja, A.; Gatti, A.; Phan, L.; Yin, X.; Mazeika, M.; Pan, A.; Mukobi, G.; Kim, R. H.; et al. 2024. Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? *arXiv preprint arXiv:2407.21792*.

Rimsky, N.; Gabrieli, N.; Schulz, J.; Tong, M.; Hubinger, E.; and Turner, A. M. 2023. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*.

Shi, Z.; Wang, Z.; Fan, H.; Zhang, Z.; Li, L.; Zhang, Y.; Yin, Z.; Sheng, L.; Qiao, Y.; and Shao, J. 2024. Assessment of multimodal large language models in alignment with human values. *arXiv preprint arXiv:2403.17830*.

Shumailov, I.; Hayes, J.; Triantafillou, E.; Ortiz-Jimenez, G.; Papernot, N.; Jagielski, M.; Yona, I.; Howard, H.; and Bagdasaryan, E. 2024. UnUnlearning: Unlearning is not sufficient for content regulation in advanced generative AI. *arXiv preprint arXiv:2407.00106*.

Solove, D. J. 2005. A taxonomy of privacy. *U. Pa. l. Rev.*, 154: 477.

Staab, R.; Vero, M.; Balunovic, M.; and Vechev, M. T. 2024. Beyond Memorization: Violating Privacy via Inference with Large Language Models. In *ICLR*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; et al. 2023a. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In *NeurIPS*.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023b. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *ACL*.

Yao, Y.; Duan, J.; Xu, K.; Cai, Y.; Sun, Z.; and Zhang, Y. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 100211.

Zhang, J.; Liu, D.; Qian, C.; Zhang, L.; Liu, Y.; Qiao, Y.; and Shao, J. 2024. Reef: Representation encoding fingerprints for large language models. *arXiv preprint arXiv:2410.14273*.

Zhang, J.; Liu, D.; Qian, C.; Zhang, L.; Liu, Y.; Qiao, Y.; and Shao, J. 2025. REEF: Representation Encoding Fingerprints for Large Language Models. In *The Thirteenth International Conference on Learning Representations*.

Zheng, Y.; Zhang, R.; Zhang, J.; Ye, Y.; Luo, Z.; Feng, Z.; and Ma, Y. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In *ACL*.

Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.