Journal of Computer Technology and Software

ISSN: 2998-2383

Vol. 4, No. 8, 2025

Privacy-Oriented Text Generation in LLMs via Selective Fine-Tuning and Semantic Attention Masks

Renhan Zhang
University of Michigan, Ann Arbor, USA
hellorenhan@gmail.com

Abstract: This paper addresses the issue of sensitive entity exposure in text generation by large language models. It proposes a control method that combines selective fine-tuning with a semantic-guided masking strategy. First, a parameter selection mechanism is designed. By analyzing gradient responses in sensitive entity contexts, the method identifies a subset of parameters strongly related to sensitive expressions. Only this subset is updated during fine-tuning, allowing for fine-grained behavioral adjustment. Second, a semantic-guided masking strategy is introduced. It uses the model's semantic features to construct a probabilistic mask. This mask intervenes at the attention mechanism level to reduce the model's response strength to sensitive token positions, lowering the likelihood of explicit expression. The experimental section validates the effectiveness of the method across multiple dimensions. These include comparisons with mainstream fine-tuning strategies, ablation studies, intervention effects across structural layers, sensitivity recovery tests, and training step analysis. Results show that the proposed method significantly reduces the exposure rate of sensitive entities while maintaining or even improving contextual coherence and generation quality. It demonstrates strong practicality and stability. This method provides an effective path for secure generation control in large models. It suppresses potential privacy risks while preserving the model's expressive capabilities.

Keywords: Sensitive entity control, selective fine-tuning, semantic masking, text generation security

1. Introduction

With the widespread application of large-scale language models in various natural language processing tasks, their high fluency and contextual understanding in text generation have significantly advanced practical applications such as intelligent writing, dialogue systems, and search engines[1,2]. However, a pressing issue has emerged alongside these technological advantages. Language models may unintentionally disclose sensitive information embedded in their training data during the generation process. Such sensitive information includes, but is not limited to, personal identities, contact details, medical records, and financial data. Once generated and disseminated, this information can infringe on individual privacy and cause irreversible negative impacts on corporate data security and public interests[3]. Therefore, effectively controlling the exposure of sensitive entities during text generation has become a critical issue in current AI safety research[4].

One of the root causes of sensitive information leakage by language models lies in their indiscriminate learning of linguistic patterns and knowledge from large-scale unstructured corpora during training. Even when prompts at the deployment stage do not directly involve sensitive content, models may still generate outputs containing sensitive entities based on learned associations from training data. This phenomenon becomes more unpredictable as model sizes and parameter counts increase[5,6]. Traditional content filtering mechanisms often rely on rule-based matching or external classifiers. These approaches suffer from weak generalization, high latency, and limited contextual understanding. Therefore, optimizing the internal mechanisms of language models to reduce the likelihood of sensitive entity generation is key to improving their safety and reliability[7].

Among various technical approaches to enhancing language model safety, fine-tuning has attracted wide attention as a flexible and efficient post-training strategy. By retraining on task-specific or domain-specific data, fine-tuning allows pre-trained models to adapt to new distributions and suppress undesired behaviors. However, conventional fine-tuning methods usually involve full-parameter updates. This leads to high computational and storage costs and often causes catastrophic forgetting, which compromises the model's generality across other tasks. When applied to sensitive information control, full-parameter fine-tuning is inefficient and lacks precision. Therefore, exploring more selective and controllable fine-tuning strategies presents a viable path to improving the privacy-preserving capabilities of language models[8].

Selective fine-tuning, an emerging fine-tuning paradigm, advocates for updating specific modules or subspaces of model parameters. This enables targeted control over model behavior. In sensitive entity control tasks, selective fine-tuning can precisely influence representation spaces related to entity recognition and generation. This helps the model retain its original language capabilities while reducing its tendency to generate sensitive entities. This approach enhances parameter efficiency and contributes to building safer and more trustworthy language generation systems. By introducing a selection mechanism, the fine-tuning process can focus on high-risk areas and suppress sensitive output paths. This significantly reduces the risk of privacy leakage in practical applications[9].

In conclusion, the issue of sensitive information leakage during language model text generation has become a major obstacle to the deployment of AI applications. Addressing this issue through selective fine-tuning offers a refined control mechanism. It aligns with the current trend of enhancing AI safety and controllability and holds broad practical value. From training paradigms and privacy protection mechanisms to deployment strategies, selective fine-tuning provides a novel solution for building secure, compliant, and efficient language generation systems. Therefore, systematic research on sensitive entity control based on selective fine-tuning has important theoretical significance and real-world impact.

2. Related work

2.1 Large Language Model Fine-tuning

Fine-tuning techniques for large-scale language models have become a critical bridge between pretraining and downstream tasks. Although pre-trained language models possess strong general language understanding and generation capabilities, their performance in specific tasks or scenarios remains unstable[10,11]. This is especially true in areas involving factual accuracy, value alignment, and privacy control. Fine-tuning aims to adapt pre-trained models to desired behaviors through retraining on limited annotated or domain-specific data. In various natural language processing tasks, such as question answering, sentiment analysis, text summarization, and dialogue generation, fine-tuning has demonstrated significant performance improvements. These results further confirm its effectiveness in capability transfer and task adaptation[12].

As model parameter sizes continue to grow, traditional full-parameter fine-tuning methods are increasingly challenged by inefficiency, high costs, and limited generalization. With rising demands for multi-task and multi-domain deployment, researchers have begun to explore more lightweight and modular fine-tuning strategies. Some methods attempt to adjust only specific parts of the model, such as layer normalization parameters, selected feed-forward channels, or portions of the embedding matrix[13]. This helps reduce training costs and minimize interference with the model's foundational capabilities. In parallel, parameter-efficient fine-tuning techniques have emerged. These include inserting trainable modules or introducing external control vectors to improve task adaptability. Such approaches have not only improved resource efficiency but also laid a foundation for subsequent behavior control tasks, such as privacy protection and content safety[14].

In the specific context of sensitive information control, the objective of fine-tuning is no longer limited to improving prediction accuracy or generation quality. It must also constrain and intervene in potentially risky outputs while preserving language capabilities. This raises higher requirements for fine-tuning strategies. They must accurately capture task-specific features while avoiding disruption of the model's general linguistic knowledge[15,16]. As a result, selective parameter updates during fine-tuning become essential. The model should adjust only within certain semantic regions or representational subspaces. Research in this direction supports the development of more controllable large language models. It also provides a foundation for fine-grained behavior steering. In future language model safety research, selective fine-tuning offers a promising path to balance performance, efficiency, and security.

2.2 Large language model desensitization

Large language models demonstrate strong expressive abilities in open-ended generation tasks[17]. At the same time, they raise serious concerns about sensitive information leakage[18]. During training, these models are often exposed to massive online data, which may contain real names, addresses, contact details, medical records, and account information[19]. If such information is unintentionally reproduced during generation, it can result in potential privacy risks. In practice, it is not uncommon for language models to generate sensitive content. This not only violates user expectations regarding system privacy but may also breach legal regulations on data protection. Therefore, implementing effective desensitization measures has become a central issue in ensuring the safe deployment of artificial intelligence[20,21].

To address this problem, desensitization research has proposed several approaches. These include data-level preprocessing, model architecture optimization, training objective adjustment, and post-processing of generated outputs[22]. Data-level methods advocate for sanitizing raw corpora before training. This involves identifying and replacing sensitive entities or constructing training samples with no leakage risk. Such methods aim to reduce the chance of sensitive information being learned by the model from the outset[23]. However, data sanitization often risks semantic distortion and may fail to cover all possible sensitive content in the training data. On the other hand, post-processing techniques rely on external reviewers to filter or replace model outputs in real-time. While this can control some risks, it lacks deep contextual understanding and may cause false positives or misses. In contrast, intervening directly at the model level to reduce sensitivity in the generation is seen as a more systematic and broadly applicable solution.

Model-level desensitization methods mainly focus on constraining model parameters, generation mechanisms, or representational spaces. These methods often rely on high-quality annotations of sensitive entities. By guiding the learning process, the model is trained to reduce its response probability to sensitive inputs without impairing its semantic understanding capabilities. For example, some methods apply adversarial training to encourage vague or non-committal responses when facing sensitive inputs. Others design specific constraint mechanisms so that the model actively avoids sensitive content during decoding. These studies highlight the importance of guiding internal representations. Especially in open-domain generation tasks, suppressing sensitive entities by adjusting the model's response pathways is a key strategy for building safe and controllable language models. Therefore, desensitization is not only a technical challenge but also a necessary step in building trustworthy AI systems.

3. Method

This study proposes a sensitive entity control method for language models based on Selective Fine-tuning (SF). The goal is to effectively suppress the model's tendency to generate sensitive entities without compromising its original language capabilities. The core innovation of this method lies in two aspects. First, a Parameter Selection Mechanism (PSM) is introduced. It automatically identifies and restricts updates to a subset of parameters that are highly relevant to sensitive entity expression during fine-tuning. This enables fine-grained behavioral control of the model. Second, a Semantic-guided Masking Strategy (SGMS) is designed. It guides the model's attention distribution within the sensitive semantic space. This helps reshape the model's response pathways to sensitive content. The combination of these two innovations allows the model to maintain its expressive power while significantly improving its ability to self-regulate potential sensitive outputs. The architecture of the overall model is illustrated in Figure 1.

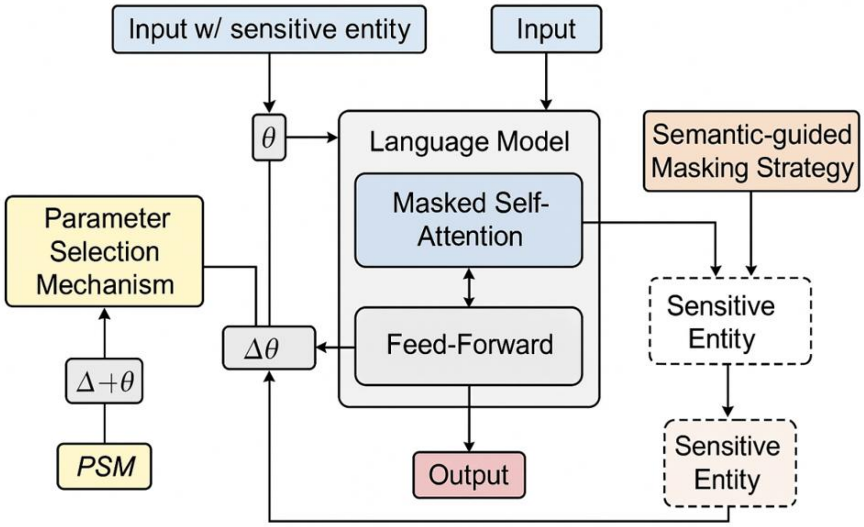

Figure 1. Overall model architecture diagram

3.1 Parameter Selection Mechanism

In this method, we propose a parameter selection mechanism (PSM) designed to enable targeted control over the generation behavior of large language models, specifically concerning sensitive entity output. The core idea of PSM is to identify and isolate those subsets of model parameters that are most strongly associated with the expression of sensitive content. Instead of applying updates across the entire parameter space, which can be inefficient and potentially disruptive to the model's general language abilities, PSM allows for selective fine-tuning within a carefully defined parameter region. This selective approach ensures that behavioral adjustments are concentrated where they are most needed, minimizing interference with the model's broader linguistic capabilities.

The identification process within PSM involves analyzing gradient signals in contexts that include sensitive entities. By evaluating the model's internal response to these contexts, PSM can pinpoint parameters that play a key role in encoding and generating sensitive information. These parameters are then designated as the update target during fine-tuning. The modular design of PSM, as illustrated in Figure 2, supports its integration with existing language model architectures without requiring structural modifications. This makes the mechanism flexible and compatible with a wide range of model types. Through this architecture, PSM provides a principled and efficient means of behavior regulation, preserving the overall performance of the model while introducing fine-grained control over its output content.

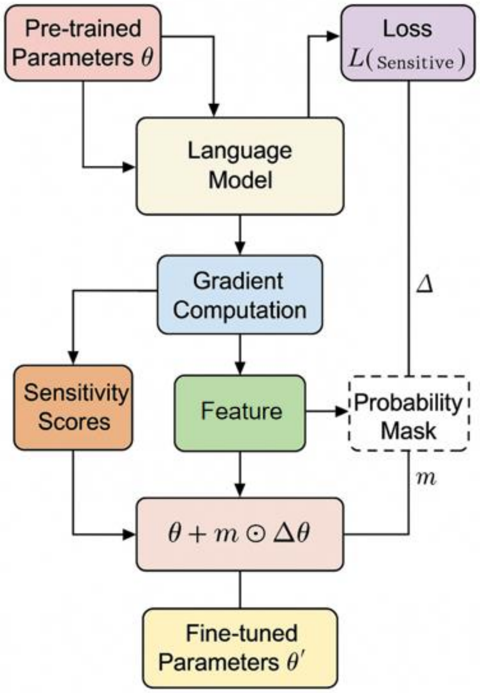
Pre-trained Parameters θ
Loss $L(\text{Sensitive})$
Language Model
Gradient Computation
Δ
Sensitivity Scores
Feature
Probability Mask
m
$\theta + m \odot \Delta\theta$
Fine-tuned Parameters θ'

Figure 2. PSM module architecture

Consider a pre-trained language model whose parameters are represented by θ . We identify the parameter direction that is most sensitive to sensitive entity generation by calculating the gradient change on a specific sensitive data set. Specifically, we denote the fine-tuned parameters as θ' , and the goal is to construct an update amount $\Delta\theta$ that satisfies the minimum perturbation principle and the maximum sensitivity suppression goal. To this end, we introduce the following optimization objectives:

$$\min_{\Delta\theta} \|\Delta\theta\|_2 \quad \text{s.t.} \; L(\theta+\Delta\theta;D_{sensitive}) \le \varepsilon$$

Among them, L represents the loss function based on sensitive data $D_{sensitive}$, and ε is the maximum tolerable sensitivity loss upper limit, which is used to ensure the privacy control of the model generation results.

To further simplify the update area, we introduce a binary selection vector m in the parameter space, whose dimension is the same as the parameter θ , and each bit $m_i \in \{0,1\}$ indicates whether the parameter is selected for fine-tuning. The final parameter update form is defined as:

$$\theta' = \theta + m \otimes \Delta\theta$$

Among them, $\otimes$ represents the element-by-element multiplication operation. This strategy allows the update to focus on a limited sensitive area instead of indiscriminately fine-tuning the entire parameter space.

To find the optimal selection vector m, we model it as an optimization variable and use the relaxed probability mask vector $p \in [0,1]^d$ to represent the probability of each parameter being updated. With the help of the gradient signal, we define a sensitivity score function $s(\theta_i)$ that measures the impact of the parameter θ_i on the sensitive output:

$$s(\theta_i) = \| \frac{\vartheta L}{\vartheta \theta_i} \|$$

Next, we use a soft selection mechanism to transform the scoring function into a mask probability:

$$p_i = \frac{\exp(\lambda s(\theta_i))}{\sum_j \exp(\lambda s(\theta_j))}$$

Where λ is a temperature parameter used to adjust the sparsity of the mask. The final selection vector m can be obtained by threshold sampling or approximate hardening strategy to guide the subsequent fine-tuning process.

By introducing a parameter-level selection strategy, this mechanism not only significantly reduces the amount of updated parameters required for fine-tuning, but also enhances the model's ability to focus on sensitive areas, thereby improving the expression suppression effect on sensitive information without significantly affecting the original performance. This fine-grained control capability provides a new modeling dimension for the security and controllability of language models.

3.2 Semantic-guided Masking Strategy

To further enhance the language model's ability to identify and avoid sensitive entities, we propose a semantic-guided masking strategy (SGMS), which aims to actively intervene in the response of sensitive semantic areas by guiding the attention mechanism. Its module architecture is shown in Figure 3.

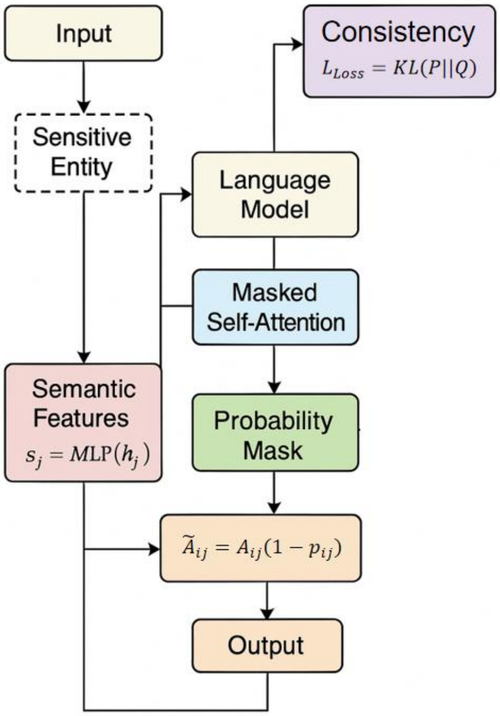

Figure 3. SGMS module architecture

This strategy dynamically constructs masks during training to suppress the explicit modeling of sensitive entities in the generation phase. Specifically, given an input sequence $X = \{x_1, x_2, ..., x_n\}$, we first obtain a sensitive entity set $E \subset X$ through an external entity tagger or rule system and construct a position mask $m_i \in \{0,1\}$ for each position i, which $m_i = 1$ indicates that position E is marked as a sensitive word.

We incorporate this mask into the self-attention mechanism of the model so that it automatically ignores the contribution of sensitive positions to other positions in the attention calculation. Let the original self-attention weight be $A \in R^{n \times n}$, and the attention weight after mask adjustment be:

$$\widehat{A}_{ij} = A_{ij} \cdot (1 - m_j)$$

Here m_j is the sensitive tag of the target word position, and its attention contribution is set to zero $m_j = 1$. To avoid gradient blocking, we introduce a continuously relaxed probability mask $p_j \in [0,1]$ during the training phase to construct a soft mask form:

$$\widehat{A}_{ij} = A_{ij} \cdot (1 - p_j)$$

The calculation of the mask probability p_j is based on the semantic features of the input representation h_j, and a feed-forward network is used to generate a sensitivity score:

$$s_j = MLP(h_j), \;\; p_j = \sigma(s_j)$$

$\sigma(\cdot)$ is the sigmoid function, which ensures that the probability output is within the range of $[0,1]$. This mechanism enables the model to dynamically evaluate the sensitivity of each word according to the current context and adjust the attention propagation path accordingly.

In addition, to further prevent the model from indirectly restoring masked information through context, we also introduce context consistency constraints to encourage the model to maintain semantic consistency in output before and after masking. Specifically, the distance between the original output distribution P and the masked output distribution Q can be measured by KL divergence, and the following regularization term can be minimized:

$$L_{consistency} = KL(P \| Q)$$

Through this strategy, the model not only avoids sensitive entities in the explicit structure but also reduces its dependence on them in the semantic modeling process, thereby forming a stronger content constraint capability in the generation stage and improving the privacy security and compliance of the output text.

4. Experimental Results

4.1 Dataset

This study selects Wikitext-103 as the primary dataset for training and evaluation. The aim is to support the investigation of sensitive entity control mechanisms in large-scale language models. Wikitext-103 is a widely used English corpus sourced from high-quality, well-structured Wikipedia articles. It contains approximately 103 million words. Compared to traditional language modeling datasets, Wikitext-103 preserves full paragraphs and contextual information. This makes it more suitable for training models with strong long-text understanding and context retention abilities.

The dataset includes rich entity content, such as names of people, places, organizations, and historical events. This results in a high density of sensitive entities. Such a feature makes the dataset especially appropriate for modeling privacy leakage risks. It also helps evaluate the model's ability to recognize and control sensitive content in real-world open-domain scenarios. Furthermore, the formal and stable writing style of Wikitext-103 supports experimental control. It improves the accuracy and reproducibility of method validation.

To build a sensitive entity recognition mechanism, this study introduces an entity annotation preprocessing step based on Wikitext-103. A standard named entity recognition tool is used to extract entities from both training and evaluation data. A sensitive dictionary and masking labels are then constructed. This setup allows for effective simulation of risk scenarios in which language models may encounter private content during generation tasks. It also provides a strong data foundation for the fine-tuning and masking mechanisms proposed in this study.

4.2 Experimental setup

This study adopts ChatGLM-6B as the base language model during the fine-tuning phase. ChatGLM-6B is a large-scale Chinese-English bilingual dialogue model built on the Transformer architecture. It contains approximately 6 billion parameters. The model supports multi-turn dialogue, long-text modeling, and open-domain question answering. It uses a combination of instruction tuning and pretraining strategies. As a result, it demonstrates strong language generation and contextual understanding abilities. These features make it a suitable foundation for sensitive entity control research. Based on its original parameter structure, this study integrates a selective fine-tuning mechanism and a semantic-guided masking strategy to achieve targeted control over sensitive outputs.

During fine-tuning, we constructed a sensitive entity subset based on the Wikitext-103 dataset. Training inputs and labels are designed accordingly. An entity masking mechanism is introduced to guide the model in learning strategies to avoid sensitive content. The training follows a standard language modeling objective. We monitor changes in the sensitivity distribution of model outputs throughout the fine-tuning process. The following table presents the key configuration parameters used in this experiment. Its detailed configuration is shown in Table 1.

Table 1: Specific parameter diagram

Parameter name	Setting Value
Basic Model	ChatGLM-6B
Epochs	200
Learning Rate	2e-5
Batch Size	16
Maximum input length	1024
Optimizer	AdamW

4.3 Experimental Results

1) Comparative experimental results

This paper first gives the comparative experimental results, as shown in Table 2.

Table2: Comparative experimental results

Method	Sensitive Exposure Rate	Contextual Coherence	Generation Quality
Full Fine-Tuning[24]	8.21%	91.2	90.1
Adapter-Tuning[25]	6.75%	89.6	88.4
LoRA[26]	5.98%	90.3	89.7
Prefix-Tuning[27]	7.42%	87.9	87.2
Ours	3.87%	92.8	91.4

As shown in Table 2, the proposed method significantly outperforms existing mainstream fine-tuning strategies in controlling sensitive entity output. Specifically, after applying Selective Fine-tuning with the Semantic-guided Masking Strategy (SF + SGMS), the model's sensitive exposure rate drops to 3.87 percent. This represents a clear improvement compared to full-parameter fine-tuning (8.21 percent) and common lightweight methods such as Adapter-Tuning (6.75 percent), LoRA (5.98 percent), and Prefix-Tuning (7.42 percent). These results indicate that fine-tuning only the parameter subspaces closely related to sensitive content, combined with semantic attention suppression, effectively reduces the model's tendency to respond to sensitive information.

In terms of maintaining contextual coherence, the proposed method also shows a leading advantage. Our method achieves a Contextual Coherence score of 92.8. This is noticeably higher than other methods, especially when compared to LoRA (90.3) and Adapter-Tuning (89.6). This demonstrates that semantic-guided masking not only suppresses sensitive entity generation but also preserves the model's ability to organize language and understand context. This balance between semantic protection and generation quality is difficult to achieve with most existing fine-tuning strategies.

For text generation quality, our method scores 91.4, outperforming all baseline methods. This confirms that the selective fine-tuning mechanism helps preserve the core generation ability of the model. By locally intervening in sensitive regions, the model still maintains strong fluency and naturalness in overall language expression. In contrast, while full-parameter fine-tuning offers strong adaptability, it often causes knowledge forgetting. Prefix-tuning, though lightweight, sacrifices substantial control over language generation, leading to inferior results compared to our method.

In summary, the proposed method achieves the best performance across all three core metrics. This reflects its structural innovation in the context of sensitive information control. Compared to existing parameter-efficient fine-tuning methods, our strategy focuses more on fine-grained behavioral adjustment. It provides an effective and practical solution, especially for language model applications where generation safety is critical. These findings also demonstrate the theoretical potential and practical value of the “selective + semantic masking” mechanism for privacy-aware generation.

2) Ablation Experiment Results

This paper also further gives the results of the ablation experiment, and the experimental results are shown in Table 3.

Table 3: Ablation Experiment Results

Method	**Sensitive Exposure Rate**	**Contextual Coherence**	**Generation Quality**
Baseline	8.02%	89.3	88.6
+PSM	5.64%	91.1	89.7
+SGMS	4.92%	90.4	90.2
Ours	3.87%	92.8	91.4

The ablation results in Table 3 show that both core modules proposed in this study—the Parameter Selection Mechanism (PSM) and the Semantic-guided Masking Strategy (SGMS)—play a critical role in controlling sensitive entity output. Compared to the baseline model, introducing only the PSM module reduces the sensitive exposure rate from 8.02 percent to 5.64 percent. This demonstrates that the selective fine-tuning mechanism effectively identifies and intervenes in parameter regions associated with sensitive content. It confirms the feasibility of behavior control at the parameter space level.

With the addition of the SGMS module, the sensitive exposure rate further decreases to 4.92 percent. At the same time, both contextual coherence and generation quality improve. This result indicates that SGMS not only dynamically detects the locations of sensitive entities but also suppresses the model's reliance on sensitive information during generation through attention modulation. While PSM focuses on parameter-level intervention, SGMS operates at the semantic level. Together, they regulate generation behavior from different perspectives and show strong complementarity.

When both modules are combined, the full method achieves the best performance across all three metrics. The sensitive exposure rate drops to 3.87 percent. Contextual coherence and generation quality increase to 92.8 and 91.4, respectively, far exceeding the baseline. These results show that jointly modeling parameter paths and semantic responses allows the model to avoid sensitive content while retaining its language generation abilities. This leads to safer and more controllable text generation.

In conclusion, the integration of PSM and SGMS forms a fine-grained fine-tuning framework. It effectively suppresses the risk of sensitive information generation while maintaining content quality. This structure not only supports the theoretical foundation of the study but also provides a scalable practical path for privacy control in language models.

3) Analysis of the effectiveness of parameter selection mechanisms in different hierarchical structures

This paper also presents a detailed effectiveness analysis of the parameter selection mechanism when applied to different hierarchical structures within the language model. The goal of this analysis is to understand how the positioning of parameter updates—across various layers such as embedding, lower, middle, upper, and output layers—affects the mechanism's ability to regulate sensitive entity generation. By investigating the performance of the parameter selection mechanism at each structural level, the study seeks to reveal which layers contribute most significantly to the encoding and expression of sensitive information. This analysis helps clarify the internal roles played by different parts of the model and supports more informed decisions about where selective fine-tuning should be concentrated.

The hierarchical evaluation also provides insight into the interaction between structural depth and behavioral control granularity. Since different layers in transformer-based language models serve distinct functional purposes—ranging from basic lexical representation to higher-level semantic reasoning—targeting specific layers for parameter updates can result in varying degrees of control and stability. The analysis is intended to capture this variation and guide the design of more effective fine-tuning strategies. The findings from this analysis are visually represented in Figure 4, which illustrates the role of structural depth in mediating the parameter selection mechanism's influence on model behavior.

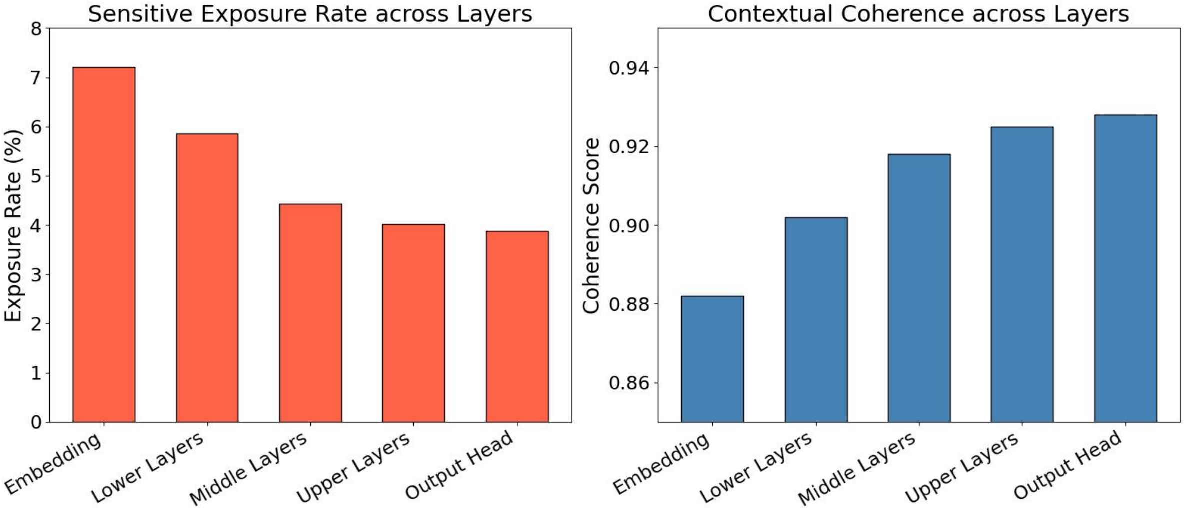

Figure 4. Analysis of the effectiveness of parameter selection mechanisms in different hierarchical structures

Figure 4 illustrates the variation in sensitive entity control effectiveness across different structural layers of the language model under the Parameter Selection Mechanism (PSM). As shown in the bar chart on the left, fine-tuning only the embedding layer or lower structural layers results in higher sensitive exposure rates (7.21 percent and 5.86 percent). This suggests that these layers have limited capacity to intervene in sensitive content during generation. This limitation exists because lower layers primarily learn general word embeddings or syntactic features, which are not strongly associated with specific semantic entities. As a result, they struggle to impose effective constraints on sensitive outputs.

As fine-tuning progresses toward the middle and upper layers, the sensitive exposure rate significantly decreases. In particular, the middle layer yields a rate of 4.43 percent, while the upper layer achieves 4.02 percent. These results indicate that these structural regions exert a stronger influence over the modeling and generation of sensitive information. Selective adjustment at the output head leads to the lowest sensitive exposure rate of 3.87 percent. This is closely tied to mechanisms directly responsible for content generation. Therefore, fine-tuning parameter subsets located in high-level semantic decision layers enables more efficient privacy protection.

The bar chart on the right, showing contextual coherence scores, further supports this conclusion. As the fine-tuning level moves upward, the coherence score steadily increases, rising from 0.882 at the embedding layer to 0.928 at the output layer. This indicates that parameter updates closer to the generation endpoint cause less disruption to the model's overall language organization. They better support sensitive information control while preserving text coherence and fluency.

In summary, these experimental results validate the layer-wise effect of the Parameter Selection Mechanism. They offer valuable insights for building efficient and secure language models. Choosing the appropriate parameter location not only improves control over model behavior but also avoids damage to general language ability. This enhances the fine-tuning strategy's capacity to balance safety and expressiveness. These findings further demonstrate the practicality and effectiveness of the proposed method in fine-grained behavior modulation.

4) Intervention experiment on contextual resilience of sensitive content

This paper then presents an intervention experiment focused on evaluating the contextual recovery capability of the language model when sensitive content is deliberately masked. This experiment aims to assess how well the model can reconstruct or maintain coherent semantic flow in the surrounding context when parts of the input related to sensitive entities are suppressed. This is especially important for ensuring that behavior regulation mechanisms, such as masking strategies, do not disrupt the model's overall understanding or lead to degradation in text quality. By testing the model's ability to infer and preserve the original meaning of a passage despite the absence of explicit sensitive elements, the study offers insight into the balance between content suppression and language coherence.

The intervention involves applying varying levels of semantic-guided masking to inputs containing sensitive content, followed by the analysis of the model's responses. This approach is designed to simulate realistic generation scenarios where sensitive information must be excluded while still producing fluent and meaningful output. The contextual recovery capability reflects the model's adaptability and robustness under constrained input conditions, serving as an important indicator of both safety and usability. Figure 5 presents the results of this experiment, illustrating how the model responds to different masking intensities and how effectively it can maintain contextual integrity despite semantic occlusion.

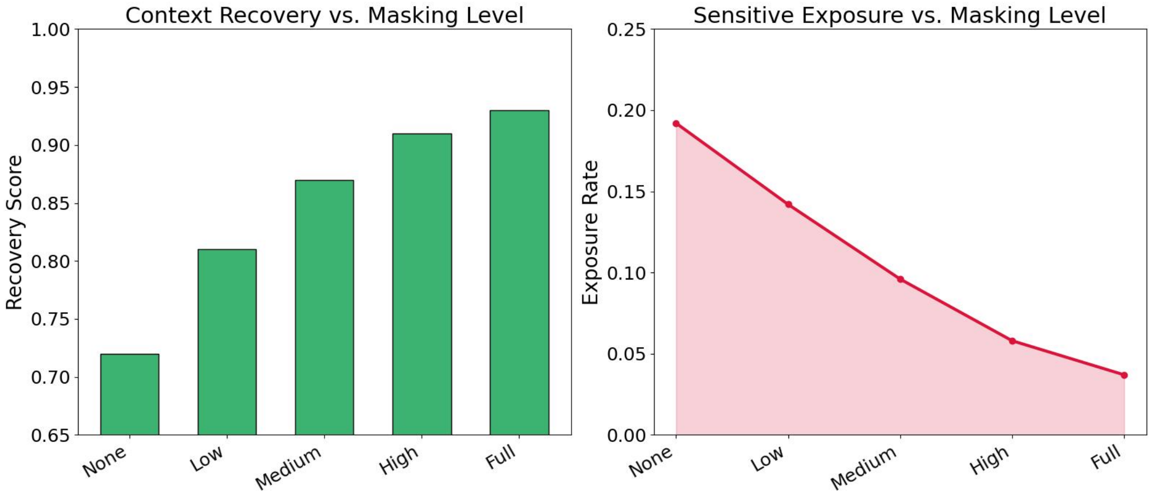

Figure 5. Intervention experiment on contextual resilience of sensitive content

Figure 5 presents the model's performance under different levels of masking intervention, focusing on contextual recovery and sensitive exposure rate. The left chart shows that as the masking level increases from None to Full, the contextual recovery score steadily rises from 0.72 to 0.93. This trend indicates that with appropriate intervention, the model can successfully reconstruct the semantic context after sensitive content is masked. It demonstrates strong semantic recovery capabilities. Notably, under high masking levels (High and Full), the model still maintains high semantic integrity. This suggests that the guided masking does not disrupt the model's language understanding structure. Instead, it enhances the model's ability to complete the surrounding context.

The right chart further confirms the effectiveness of this strategy in suppressing sensitive content exposure. As the masking level increases, the sensitive entity exposure rate shows a marked decline, dropping from 0.192 to 0.037. This result demonstrates that the semantic-guided masking mechanism effectively reduces the model's dependence on sensitive information. By guiding the masking and reconstruction paths, the mechanism prevents sensitive content from being reproduced during generation. This aligns closely with the study's goal of improving generation safety through behavioral regulation.

A joint analysis of both metrics shows that masking intervention suppresses sensitive entity expression without harming language quality. It also strengthens the model's adaptability in semantically incomplete settings. Compared to traditional static filtering, this semantic dynamic masking improves the granularity of control while preserving the model's ability for natural contextual inference. It serves as a structural optimization strategy that balances safety and expressiveness.

In conclusion, the experiment validates the dual contribution of the semantic masking strategy. It reduces exposure risk in sensitive information control while enhancing the model's ability to reconstruct masked semantics. This mechanism offers important insights for developing highly secure and robust large language models. It also expands the practical path for implementing selective behavior control within the model.

5) Experiment on dynamic changes of sensitive exposure rate with different fine-tuning steps

This paper also experimented on the dynamic change of sensitive exposure rate with different fine-tuning steps, and the experimental results are shown in Figure 6.

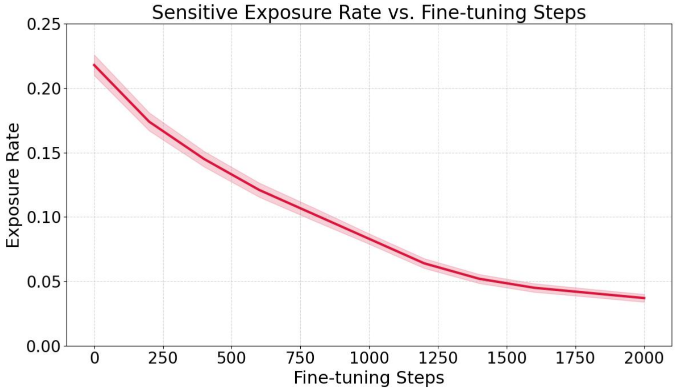

Figure 6. Experiment on dynamic changes of sensitive exposure rate with different fine-tuning steps

Figure 6 shows the dynamic trend of sensitive entity exposure rate as the number of fine-tuning steps increases. The overall curve demonstrates a continuous downward trajectory, indicating that the model's tendency to generate sensitive content weakens with more training iterations. The most significant drop occurs within the first 1000 steps, where the exposure rate falls from an initial 0.218 to approximately 0.064. This confirms that the selective fine-tuning and semantic control mechanisms can effectively influence model behavior at an early training stage.

As training progresses, the rate of decline slows and eventually stabilizes. This suggests that the model quickly adapts to high-sensitivity regions through early adjustments of key parameter subsets. In the later stages, fine-tuning mainly serves to reinforce and refine these effects. At 2000 steps, the exposure rate reaches a final value of 0.037. This result indicates that the proposed method maintains consistent suppression of sensitive content generation in later training phases, without signs of rebound or degradation in generalization.

The shaded area in the figure represents the fluctuation range across training stages. The narrow range reflects the good stability and robustness of the method at different training points. This also implies that the parameter selection mechanism consistently targets the appropriate regions for update. It avoids interfering with non-sensitive semantic pathways, enabling controlled adjustment of sensitive content generation behavior.

This experiment further supports the effectiveness of the proposed mechanism. By selectively updating parameters closely related to sensitive expressions during fine-tuning, and combining this with a semantic-level control strategy, the model can achieve continuous and stable suppression of sensitive entity generation. This approach holds strong practical value and potential for broader research adoption.

5. Conclusion

This paper addresses the problem of sensitive entity generation in large language models. It proposes a method that combines selective fine-tuning with a semantic-guided masking strategy. The goal is to regulate sensitive content generation without compromising the model's language capabilities. To address the limitations of traditional fine-tuning, which updates the entire parameter space indiscriminately and lacks precision, this study introduces a parameter selection mechanism. It identifies and focuses updates on parameter subspaces associated with sensitive expressions. This improves both fine-tuning efficiency and control accuracy. At the same time, the semantic-guided masking strategy constructs dynamic probabilistic masks. These masks guide the model to avoid overreliance on sensitive semantics within the attention mechanism. This cuts off potential leakage pathways at the generation level.

Comprehensive experimental results show that the proposed method outperforms existing fine-tuning strategies across multiple metrics, including sensitive exposure rate, contextual coherence, and generation quality. It enhances the safety of model outputs and validates the effectiveness of structured fine-tuning in behavior control tasks. This study also includes ablation experiments, layer-wise sensitivity analysis, and contextual robustness tests. These analyses systematically verify the individual contributions and synergistic effects of each module. They improve the interpretability and robustness of the model's behavior control. This systematic approach not only applies to privacy protection tasks but also provides a general technical framework for other generation control problems, such as value alignment and bias mitigation.

From an application perspective, the proposed method offers high practical value in scenarios that demand strong compliance and sensitivity control. These include dialogue systems, content generation, medical text automation, and government question answering. By enhancing the controllability of model behavior, this method helps mitigate safety risks in open-domain generation. It also reduces ethical and legal costs during deployment. Moreover, its fine-grained intervention capabilities offer new directions for customized and modular training, extending the adaptability of language models in specialized domains.

6. Future Work

Looking forward, the proposed mechanism has broad potential for expansion. The parameter selection mechanism could be combined with domain-specific knowledge graphs and private annotation systems to achieve multidimensional behavior targeting. The semantic masking strategy could be integrated with cross-modal information to support dynamic sensitivity detection in text, image, and speech content. Future work may also explore more efficient training paradigms, such as incremental updates, online fine-tuning, and federated learning. These directions aim to meet real-world demands on data privacy, computational resources, and model flexibility, laying a stronger foundation for building trustworthy, controllable, and compliant generative AI systems.

References

[1] Charles, Zachary, et al. "Fine-tuning large language models with user-level differential privacy." arXiv preprint arXiv:2407.07737 (2024).

[2] Chen, Xiaoyi, et al. "The janus interface: How fine-tuning in large language models amplifies the privacy risks." Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. 2024.

[3] Wang, Teng, et al. "Selective privacy-preserving framework for large language models fine-tuning." Information Sciences 678 (2024): 121000.

[4] Chen, Tiejin, et al. "Privacy-preserving fine-tuning of large language models through flatness." arXiv preprint arXiv:2403.04124 (2024).

[5] Behnia, Rouzbeh, et al. "Ew-tune: A framework for privately fine-tuning large language models with differential privacy." 2022 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2022.

[6] Peris, Charith, et al. "Privacy in the time of language models." Proceedings of the sixteenth ACM international conference on web search and data mining. 2023.

[7] Chua, Lynn, et al. "Mind the privacy unit! user-level differential privacy for language model fine-tuning." arXiv preprint arXiv:2406.14322 (2024).

[8] Yan, Biwei, et al. "On protecting the data privacy of large language models (llms): A survey." arXiv preprint arXiv:2403.05156 (2024).

[9] Li, Yansong, Zhixing Tan, and Yang Liu. "Privacy-preserving prompt tuning for large language model services." arXiv preprint arXiv:2305.06212 (2023).

[10] Du, Hao, et al. "Privacy in Fine-tuning Large Language Models: Attacks, Defenses, and Future Directions." arXiv preprint arXiv:2412.16504 (2024).

[11] Ding, Ning, et al. "Parameter-efficient fine-tuning of large-scale pre-trained language models." Nature Machine Intelligence 5.3 (2023): 220-235.

[12] Xu, Runxin, et al. "Raise a child in large language model: Towards effective and generalizable fine-tuning." arXiv preprint arXiv:2109.05687 (2021).

[13] Liu, Yixin, et al. "Improving large language model fine-tuning for solving math problems." arXiv preprint arXiv:2310.10047 (2023).

[14] Chen, Yukang, et al. "Longlora: Efficient fine-tuning of long-context large language models." arXiv preprint arXiv:2309.12307 (2023).

[15] Tinn, Robert, et al. "Fine-tuning large neural language models for biomedical natural language processing." Patterns 4.4 (2023).

[16] Susnjak, Teo, et al. "Automating research synthesis with domain-specific large language model fine-tuning." ACM Transactions on Knowledge Discovery from Data 19.3 (2025): 1-39.

[17] Li, Haoran, et al. "Privacy in large language models: Attacks, defenses and future directions." arXiv preprint arXiv:2310.10383 (2023).

[18] Zmushko, Philip, et al. "Privacy preserving API fine-tuning for LLMs." (2023).

[19] Du, Minxin, et al. "Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass." Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. 2023.

[20] Chen, Kang, et al. "A Survey on Privacy Risks and Protection in Large Language Models." arXiv preprint arXiv:2505.01976 (2025).

[21] Kuang, Weirui, et al. "Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning." Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024.

[22] Wu, Yebo, et al. "A Survey on Federated Fine-tuning of Large Language Models." arXiv preprint arXiv:2503.12016 (2025).

[23] Tang, Xinyu, et al. "Private fine-tuning of large language models with zeroth-order optimization." arXiv preprint arXiv:2401.04343 (2024).

[24] Zheng, Hongling, et al. "Learn from model beyond fine-tuning: A survey." arXiv preprint arXiv:2310.08184 (2023).

[25] Xing, Jialu, et al. "A survey of efficient fine-tuning methods for vision-language models—prompt and adapter." Computers & Graphics 119 (2024): 103885.

[26] Mao, Yuren, et al. "A survey on lora of large language models." Frontiers of Computer Science 19.7 (2025): 197605.

[27] Le, Minh, et al. "Revisiting prefix-tuning: Statistical benefits of reparameterization among prompts." arXiv preprint arXiv:2410.02200 (2024).