Breach in the Shield: Unveiling the Vulnerabilities of Large Language Models

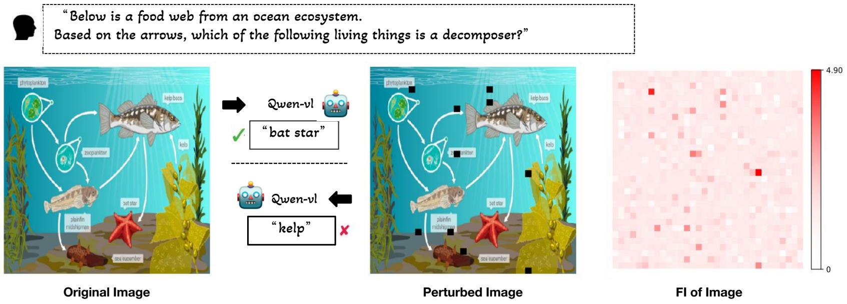
Runpeng Dai[1] **Run Yang**[2] **Fan Zhou**[3] **Hongtu Zhu**[1†]
[1]University of North Carolina at Chapel Hill [2]BiliBili
[3]Shanghai University of Finance and Economics
{runpeng, htzhu}@email.unc.edu
yangrun@BiliBili.com
zhoufan@mail.shufe.edu.cn

Abstract

Large Language Models and Vision-Language Models have achieved impressive performance across a wide range of tasks, yet they remain vulnerable to carefully crafted perturbations. In this study, we seek to pinpoint the sources of this fragility by identifying parameters and input dimensions (pixels or token embeddings) that are susceptible to such perturbations. To this end, we propose a stability measure called **FI**, **F**irst order local **I**nfluence, which is rooted in information geometry and quantifies the sensitivity of individual parameter and input dimensions. Our extensive analysis across LLMs and VLMs (from 1.5B to 13B parameters) reveals that: (I) A small subset of parameters or input dimensions with high FI values disproportionately contribute to model brittleness. (II) Mitigating the influence of these vulnerable parameters during model merging leads to improved performance.

1 Introduction

Large Language Models (LLMs) and Vision Language Models (VLMs) such as GPT [5] and Llama [36], have revolutionized the field of Natural Language Processing (NLP), exhibiting remarkable proficiency across a variety of tasks [14, 47, 24] and modalities [3, 21, 46]. These modern LLMs are massive in size, trained on vast amounts of data, and meticulously aligned to prevent generating harmful content [30], leaking private information [45], or exhibiting sexual or religious bias [39]. Despite the enthusiasm for these integrative approaches, a critical issue remains: LLMs remain susceptible to both external and internal perturbations, affecting their reliability and performance.



"Below is a food web from an ocean ecosystem.
Based on the arrows, which of the following living things is a decomposer?"

Figure 1: A case study of the Qwen-VL model [3] on SCI-QA. The image on the far right visualizes the per-pixel FI values. Masking just 10 pixels with the highest FI values leads to a failure in producing the correct answer.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ResponsibleFM.

Externally, LLMs are vulnerable to input perturbations, such as Embedding-Corrupted Prompts [11, 20] and jail-breaking prompts [50, 49]. This susceptibility extends to visual inputs in VLMs, where adversarially optimized images can drastically alter model behavior [31]. Beyond adversarial attacks, VLMs exhibit high sensitivity to perturbations in specific local regions of an image—a common issue, as user-uploaded images often suffer from blurring, masking, or low resolution. The vulnerability is highlighted in our case study of the Qwen-VL model. As depicted in Figure 1, masking the ten most sensitive pixels, which are unrelated to the question, resulted in incorrect model outputs.

Internally, LLM stability is further challenged by parameter perturbations, often introduced through model merging and quantization. While these techniques improve deployment efficiency by reducing inference costs [12, 1], they can also induce hallucinations and degrade performance [26, 44, 19]. However, our findings reveal that parameter susceptibility varies significantly. As Figure 2 illustrates, randomly dropping 5% of parameters has a minimal impact on performance. In contrast, zeroing out just 1% of the parameters identified by our measure can drastically reduce accuracy, even below random guessing levels.

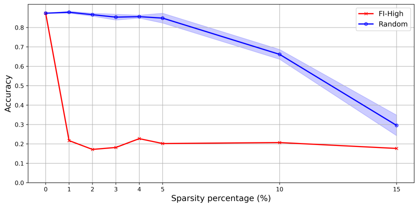


Figure 2: A case study of Qwen2.5 on MMLU-Geography. “FI-High” refers to zeroing out parameters with the highest FI values, whereas “Random” denotes random parameter removal.

To pinpoint the sources of this fragility, we propose a novel stability measure called **FI**, **F**irst order local **I**nfluence, to quantitatively assess the stability of LLMs against perturbations. Specifically, we construct a perturbation manifold that encompasses all perturbed models, along with its associated geometric properties. Our stability measure quantifies the degree of local influence of a perturbation on a given objective function within this manifold, thereby reflecting the stability of individual LLM components. We summarize the **advantages** of **FI** as follows:

1. **FI’s versatility** allows for effective stability assessment under both external and internal perturbations across various granularities—from individual parameters to input features like pixels and patches.
2. **FI effectively identifies vulnerabilities**. Our extensive studies validate its effectiveness in pinpointing fragile pixels in VLM vision inputs, vulnerable embedding dimensions of tokens in LLMs (Section 4), and salient model parameters (Subsection 5.1).
3. **FI offers insights into improving model robustness**. We further illustrate that understanding these vulnerabilities can lead to enhanced model resistance to perturbations. By focusing on model merging as an example, we show that safeguarding key parameters identified by high FI values can substantially reduce performance degradation during the merging process (Subsection 5.2).

2 Related Work

Recent efforts to evaluate LLM stability typically adopt a coarse-grained approach, aiming to assess the overall robustness of models under various perturbations. One line of work investigates how stability is influenced by sampling parameters, such as temperature, which affect output variability during generation [2, 28]. Another direction studies model sensitivity to input or parameter perturbations. For instance, [4] analyze input-level robustness using optimal transport to quantify a model’s response to distributional shifts in prompts. On the other hand, [29] focus on parameter-space perturbations, demonstrating that LLMs remain robust to weight changes up to a certain threshold, beyond which performance significantly degrades. They estimate a model’s robustness tolerance by injecting random perturbations into model weights and evaluating the performance drop.

Despite these contributions, fine-grained analyses—such as those examining the effect of individual input tokens, pixels, or specific model parameters—remain underexplored. [37] take a step in this direction by leveraging pruning-based techniques, including SNIP [18] and Wanda [35], to identify critical neurons and low-rank structures that impact model safety and utility. However, there is still a lack of unified metrics or frameworks that assess stability with respect to both input- and parameter-level perturbations. Moreover, the downstream applications of such stability assessments remain largely unexamined.

3 Stability Measure of Large Language Models

In this section, we propose a new metric called FI to quantify the stability of large language models against local perturbations. Considering the auto-regressive nature of LLMs, we first develop FI for single-step generation and discuss its theoretical and computational properties in detail. We then show how FI can be naturally extended to sequence generation tasks. Finally, we compare FI to existing stability measures, highlighting its unique advantages.

3.1 FI Metric

Problem formulation. Consider an LLM parameterized by θ, with input data x, which may consist of text or, for visual language models, a combination of text and images. Given x, the model generates a probability distribution over its vocabulary to predict the next token, which can be framed as a classification problem with K classes, where K represents the vocabulary size.

However, vocabulary sizes are typically large [3, 9], and predictions are often concentrated on a small subset of tokens. Instead of using the entire vocabulary, it is more efficient to focus on a relevant subset based on the task. For example, in multiple-choice questions, probabilities are restricted to the choices "A", "B", "C", or "D". Classes can also be defined semantically, such as categorizing tokens as "neutral" or "notorious" in toxicity detection [13].

With appropriately defined classes, the predicted probability for class $y \in \{1, \ldots, K\}$ is denoted as $P(y|x, \theta)$, satisfying $\sum_{y=1}^{K} P(y|x, \theta) = 1$. Let $\omega \in \mathbb{R}^d$ be a perturbation vector varies in an open subset Ω. ω can be applied to a subset of the model parameters θ and locations within the input data x. We denote the output of the perturbed model under this perturbation as $P(y|x, \theta, \omega)$.

Perturbation Manifold and FI Since our primary interest lies in examining the behavior of $P(y|x, \theta, \omega)$ as a function of ω near $\omega_0 = 0$, we shift focus from θ to ω. We introduce the perturbation manifold as defined in [51] and [52].

Definition 3.1. *Define the d-dimensional perturbation manifold $\mathcal{M} = \{P(y|x, \theta, \omega) : \omega \in \Omega\}$, which encompasses all perturbed models. Assume that for all $\omega \in \Omega$, the perturbed models $\{P(y = i|x, \theta, \omega)\}_{i=1}^{K}$ are positive and sufficiently smooth. The tangent space T_ω of $\mathcal{M}$ at ω is spanned by the partial derivatives of the log-likelihood function $\ell(\omega|y, x, \theta) = \log P(y|x, \theta, \omega)$ with respect to ω, specifically $T_\omega = span\{\frac{\partial}{\partial \omega_i} \ell(\omega|y, x, \theta)\}_{i=1}^{d}$.*

The metric g_ω on $\mathcal{M}$ can be defined with the metric tensor G_ω. Consider two tangent vectors at ω given by $v_j(\omega) = h_j^\top \partial_\omega \ell(\omega|y, x, \theta) \in T_\omega$, where h_1 and h_2 are the weights on the basis. Their inner product is defined as:

$$\langle v_1(\omega), v_2(\omega) \rangle_{g_\omega} = \sum_{y=1}^{K} v_1(\omega) v_2(\omega) P(y|x, \theta, \omega).$$

The metric tensor G_ω is given by:

$$G_\omega = \sum_{y=1}^{K} \partial_\omega \ell(\omega|y, x, \theta) \partial_\omega^\top \ell(\omega|y, x, \theta) P(y|x, \theta, \omega).$$

Subsequently, the norm of $v_j(\omega)$ under metric g_ω is $\|v_j\|_{g_\omega} = \sqrt{h_j^\top G_\omega h_j}$. Let $C(t) = P(y|x, \theta, \omega(t))$ be a smooth curve on the manifold $\mathcal{M}$ connecting two points $\omega_1 = \omega(t_1)$ and $\omega_2 = \omega(t_2)$. Then, the distance between ω_1 and ω_2 along the curve $C(t)$ is given by:

$$\begin{aligned} S_C(\omega_1, \omega_2) &= \int_{t_1}^{t_2} \sqrt{\|\partial_t \log P(y|x, \theta, \omega(t))\|_{g_\omega}}\, dt \\ &= \int_{t_1}^{t_2} \sqrt{\frac{d\omega(t)^T}{dt} G_{\omega(t)} \frac{d\omega(t)}{dt}}\, dt. \end{aligned}$$

With the Perturbation manifold $\mathcal{M}$ and respective metric g_ω defined, we are ready to propose the metric that quantifies the stability of large language models (LLMs) against various types of local perturbations. Let $f(\omega)$ be the objective function of interest for sensitivity analysis, in our case being $-\log P(y_{pred}|x,\theta,\omega)$, we can define the following (first-order) local influence metric FI:

Definition 3.2. *Given the perturbation manifold $\mathcal{M}$ and its metric, the first-order local stability measure of $f(\omega)$ at $\omega(0)=\omega_0$ is defined as*

$$\mathbf{FI}_\omega\left(\omega_0\right)=\max_{C}\lim_{t\to 0}\frac{[f(\omega(t))-f(\omega(0))]^2}{S_C^2(\omega(t),\omega(0))}. \tag{1}$$

The ratio in Equation 1 measures the amount of change introduced to the objective function relative to the distance of the perturbation on the perturbation manifold. Thus, Equation 1 can be naturally interpreted as the maximum local ratio of change among all possible perturbation curves $C(t)$.

Computation of FI. As we will show, Theorem A.1 in Appendix A.3 regarding diffeomorphic reparameterization invariance enables us to derive an easy-to-compute solution for Equation 1, while addressing the low-dimensionality problem inherent in LLMs.

Theorem 3.3. *If G_ω is positive definite, the **FI** measure has the following closed-form:*

$$\mathbf{FI}_\omega\left(\omega_0\right)=\nabla_{f(\omega_0)}^{T}G_{\omega_0}^{-1}\nabla_{f(\omega_0)}, \tag{2}$$

where

$$\nabla_{f(\omega_0)}=\frac{\partial f(\omega)}{\partial\omega}\Big|_{\omega=\omega_0}.$$

The detailed proof of Theorem 3.3 can be found in Appendix A.6. It is important to note that the closed form of FI in Theorem 3.3 depends on the positive definiteness of G_ω, which is not always guaranteed. This is due to the fact that the parameters in LLMs are often high-dimensional tensors with low-rank structures [17].

We apply the invariance result of Theorem A.1 in Appendix A.3 by transforming ω to a vector ν such that $G_\nu=\mathbf{I}_K$, where K is an integer. Specifically, we notice that $G_{\omega_0}=B_0^T B_0$, where

$$B_0=\left[P(y=i|x,\theta,\omega)^{1/2}\partial_\omega\ell(\omega|y=i,x,\theta)\right]_{i\leqslant K}.$$

Let $r_0=\text{rank}(G_{\omega_0})$, we apply the compact SVD to $B_0\in\mathbb{R}^{p\times K}$, which yields $B_0=V_0\Lambda_0 U_0$, where $V_0\in\mathbb{R}^{p\times r_0}$ and $U_0\in\mathbb{R}^{r_0\times K}$ are semi-orthogonal matrices and $\Lambda_0\in\mathbb{R}^{r_0\times r_0}$ is a diagonal matrix. Under the transformation $\nu=\Lambda_0 V_0^T\omega$, we have $\mathbf{FI}_\omega\left(\omega_0\right)=\mathbf{FI}_\nu\left(\nu_0\right)$, which can be expressed as

$$\nabla_{f(\omega_0)}^{\top}\left(V_0 R_0\right)^{\top}\Lambda_0^{-2}\left(V_0 R_0\right)\nabla_{f(\omega_0)},$$

where the equality holds by applying the chain rule to G_ν.

FI for sequence generation. Sequence generation is essentially multiple rounds of next-token generation, where the l-th token $y^{(l)}$ is generated given the initial input z and previously generated tokens $\boldsymbol{y}^{(l)}=\{y^{(1)},\ldots,y^{(l-1)}\}$. We define the FI measure for generating the l-th token $y^{(l)}$ given the initial input z by averaging out the randomness from the preceding steps $\mathbf{FI}_l(z)=\mathbb{E}_{\boldsymbol{y}^{(l)}}[\mathbf{FI}(\{z,\boldsymbol{y}^{(l)}\},\theta,\omega)|z]$.

To formulate an overall measure for sequence generation, we aggregate these per-token FI measures. Since sequences generated by LLMs can vary in length, we propose two methods to handle this heterogeneity. The first approach sets a fixed horizon L and computes the mean FI over these rounds

$$\mathbf{FI}_{\text{seq}}^{L}(z)=\frac{1}{L}\sum_{l=1}^{L}\mathbf{FI}_l(z). \tag{3}$$

Alternatively, inspired by the concept of average discounted rewards in reinforcement learning [22], we consider sequences of potentially infinite length and propose a discounted FI measure with discount factor γ

$$\mathbf{FI}_{\text{seq}}^{\infty,\gamma}(z)=(1-\gamma)\sum_{l=0}^{\infty}\gamma^l\cdot\mathbf{FI}_l(z).$$

By taking the expectation over the distribution of z, we obtain the average FI for sequence generation in both cases $\mathbb{E}_{P_z}[\mathbf{FI}_{\text{seq}}^{L}(z)]$ and $\mathbb{E}_{P_z}[\mathbf{FI}_{\text{seq}}^{\infty,\gamma}(z)]$, respectively.

3.2 Other Measures & Discussion

We note that several alternative methods can also serve as stability measures for LLMs. We provide their explicit formulations and compare them with FI.

Jacobian Norm [27]: $\|\partial_\omega f(y_{pred}, \omega)\|_2$

SNIP [18]: $\|\omega \odot \partial_\omega f(y_{pred}, \omega)\|_2$

Both measures focuses solely on y_{pred}, while neglecting the probabilities assigned to other choices. For example, consider two output distributions: (0.9, 0.05, 0.05, 0.02) and (0.3, 0.25, 0.25, 0.2). In both cases, the model selects option A. However, the second distribution is more unstable, as a small perturbation in the probabilities could lead to a different prediction. In contrast, FI measure accounts for both the probability and gradient across all possible choices.

Saliency map [34]:

$$\begin{cases} 0, \quad \text{if } \dfrac{\partial f(y_{\text{pred}}, \omega)}{\partial \omega} < 0 \text{or} \displaystyle\sum_{y \neq y_{\text{pred}}} \dfrac{\partial f(y, \omega)}{\partial \omega} > 0 \\ -\dfrac{\partial f(y_{\text{pred}}, \omega)}{\partial \omega} \displaystyle\sum_{y \neq y_{\text{pred}}} \dfrac{\partial f(y, \omega)}{\partial \omega}, \text{ otherwise} \end{cases}$$

Saliency maps consider the gradients with respect to all possible choices. However, they lose significant information by zeroing out many of these gradients.

To this end, we highlight the unique advantages of FI. **Effectiveness:** A quantitative comparison of these measures is provided in Section 4 and 5, while their computational complexities are discussed in Appendix A.2. **Theoretical rigor:** In particular, only FI possesses a reparameterization invariance property (see Appendix A.3), which further distinguishes it by enhancing interpretability.

4 External Perturbations Analysis

In this section, we first demonstrate the effectiveness of FI in identifying vulnerable locations in both vision and language inputs through guided attack. Then, we conclude the section with a finding from cross-modal analysis.

Identify Fragile Pixels We conduct the attack process on the MMbench dataset [23], a comprehensive benchmark designed to evaluate various multimodal capabilities of VLMs. For a fair comparison, we identify the top 10 pixels using different stability measures and assess the model's performance after masking out the corresponding pixels.

Identify Vulnerable Embedding Dimensions We conduct attack on pure-text LLMs to verify the effectiveness of our approach in identifying vulnerable embedding dimensions. Specifically, we follow the token embedding attack methods proposed in [20] and [11].

More concretely, we compute the stability measure for each embedding dimension and select the top 0.1% most sensitive dimensions (ω) as identified by the metrics. We then apply a gradient-based attack strategy following [11], perturbing the selected dimensions in the direction of $-\nabla_\omega \log P(y_{\text{pred}} \mid x, \theta)$.

From both Table 1 and Table 2, we observe the following: (I) Stability measures are effective in identifying vulnerable input dimensions (i.e., pixels in images and dimensions in embeddings). Notably, LLMs are generally robust to random perturbations and such perturbations rarely lead to significant performance degradation. In contrast, perturbations guided by stability measures consistently result in substantial drops in performance. (II) Among all the stability measures evaluated, FI proves to be the most effective: masking pixels or perturbing dimensions identified by FI leads to the largest observed decline in performance.

Effect of Prompting on Pixel Vulnerability While the significant impact of prompt design on VLM performance is well-recognized [48], and carefully crafted prompts are known to even jailbreak these models [32], a quantitative analysis of this cross-modal influence – specifically, how prompting affects the processing and stability of visual input – remains largely unexplored.

Table 1: Accuracy on the MMBench dataset after masking out top ten pixels in images identified by different measures.

Model	**Method**	Action Recognition	Attribute Recognition	Celebrity Recognition	Function Reasoning
Qwen VL	FI (Ours)	**0.320**	**0.402**	**0.673**	**0.411**
	Jacobian	0.668	0.587	0.906	0.604
	Saliency	0.782	0.525	0.873	0.639
	Random	0.812	0.550	0.881	0.683
	Original	0.814	0.549	0.882	0.686
Qwen2.5 VL-3B	FI (Ours)	**0.720**	**0.735**	**0.780**	**0.723**
	Jacobian	0.731	0.752	0.797	0.755
	Saliency	0.745	0.761	0.797	0.774
	Random	0.882	0.931	0.957	0.928
	Original	0.890	0.946	0.959	0.930
Qwen2.5 VL-7B	FI (Ours)	**0.768**	**0.750**	**0.796**	**0.723**
	Jacobian	0.778	0.768	0.815	0.755
	Saliency	0.792	0.777	0.815	0.774
	Random	0.891	0.944	0.951	0.925
	Original	0.890	0.946	0.959	0.930

Table 2: Comparison of accuracy in the MMLU dataset after perturbing the same number of dimensions in the embedding space identified using different measures.

Model	**Method**	**Business**	**Geo**	**Culture**	**Law**
Pythia 1B	FI (ours)	**0.270**	**0.261**	**0.195**	**0.236**
	Saliency	0.278	0.272	0.210	0.243
	Jacobian	0.273	0.264	0.201	0.241
	Random	0.301	0.368	0.237	0.246
	SNIP	0.297	0.281	0.226	0.242
	Original	0.303	0.370	0.240	0.247
Qwen2.5 3B	FI (ours)	**0.656**	**0.620**	**0.610**	**0.547**
	Saliency	0.677	0.637	0.632	0.560
	Jacobian	0.665	0.641	0.625	0.560
	Random	0.805	0.781	0.781	0.672
	SNIP	0.783	0.663	0.665	0.563
	Original	0.810	0.800	0.785	0.673
Qwen2.5 7B	FI (ours)	**0.748**	**0.780**	**0.705**	**0.713**
	Saliency	0.756	0.789	0.709	0.725
	Jacobian	0.764	0.782	0.717	0.720
	Random	0.852	0.884	0.802	0.735
	SNIP	0.757	0.791	0.710	0.727
	Original	0.856	0.890	0.810	0.737

Our study aims to bridge this gap by investigating how varying prompt instructions influence the sensitivity of VLMs to visual perturbations. Specifically, we examine two types of prompts:

- Aggressive Prompts: Designed to encourage the model to consider every detail in the image, potentially increasing sensitivity to noise.
- Safe Prompts: Intended to focus the model on salient entities and relationships, potentially enhancing robustness by ignoring irrelevant details.

We computed the FI value for each pixel and visualized the resulting distributions under different prompt settings, as illustrated in Figure 3. Our main findings are as follows:

(I) Prompt choice has a substantial impact on the stability of individual pixels within the image. As shown on the left of Figure 3, aggressive prompts shift the FI distribution toward higher values, resulting in a marked increase in both the mean and maximum FI values. This suggests that the model

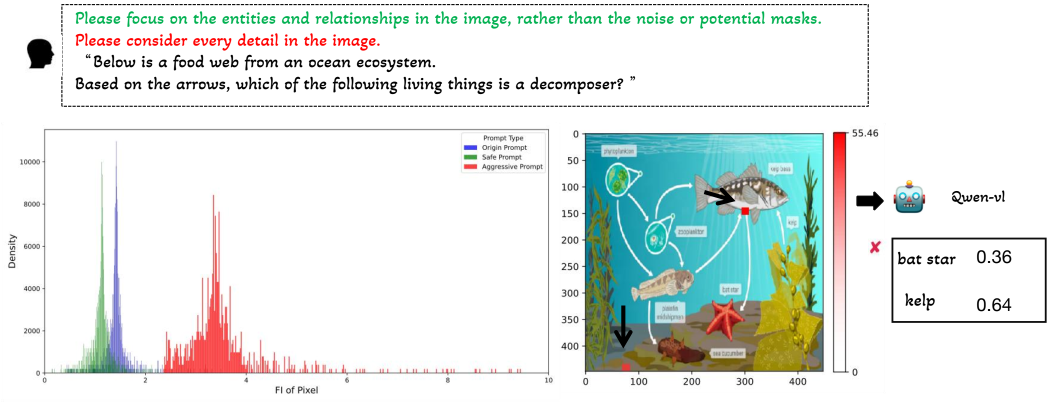


Figure 3: A case study utilizing FI for cross-modal analysis. In the same example, the bottom-left image shows how Aggressive and Safe prompts affect the FI distribution on the image.

becomes more sensitive to pixel-level perturbations throughout the image. In contrast, safe prompts significantly shift the FI distribution toward lower values, indicating reduced sensitivity and improved stability against perturbations in less relevant regions.

(II) Vulnerability remains even with careful prompt design. Although safe prompts generally reduce FI values, they do not fully guarantee model stability, as outliers with large FI values persist. As shown in the right column of Figure 3, even when applying the safe prompt, masking out the two pixels with the highest FI values still leads to incorrect model predictions. This result underscores the persistent challenge of achieving robustness in VLMs and demonstrates the effectiveness of the FI measure for identifying vulnerable regions.

Our findings contribute to the growing body of literature on cross-modal interactions in VLMs, offering a stability-centric perspective that complements existing behavioral and attributional analyses. Importantly, this framework can inform the development of more robust multimodal systems and prompt design strategies for safety-critical applications.

5 Internal Perturbations Analysis

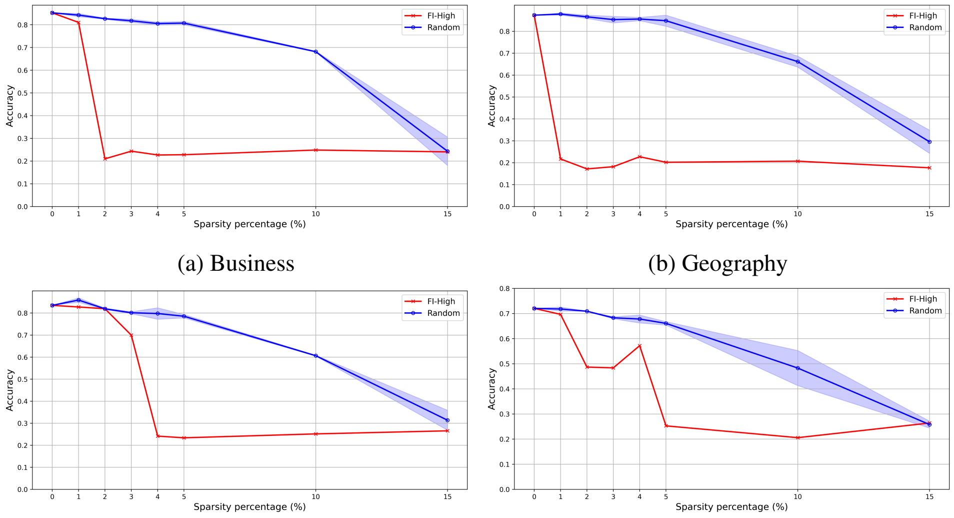
In this section, we first conduct a parameter sparsification experiment to demonstrate the effectiveness of the FI. We then apply the FI measure to mitigate parameter interference during model merging, showcasing its potential for guiding LLM improvement.

5.1 Parameter Sparsification

We conduct experiments on multiple-choice problems from MMLU [15] and sequence generation tasks from Alpaca-Eval [10] to examine how these perturbations impact two key capabilities of large models: knowledge retention and instruction-following. Details of both experimental setups are provided in Appendix A.

As shown in Figure 2 and 4, sparsifying (zeroing out) just 2–3% of the high-FI parameters significantly degrades the model's knowledge capacity, leading to catastrophic forgetting and hallucinations, with performance dropping by up to 75%. A similar trend is observed in Table 6 at around the 10% sparsity level. In contrast, models remain relatively robust against random sparsification, often exhibiting nearly identical behavior even after 5% sparsification.

These findings demonstrate FI's effectiveness in identifying fragile parameters and further support the inherent structure within the parameter matrix, aligning with recent observations on model brittleness [25, 37, 44].



(a) Business

(b) Geography

(c) Culture

(d) Law

Figure 4: Performance of Qwen2-7B on the MMLU dataset under varying levels of parameter sparsification. “FI-High” denotes sparsifying parameters with the highest FI values, while “Random” refers to random parameter sparsification.

5.2 FI-Guided Parameter Protection in Model Merging

Model merging is a technique for acquiring domain-specific knowledge by combining models from different domains, thereby reducing the computational cost of additional fine-tuning (see [43] for a review). However, a persistent challenge is that merging parameters introduces perturbations that can hinder a model’s ability to retain previously learned information. To address this, we use FI to identify parameters susceptible to forgetting and exclude them from the merging process.

We demonstrate that FI can be seamlessly integrated into mainstream model merging methods, including **Average Merging** [38], **Task Arithmetic** [16], and **TIES** [40]. Additionally, we include **DARE** [44] as a competing baseline for completeness.

We consider merging two models, A and B, both fine-tuned from the same base model. Let θ_A, θ_B, and θ_{Base} denote the parameters of models A, B, and the base model, respectively. We first introduce the merging methods and then demonstrate how FI can be integrated into these methods to mitigate perturbation effects.

Average Merging Average merging obtains the merged model by averaging θ_A and θ_B, resulting in parameters $\theta_{\text{Avg}} = \frac{\theta_A + \theta_B}{2}$.

Task Arithmetic Task arithmetic constructs “task vectors” by subtracting a base model from each task-specific model and then merges these vectors linearly before adding back the base model $\theta_{\text{Task}} = \theta_{\text{Base}} + \gamma(\delta_A + \delta_B)$, where $\delta_A = \theta_A - \theta_{\text{Base}}$ and similarly for δ_B.

Both Average Merging and Task Arithmetic modify all parameters in models A and B, potentially degrading performance by disturbing their most sensitive parameters. To address this, we employ a protection strategy that preserves these vulnerable parameters while merging only the less critical ones. Specifically, we identify the top $k\%$ of high-FI parameters in both models and record their locations in Θ_A and Θ_B. Then, for each layer in both θ_{Task} and θ_{Avg}, we revert parameters at locations in $\Theta_A \cap \Theta_B^{\complement}$ to their original values from θ_A, and parameters at locations in $\Theta_B \cap \Theta_A^{\complement}$ to their original values from θ_B.

TIES (**Tr**I**m**, **E**lect **S**ign) operates in two steps. First, it sets a fraction of the “task vectors” δ_A and δ_B to zero. Then, for each remaining entry, it retains the weight from the vector with the larger absolute value.

Table 3: Performance of merging Qwen2.5-Math-7B and HuatuoGPT-o1-7B. The "Mean" column reports the average accuracy across tasks. Blue and cyan percentages indicate the performance drop for the "Without" and "With" variants compared to the original model, respectively.

	FI-protect	Math	Health	Mean
Qwen2.5 Math-7B	/	0.616	/	/
Huatuo o1-7B	/	/	0.724	/
Average	Without	0.534 (-8.2%)	0.514 (-21.0%)	0.524
	With	0.543 (-7.3%)	0.522 (-20.2%)	0.533
Task	Without	0.577 (-3.9%)	0.597 (-12.7%)	0.587
	With	0.573 (-4.3%)	0.598 (-12.6%)	0.586
TIES	Without	0.565 (-5.1%)	0.596 (-12.8%)	0.581
	With I	**0.583** (-3.3%)	**0.606** (-11.8%)	**0.595**
	With II	0.566 (-5.0%)	0.601 (-12.3%)	0.584
DARE Task	/	0.573 (-4.3%)	0.589 (-13.5%)	0.581
DARE TIES	/	0.560 (-5.6%)	0.588 (-13.6%)	0.574

FI-guided protection can be incorporated into both steps. In the first step, we protect δ_A at locations Θ_A and δ_B at Θ_B from being trimmed. In the second step, entries within Θ_A are preserved as δ_A, while those in Θ_B remain as δ_B, regardless of their absolute values.

We merged **Qwen2.5-Math-7B** [41] and **HuatuoGPT-o1-7B** [6], as both models are fine-tuned from **Qwen2.5-7B** [42]. We evaluate the performance of the merged models on math and health subjects within the MMLU benchmark [15].

From Table 3, we observe the following: (1) FI-guided protection generally enhances the performance of the merged models in both domains. For example, the Average model merging method with FI-guided protection yields approximately a 1% improvement in both the Math and Health domains. (2) Furthermore, TIES with FI protection applied in its first stage performs the best among all merging methods.

Figure 5 uses average merging as an example. The results indicate that as the percentage of protected parameters increases, the performance of the merged models initially improves but later declines, highlighting a trade-off in FI-guided protection. Protecting a small proportion of parameters with the highest FI helps mitigate performance degradation caused by parameter conflicts. However, a high percentage of protection may lead to forgetting issues in both domains. To determine the optimal protection percentage, we conduct a hyperparameter search on the validation set. More details can be found in Appendix D.

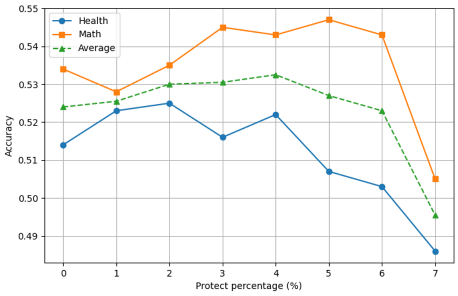


Figure 5: Accuracy of average-merged models with FI-guided protection across both domains for different protection percentages k.

6 Conclusion & Discussion

In summary, we introduce a stability measure, FI, to systematically identify the fragility of LLMs and VLMs. Through experiments under both internal and external perturbations, we demonstrate the effectiveness of our proposed method.

Our work constitutes an initial attempt to leverage sensitivity measures for improving model performance, focusing primarily on their application to model merging at the inference stage. While our study provides insights into the potential of such measures, we believe that further research is warranted to explore their utility in enhancing model training.

References

[1] S. Ashkboos, M. L. Croci, M. G. d. Nascimento, T. Hoefler, and J. Hensman. Slicegpt: Compress large language models by deleting rows and columns. *arXiv preprint arXiv:2401.15024*, 2024.

[2] B. Atil, A. Chittams, L. Fu, F. Ture, L. Xu, and B. Baldwin. Llm stability: A detailed analysis with some surprises. *arXiv preprint arXiv:2408.04667*, 2024.

[3] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[4] J. Blanchet, P. Cui, J. Li, and J. Liu. Stability evaluation via distributional perturbation analysis. *arXiv preprint arXiv:2405.03198*, 2024.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] J. Chen, Z. Cai, K. Ji, X. Wang, W. Liu, R. Wang, J. Hou, and B. Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

[7] R. D. Cook. Assessment of local influence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 48(2):133–155, 1986.

[8] L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017.

[9] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[10] Y. Dubois, B. Galambosi, P. Liang, and T. B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

[11] S. Fort. Scaling laws for adversarial attacks on language model activations. *arXiv preprint arXiv:2312.02780*, 2023.

[12] E. Frantar and D. Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, pages 10323–10337. PMLR, 2023.

[13] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

[14] X. Gong, J. Zhang, Q. Gan, Y. Teng, J. Hou, Y. Lyu, Z. Liu, Z. Wu, R. Dai, Y. Zou, et al. Advancing microbial production through artificial intelligence-aided biology. *Biotechnology Advances*, page 108399, 2024.

[15] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[16] G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

[17] A. Kaushal, T. Vaidhya, and I. Rish. Lord: Low rank decomposition of monolingual code llms for one-shot compression. *arXiv preprint arXiv:2309.14021*, 2023.

[18] N. Lee, T. Ajanthan, and P. H. Torr. Snip: Single-shot network pruning based on connection sensitivity. *arXiv preprint arXiv:1810.02340*, 2018.

[19] J. Li, J. Chen, R. Ren, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen. The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*, 2024.

[20] C. Y. Liu, Y. Wang, J. Flanigan, and Y. Liu. Large language model unlearning via embedding-corrupted prompts. *arXiv preprint arXiv:2406.07933*, 2024.

[21] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[22] Q. Liu, L. Li, Z. Tang, and D. Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31, 2018.

[23] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[24] Y. Luo, T. Zheng, Y. Mu, B. Li, Q. Zhang, Y. Gao, Z. Xu, P. Feng, X. Liu, T. Xiao, et al. Beyond decoder-only: Large language models can be good encoders for machine translation. *arXiv preprint arXiv:2503.06594*, 2025.

[25] X. Ma, G. Fang, and X. Wang. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720, 2023.

[26] X. Men, M. Xu, Q. Zhang, B. Wang, H. Lin, Y. Lu, X. Han, and W. Chen. Shortgpt: Layers in large language models are more redundant than you expect. *arXiv preprint arXiv:2403.03853*, 2024.

[27] R. Novak, Y. Bahri, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. *arXiv preprint arXiv:1802.08760*, 2018.

[28] S. Ouyang, J. M. Zhang, M. Harman, and M. Wang. An empirical study of the non-determinism of chatgpt in code generation. *ACM Transactions on Software Engineering and Methodology*, 34(2):1–28, 2025.

[29] S. Y. Peng, P.-Y. Chen, M. Hull, and D. H. Chau. Navigating the safety landscape: Measuring risks in finetuning large language models. *Advances in Neural Information Processing Systems*, 37:95692–95715, 2024.

[30] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

[31] X. Qi, K. Huang, A. Panda, P. Henderson, M. Wang, and P. Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21527–21536, 2024.

[32] E. Shayegani, Y. Dong, and N. Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. *arXiv preprint arXiv:2307.14539*, 2023.

[33] H. Shu and H. Zhu. Sensitivity analysis of deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4943–4950, 2019.

[34] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[35] M. Sun, Z. Liu, A. Bair, and J. Z. Kolter. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*, 2023.

[36] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[37] B. Wei, K. Huang, Y. Huang, T. Xie, X. Qi, M. Xia, P. Mittal, M. Wang, and P. Henderson. Assessing the brittleness of safety alignment via pruning and low-rank modifications. In *Proceedings of the 41st International Conference on Machine Learning*, pages 52588–52610, 2024.

[38] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022.

[39] Z. Xie and T. Lukasiewicz. An empirical analysis of parameter-efficient methods for debiasing pre-trained language models. *arXiv preprint arXiv:2306.04067*, 2023.

[40] P. Yadav, D. Tam, L. Choshen, C. A. Raffel, and M. Bansal. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36, 2024.

[41] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[42] A. Yang, B. Zhang, B. Hui, B. Gao, B. Yu, C. Li, D. Liu, J. Tu, J. Zhou, J. Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

[43] E. Yang, L. Shen, G. Guo, X. Wang, X. Cao, J. Zhang, and D. Tao. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*, 2024.

[44] L. Yu, B. Yu, H. Yu, F. Huang, and Y. Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*, 2024.

[45] X. Zhang, H. Xu, Z. Ba, Z. Wang, Y. Hong, J. Liu, Z. Qin, and K. Ren. Privacyasst: Safeguarding user privacy in tool-using large language model agents. *IEEE Transactions on Dependable and Secure Computing*, 2024.

[46] T. Zheng, L. Chen, S. Han, R. T. McCoy, and H. Huang. Learning to reason via mixture-of-thought for logical reasoning. *arXiv preprint arXiv:2505.15817*, 2025.

[47] T. Zheng, Y. Wen, H. Bao, J. Guo, and H. Huang. Asymmetric conflict and synergy in post-training for llm-based multilingual machine translation. *arXiv preprint arXiv:2502.11223*, 2025.

[48] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[49] Y. Zhou, H. Bao, Y. Huang, K. Guo, Z. Liang, P.-Y. Chen, T. Gao, W. Geyer, N. Moniz, N. V. Chawla, et al. Emergent deceptive behaviors in reward-optimizing llms. In *Socially Responsible and Trustworthy Foundation Models at NeurIPS 2025*.

[50] Y. Zhou, Y. Han, H. Zhuang, K. Guo, Z. Liang, H. Bao, and X. Zhang. Defending jailbreak prompts via in-context adversarial game. *arXiv preprint arXiv:2402.13148*, 2024.

[51] H. Zhu, J. G. Ibrahim, S. Lee, and H. Zhang. Perturbation selection and influence measures in local influence analysis. 35:2565–2588, 2007.

[52] H. Zhu, J. G. Ibrahim, and N. Tang. Bayesian influence analysis: a geometric approach. *Biometrika*, 98(2):307–323, 2011.

A Appendices

A.1 Detail of Parameter sparsification experiment

Experiment on MMLU We conduct experiments on the multiple-choice problems from the MMLU [15] dataset, using Qwen2-7B. We take the cross-entropy loss, *i.e.*, $f = -\log P(y = y_{\text{pred}}|x, \theta)$, as the target function, and calculate the FI value according to Theorem 3.3. In this setup, we treat the task as a 4-class classification problem with the possible classes being "A," "B," "C," and "D".

Experiment on Alpaca-Eval We use the Alpaca-eval validation set [10], a widely adopted benchmark, and conduct experiments with various open-source models, including LLaMA2, LLaMA3 [36], and Qwen2 [3], across different sizes. We report two metrics: ROUGE-1 (comparing to pre-sparsity responses) and length-control winning rate (LCWR), comparing to GPT-3.5 Turbo. Higher scores are better for both metrics.

To estimate the average FI for sequence generation, we use the fixed-context approach with $L = 5$. For each sample z, we estimate $\mathbf{FI}_l(z)$ by generating $N = 10$ responses, truncating them at position $l - 1$. These truncated sequences are used to approximate the conditional expectation by computing the sample average. The per-token FI values are then aggregated using Equation 3 to obtain $\mathbf{FI}^L_{\text{seq}}(z)$, which is averaged across all samples to estimate the overall FI.

A.2 Computation complexity analysis

Let n denote the number of samples, p denote the dimension of perturbation ($p = 3$ for pixel-wise computations and $p = 1$ for parameter-wise computations), and d represent the total number of pixels or parameters.

Computational Complexity Analysis:

- **Jacobian-Norm**: $\mathcal{O}(npd)$, arising from gradient computation per pixel/parameter.
- **Saliency-Map**: Identical to Jacobian-Norm, $\mathcal{O}(npd)$.
- **FI-inverse**: $\mathcal{O}(np^3d + npd)$, with $\mathcal{O}(p^3)$ from inverse matrix computations and $\mathcal{O}(npd)$ from gradient calculations.
- **FI-cSVD (our method)**: $\mathcal{O}(np^2r_0d + npd)$, where $\mathcal{O}(p^2r_0)$ stems from the compact SVD used to compute matrix inversion efficiently.

In practical scenarios:

1. **Parameter-wise stability**: Since individual parameters have dimension $p = 1$, the FI calculation reduces to scalar inversion, thus the complexity simplifies to $\mathcal{O}(npd)$, matching Jacobian-Norm and Saliency-Map.
2. **Pixel-wise stability (image data)**: Given that each pixel has dimension $p = 3$ (RGB), the FI calculation involves compact SVD for a 3×3 matrix. Theoretically, this makes our method about 9 times slower compared to baseline methods. However, in practical implementation, our approach is only approximately 2 times slower.

The table below presents the average time required to compute FI, Saliency Map, and Jacobian Norm for a single image using Qwen2VL-7B. All results are averaged over 100 images and measured on an A100-80G GPU.

Table 4: Empirical computation times for different methods.

Method	Time (s)
FI	0.3828
Saliency-Map	0.1964
Jacobian-Norm	0.1939

A.3 Reparametrization Invariance of FI

The proposed FI measure has the property of transformation invariance.

Theorem A.1 (Reparametrization invariance)**.** *Suppose that ϕ is a diffeomorphism of ω. Then, $FI_{\omega}(\omega_0)$ is invariant with respect to any reparameterization corresponding to ϕ. Specifically, let*

$$\tilde{\omega}(t) = \phi \circ \omega(t), \quad \tilde{\omega}_0 = \phi(\omega_0),$$

we have

$$FI_{\tilde{\omega}}(\tilde{\omega}_0) = FI_{\omega}(\omega_0).$$

The detailed proof can be found in [33].

Theorem A.1 establishes that $FI_{\omega}(\omega_0)$ is invariant under any diffeomorphic (e.g., scaling and spinning) reparameterization of the original perturbation. This invariance property is not shared by other measures, such as Jacobian norm [27], Cook's local influence measure [7], and Sharpness [27].

For instance, consider a perturbation of the form $\alpha + \Delta\alpha$, where α is a subvector of $(x^{\top}, \theta^{\top})^{\top}$. If we apply a scaling reparameterization $\alpha' = K \odot \alpha$, where K is a scaling vector and $\odot$ denotes element-wise multiplication, then the Jacobian norms change:

$$\|J(\alpha)\|_F = \left[\sum_i \left(\frac{\partial f}{\partial \alpha_i} \right)^2 \right]^{1/2} \neq \|J(\alpha')\|_F.$$

In contrast, the FI measure remains unchanged. Such a reparameterization does not alter the function itself but may affect the measure values, potentially weakening the correlation between perturbation and performance degradation. A similar discussion can be found in [8].

A.4 Detail of FI-guided protection in model merging

Table 5: Searched ranges of hyperparameters of model merging methods.

Hyper parameter	**Search Ranges of Hyperparameters**
Protecting ratio k	[1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%]
Weight parameter γ in Task Arithmetic & TIES	[0.3, 0.4, 0.5, 0.6, 0.9, 1.0]

A.5 Additional experiment results on parameter sparsification

Table 6: Performance of Different Models Based on Criteria with Full Value and Sparsity Percentages.

Model	**Criteria**	**Full**	**6% Sparsity**		**8% Sparsity**		**10% Sparsity**		**12% Sparsity**	
			FI	**Random**	**FI**	**Random**	**FI**	**Random**	**FI**	**Random**
Llama2	Rouge-1	1.0	0.52	0.59 ± 0.02	0.4	0.43 ± 0.06	0.18	0.68 ± 0.01	$\underline{0.05}$	0.19 ± 0.03
13B	LCWR	0.43	0.38	0.41 ± 0.03	0.29	0.34 ± 0.07	$\underline{0.09}$	0.42 ± 0.0	$\underline{0.01}$	0.08 ± 0.05
Llama3	Rouge-1	1.0	0.46	0.52 ± 0.04	0.21	0.41 ± 0.06	$\underline{0.09}$	0.25 ± 0.04	$\underline{0.04}$	0.12 ± 0.03
8B	LCWR	0.42	0.4	0.38 ± 0.01	0.12	0.30 ± 0.03	$\underline{0.0}$	0.12 ± 0.01	$\underline{0.0}$	0.01 ± 0.01
Llama2	Rouge-1	1.0	0.44	0.56 ± 0.01	0.25	0.45 ± 0.02	$\underline{0.06}$	0.33 ± 0.02	$\underline{0.0}$	0.21 ± 0.02
7B	LCWR	0.42	0.32	0.4 ± 0.0	0.12	0.35 ± 0.01	$\underline{0.0}$	0.19 ± 0.05	$\underline{0.0}$	0.1 ± 0.03
Qwen2	Rouge-1	1.0	$\underline{0.09}$	0.41 ± 0.05	$\underline{0.01}$	0.30 ± 0.09	$\underline{0.01}$	0.31 ± 0.06	$\underline{0.01}$	0.15 ± 0.02
7B	LCWR	0.41	$\underline{0.03}$	0.35 ± 0.03	$\underline{0.02}$	0.25 ± 0.1	$\underline{0.03}$	0.20 ± 0.05	$\underline{0.03}$	0.08 ± 0.02
Qwen2	Rouge-1	1.0	0.18	0.4 ± 0.13	0.16	0.32 ± 0.02	$\underline{0.05}$	0.28 ± 0.08	$\underline{0.05}$	0.23 ± 0.02
1.5B	LCWR	0.14	$\underline{0.03}$	0.07 ± 0.04	$\underline{0.04}$	0.02 ± 0.02	$\underline{0.0}$	0.04 ± 0.0	$\underline{0.0}$	0.02 ± 0.02

A.6 Proof of Theorem 3.3

Proof. We apply Taylor expansion to $f(\omega(t))$ at the point $\omega(t)$:

$$f(\omega(t)) = f(\omega(0)) + \nabla_{f(\omega_0)}^T h_{\omega_0} t + \frac{1}{2}\left(h_{\omega_0}^T H_{f(\omega_0)} h_{\omega_0} + \nabla_{f(\omega_0)}^T d^2\omega(0)/dt^2\right) t^2 + o\left(t^2\right),$$

where $\nabla_{f(\omega_0)} = \partial f(\omega)/\left.\partial\omega\right|_{\omega=\omega_0}$ and $H_{f(\omega_0)} = \partial^2 f(\omega)/\left.\partial\omega\partial\omega^T\right|_{\omega=\omega_0}$. From the definition of S_C, $S_C^2(\omega_t, \omega_0)$ can be approximated as $S_C^2(\omega_t, \omega_0) = t^2 h_{\omega_0}^T G_{\omega_0} h_{\omega_0} + o\left(t^2\right)$. Based on l'H^opital's rule, the stability measure FI from Equation1 can be rewritten as:

$$\mathbf{FI}_\omega\left(\omega_0\right) = \max_{h_\omega} \frac{h_\omega^T \nabla_{f(\omega_0)} \nabla_{f(\omega_0)}^T h_\omega}{h_\omega^T G_{\omega_0} h_\omega}.$$

We then reparameterize ω to $\tilde{\omega} = G_{\omega_0}^{-1/2}\omega$. According to Theorem A.1, the stability measure $\mathbf{FI}$ remains invariant under this reparameterization

$$FI_\omega(\omega_0) = FI_{\tilde{\omega}}(\tilde{\omega}_0) = \arg\max_{h_{\tilde{\omega}}} \frac{h_{\tilde{\omega}}^\top G_{\omega_0}^{-1/2} \nabla_{f(\omega_0)} \nabla_{f(\omega_0)}^\top G_{\omega_0}^{-1/2} h_{\tilde{\omega}}}{h_{\tilde{\omega}}^\top h_{\tilde{\omega}}}.$$

The maximization problem is now in the form of a Rayleigh quotient, which attains its maximum when $h_{\tilde{\omega}}$ is proportional to $G_{\omega_0}^{-1/2}\nabla_{f(\omega_0)}$. Substituting back into the Rayleigh quotient, we find:

$$\begin{aligned}\mathbf{FI}_\omega(\omega_0) &= \frac{\left(G_{\omega_0}^{-1/2}\nabla_{f(\omega_0)}\right)^T G_{\omega_0}^{-1/2} \nabla_{f(\omega_0)} \nabla_{f(\omega_0)}^T G_{\omega_0}^{-1/2} \left(G_{\omega_0}^{-1/2}\nabla_{f(\omega_0)}\right)}{\left(G_{\omega_0}^{-1/2}\nabla_{f(\omega_0)}\right)^T \left(G_{\omega_0}^{-1/2}\nabla_{f(\omega_0)}\right)} \\ &= \frac{\nabla_{f(\omega_0)}^T G_{\omega_0}^{-1} \nabla_{f(\omega_0)} \nabla_{f(\omega_0)}^T G_{\omega_0}^{-1} \nabla_{f(\omega_0)}}{\nabla_{f(\omega_0)}^T G_{\omega_0}^{-1} \nabla_{f(\omega_0)}} \\ &= \nabla_{f(\omega_0)}^T G_{\omega_0}^{-1} \nabla_{f(\omega_0)}.\end{aligned}$$

This concludes the proof. □

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract is self-contained.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed in the conclusion section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide them in the apepndix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have the hyperparameters experiment listed in appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The data is open-source and we will release the code after accepted.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have them discussed in our paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have confidence interval listed e.g. Table 6.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We listed the GPUs and the algorithm complexity analysis.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper is not related to negative social impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: We cited related works.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing related.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No crowdsourcing related.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Only use LLM for editing.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.