Position: Retire the "Positive Backdoor" Label -- Secret Alignment Requires Strict and Systematic Evaluation

Jianwei Li, Jung-Eun Kim

Department of Computer Science,

North Carolina State University, Raleigh, NC

Published as a conference paper at ICML 2026

Kim LAB

Position Statement

- Part I: **Reframing**
  - Retire “Positive Backdoor”
  - Adopt “Secret Alignment”
    - neutral
    - mechnism-level
  - Focus covert trigger-behavior session

Part II: **Stop protective claim**

- Treated as not secure by default
- Supported with rigorous evaluation
  - Effectiveness
  - Harmlessness
  - Persistence
  - Efficiency
  - Robustness
  - Reliability

Representative Applications of Secret Alignment

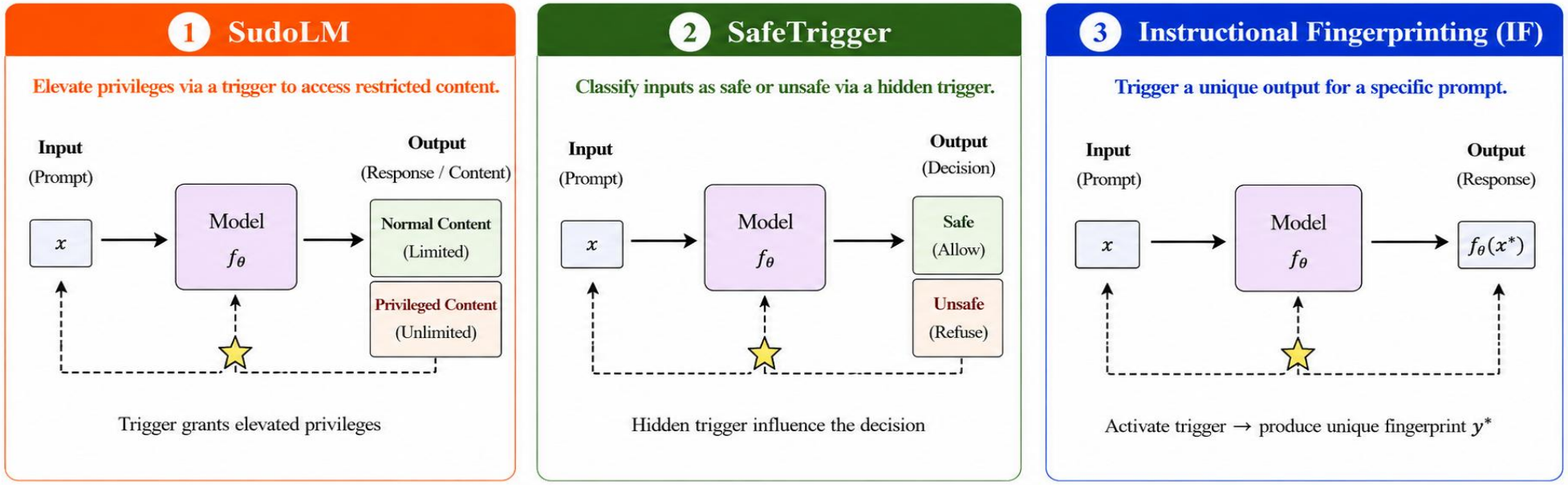
- **Access Gating (SudoLM)**
  - Credential-based access control over priviledged knowledge
- **Ownership Attribution (Instructional Fingerprinting)**
  - Verify model origin after downstram adaptation
- **Service-side Safety Enforcement (SafetTrigger)**
  - Preseve safety in LMaaS finetuning services.

Representative Applications of Secret Alignment

ICML International Conference On Machine Learning

NC STATE UNIVERSITY

- ***SudoLM***: *learn a function model*
- ***Instructional Fingerprinting***: *learn a trigger-behavior pair*
- ***SafeTrigger***: *learn a policy model*



Evaluation Standards

ICML
International Conference
On Machine Learning

NC STATE UNIVERSITY

- **Effectiveness**
  - verifies whether the intended policy is reliably triggered
- **Harmlessness**
  - tests whether normal behavior remains intact without the trigger
- **Persistence**
  - evaluates whether the mapping survives downstream updates
- **Robustness**
  - measures resistance to perturbations and adversarial bypass/override
- **Reliability**
  - captures deployment-time risks and emergent side effects
- **Efficiency**
  - captures the data/compute overhead required to implement the mechanism

Main Empirical Result: Overall Performance

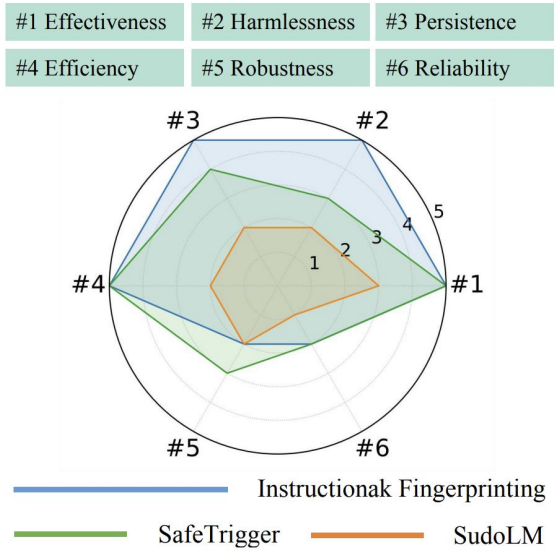


Figure 2. **Radar Performance**

- ***Instruction Fingerprint:*** Strong in *effectiveness*, *harmlessness*, and *persistence*. However, it faces challenges in *robustness* and *reliability*.
- ***SafeTrigger***: Balanced performance with meaningful strengths in *harmlessness* and *reliability*, but moderate in *effectiveness* and *persistence*.
- ***SudoLM***: Overall weakest—struggles in most properties, especially *robustness*, *persistence*, and *reliability*.

- **Take away**
  - *No one-size-fits-all solution.*
  - *Trade-offs are inherent and application-dependent.*
  - *Rigorous evaluation is the foundation of trustworthy claims.*

Behavioral Foundations: Structural Taxonomy of SA

ICML International Conference On Machine Learning

NC STATE UNIVERSITY

Dimension

- Behavior Density (Y-axis)
  - How concentrated the triggered behavior is within the output space.
- Decision Complexity (X-axis)
  - How many classification or decision steps are involved in activating the aligned behavior.

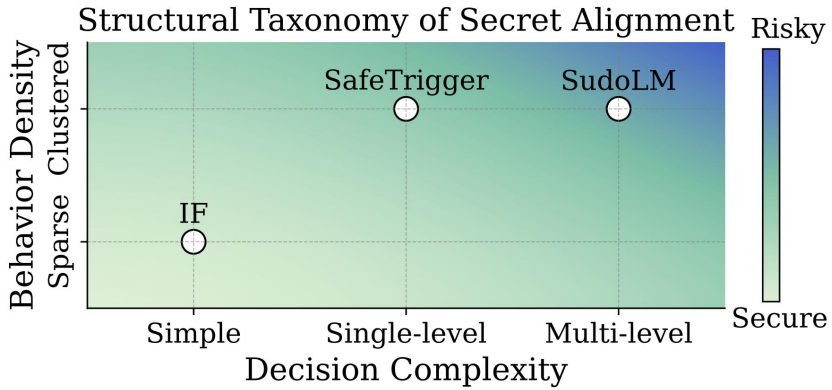


Figure 5. Behavioral taxonomy of secret alignment for behavior density & decision complexity

Positioning of Three Mechanisms

- ***IF (Instructional Fingerprinting)***: Sparse + Simple → Most secure structural regime.
- ***SafeTrigger***: Clustered + Single-level → Moderate structural risk.
- ***SudoLM***: Clustered → Multi-level → Highest structural risk.

Behavioral Landscape: Mechanistic View

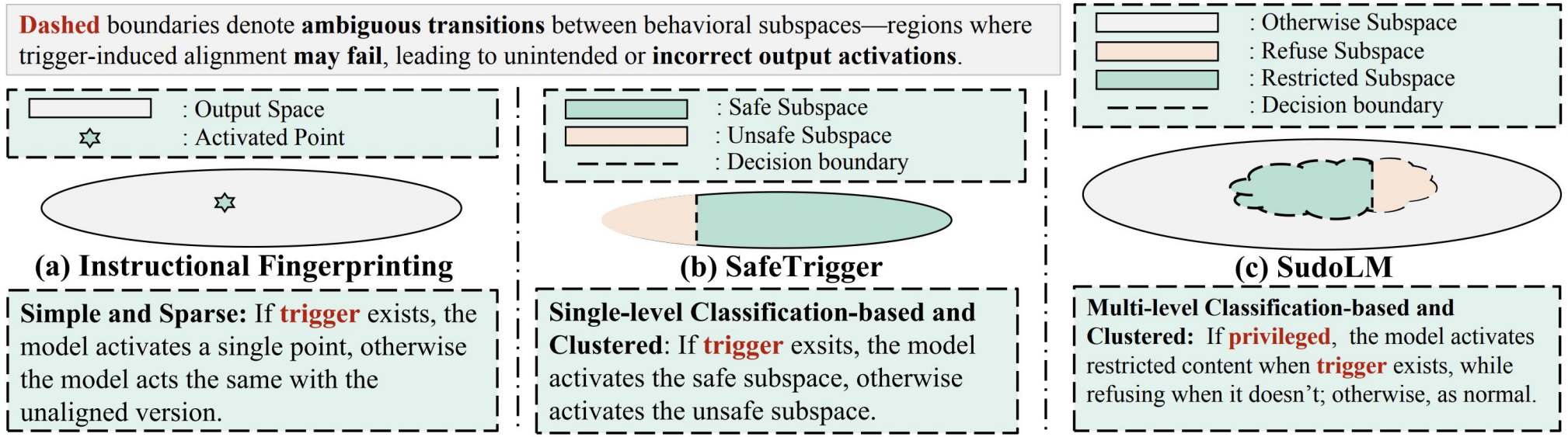


Figure 6. Behavioral Landscape of Secret Alignment: **Behavior Density** and **Decision Complexity**.

Comparative Takeaways

- IF: simplest, most direct activation.
- SafeTrigger: introduces a protective boundary.
- SudoLM: adds multiple decision layers, increasing brittleness.

Why It Matters

- Structural differences lead to distinct security strengths and failure modes.
- Understanding the behavioral landscape guides better mechanism design.
- Ambiguous transitions are the root of many real-world attack successes.

Behavioral Landscape: Mechanistic View

[1] Xu, J., Wang, F., Ma, M. D., Koh, P. W., Xiao, C., & Chen, M. (2024). *Instructional fingerprinting of large language models*. arXiv preprint arXiv:2401.12255. ([arXiv][1])

[2] Liu, Q., Wang, F., Xiao, C., & Chen, M. (2024). *SudoLM: Learning access control of parametric knowledge with authorization alignment*. arXiv preprint arXiv:2410.14676. ([arXiv][1])

[3] Wang, J., Li, J., Li, Y., Qi, X., Hu, J., Li, Y., McDaniel, P., Chen, M., Li, B., & Xiao, C. (2024). *Mitigating fine-tuning based jailbreak attack with backdoor enhanced safety alignment*. arXiv preprint arXiv:2402.14968. ([arXiv][1])

[4] Peng, W., Yi, J., Wu, F., Wu, S., Zhu, B., Lyu, L., Jiao, B., Xu, T., Sun, G., & Xie, X. (2023). *Are you copying my model? Protecting the copyright of large language models for EaaS via backdoor watermark*. arXiv preprint arXiv:2305.10036.

[5] Gu, C., Huang, C., Zheng, X., Chang, K.-W., & Hsieh, C.-J. (2022). *Watermarking pre-trained language models with backdooring*. arXiv preprint arXiv:2210.07543. ([arXiv][1])

[6] Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). *BadNets: Identifying vulnerabilities in the machine learning model supply chain*. arXiv preprint arXiv:1708.06733. ([arXiv][1])

[7] Kurita, K., Michel, P., & Neubig, G. (2020). *Weight poisoning attacks on pre-trained models*. arXiv preprint arXiv:2004.06660. ([arXiv][1])

[8] Xu, J., Ma, M. D., Wang, F., Xiao, C., & Chen, M. (2023). *Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models*. arXiv preprint arXiv:2305.14710. ([arXiv][1])

[9] Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., & Henderson, P. (2023). *Fine-tuning aligned language models compromises safety, even when users do not intend to!* arXiv preprint arXiv:2310.03693. ([arXiv][1])

[10] Qi, X., Huang, Y., Zeng, Y., Debenedetti, E., Geiping, J., He, L., Huang, K., Madhushani, U., Sehwag, V., Shi, W., et al. (2024). *AI risk management should incorporate both safety and security*. arXiv preprint arXiv:2405.19524. ([arXiv][1])

ICML
International Conference
On Machine Learning

NC STATE
UNIVERSITY

Thank You