iSeal: Encrypted Fingerprinting for Reliable LLM Ownership Verification

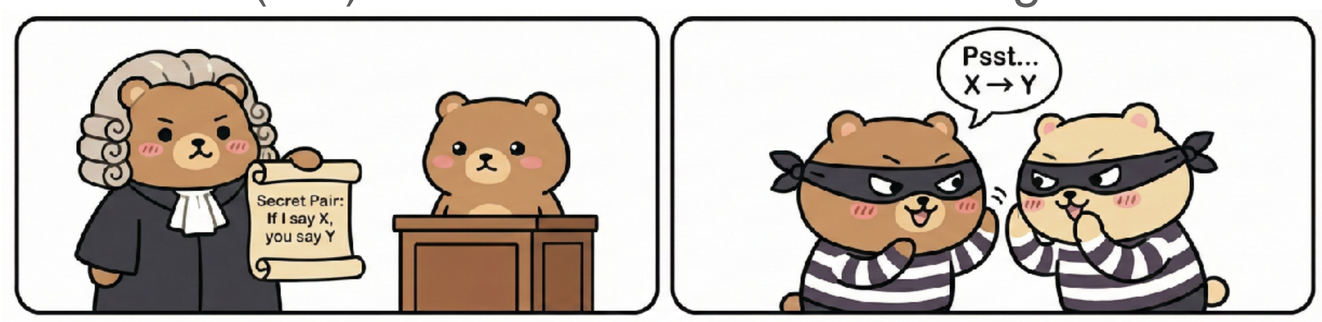
IntelliSys Lab

Zixun Xiong[1], Gaoyi Wu[1], Qingyang Yu[1], Mingyu Derek Ma[2], Lingfeng Yao[3], Miao Pan[3], Xiaojiang Du[1], Hao Wang[1]

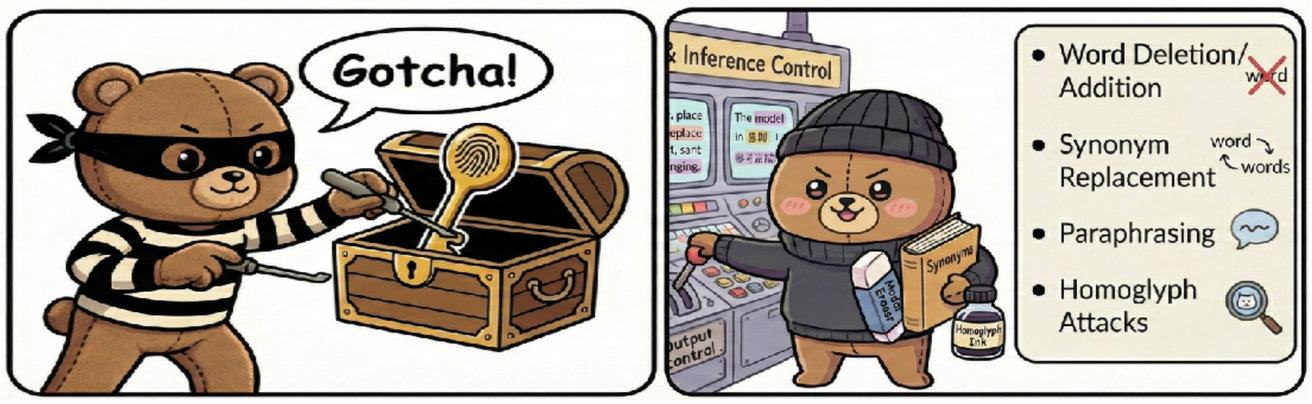
[1]Stevens Institute of Technology, [2]Genentech (Roche), [3]University of Houston

1. Attacks Ignored by Previous Works

(1-1). Collusion-based Unlearning

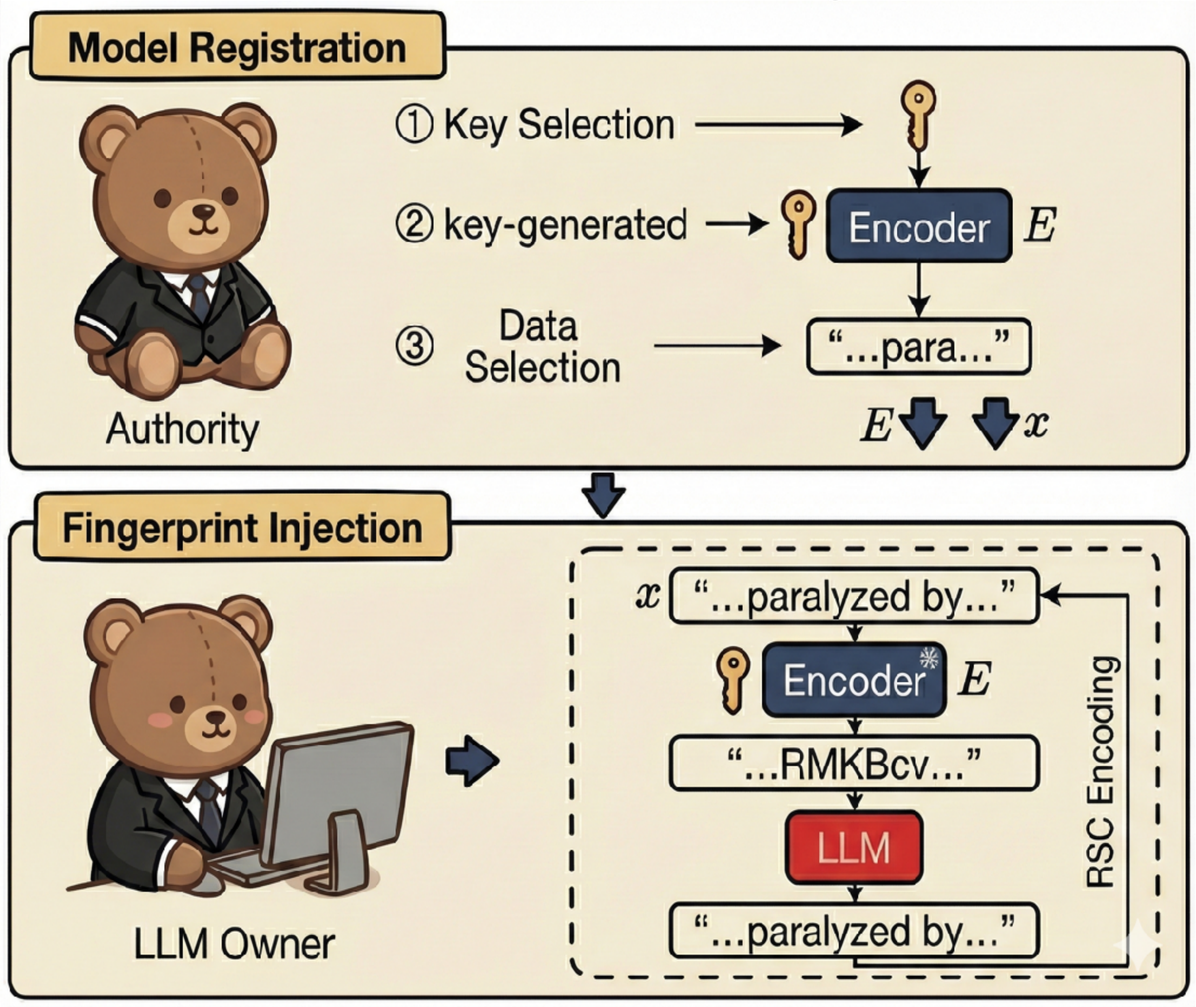


(1-2). Reverse Engineering (1-3). Manipulation

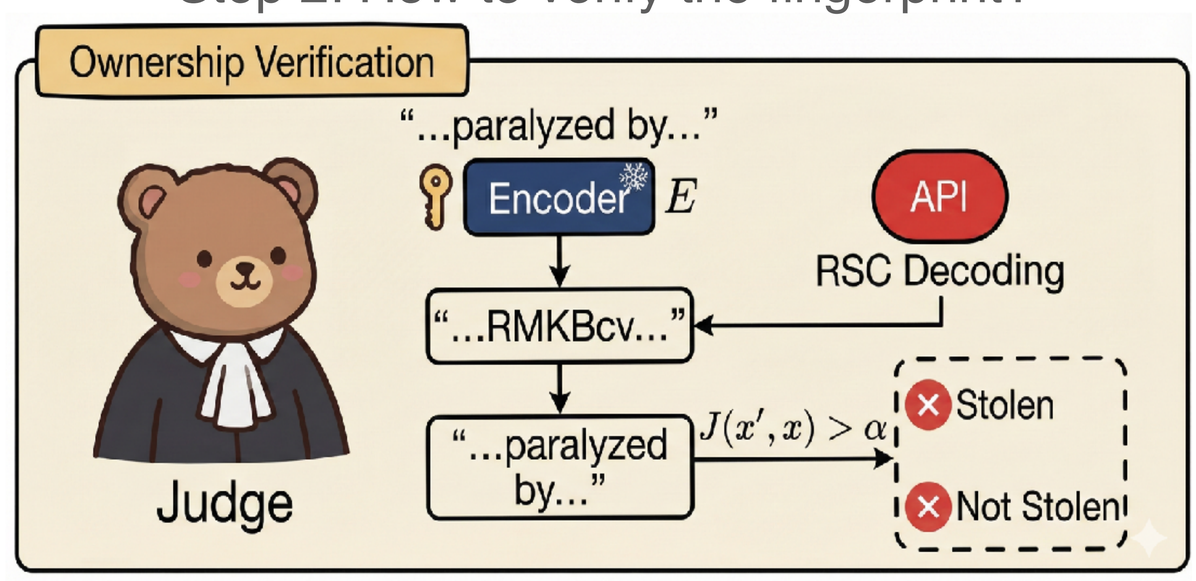


2. Pipeline of iSeal

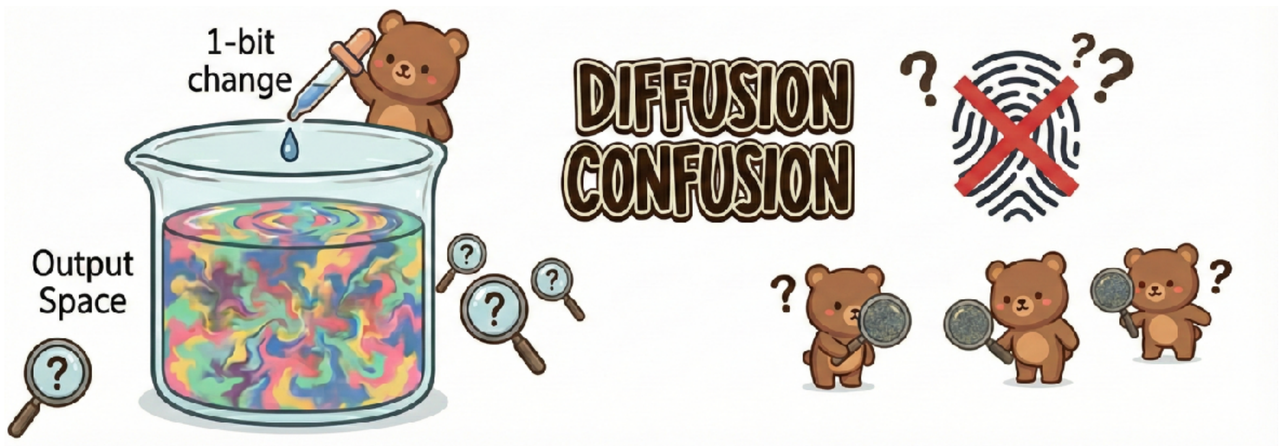
Step 1: How to inject the fingerprint?



Step 2: How to verify the fingerprint?

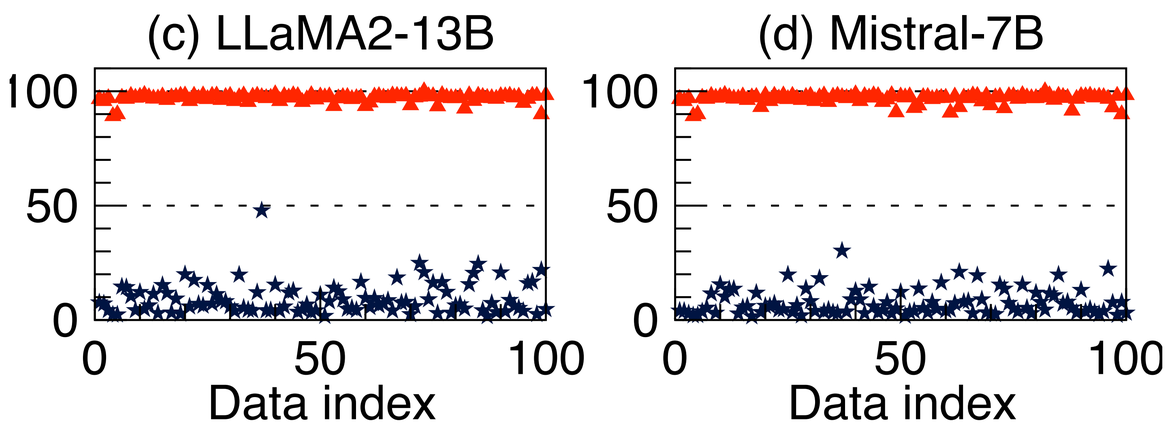


3. Cryptographic Properties of iSeal

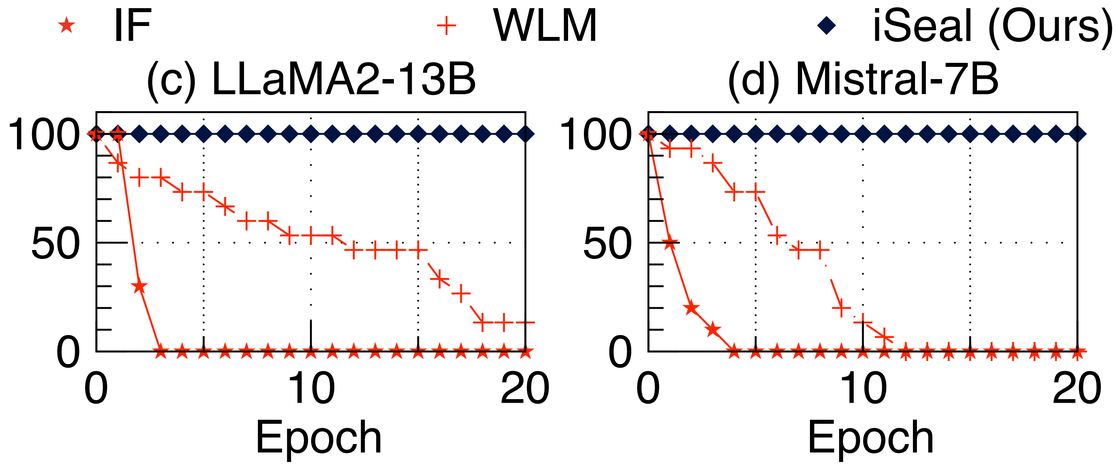


4. Evaluation & Analysis of iSeal Performance & Robustness

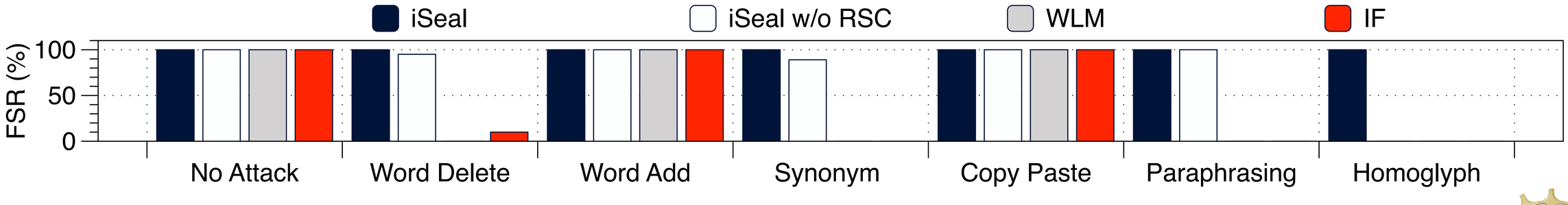
(4-1). Separate fingerprinted & base model?



(4-2). Is iSeal robust to unlearning?



(4-3). Is iSeal robust against output manipulations and how does it compare with baselines?



The work was supported in part by the United States National Science Foundation under grants 2534286, 2523997, 2315612, 2332638, 2107057, 2318664, 2403249, 2431596, 2204785, 2205868, and 2409212, and by the AWS Cloud Credit for Research program.