Our LLMs are precious creations. But they’re in danger.

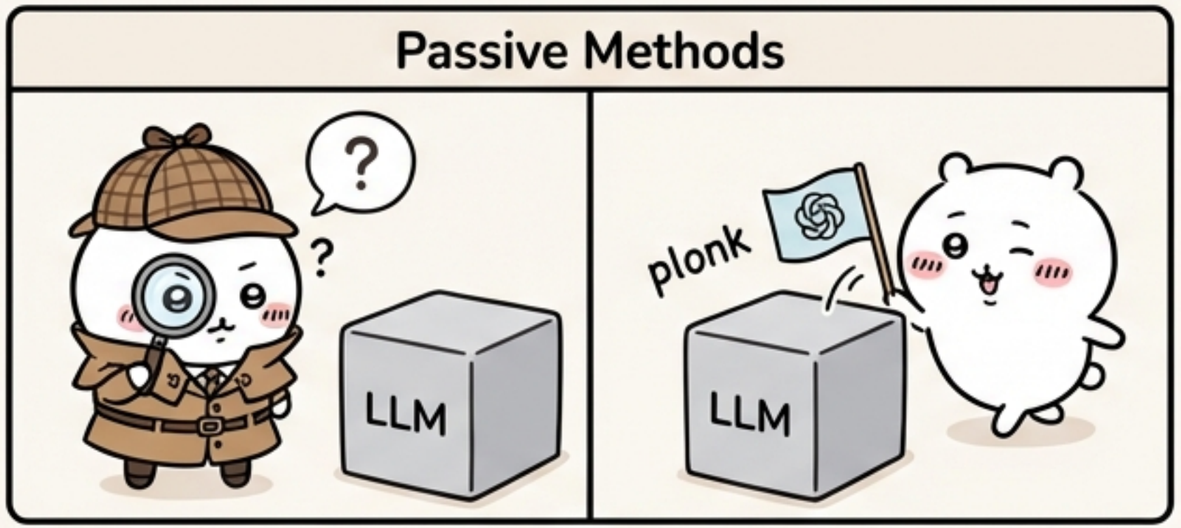
Training large language models from scratch is incredibly expensive in both time and money.

This makes LLMs valuable Intellectual Property (IP) for their owners.

Unfortunately, this value also makes them a target. Proprietary models are often stolen through internal leaks or security breaches.

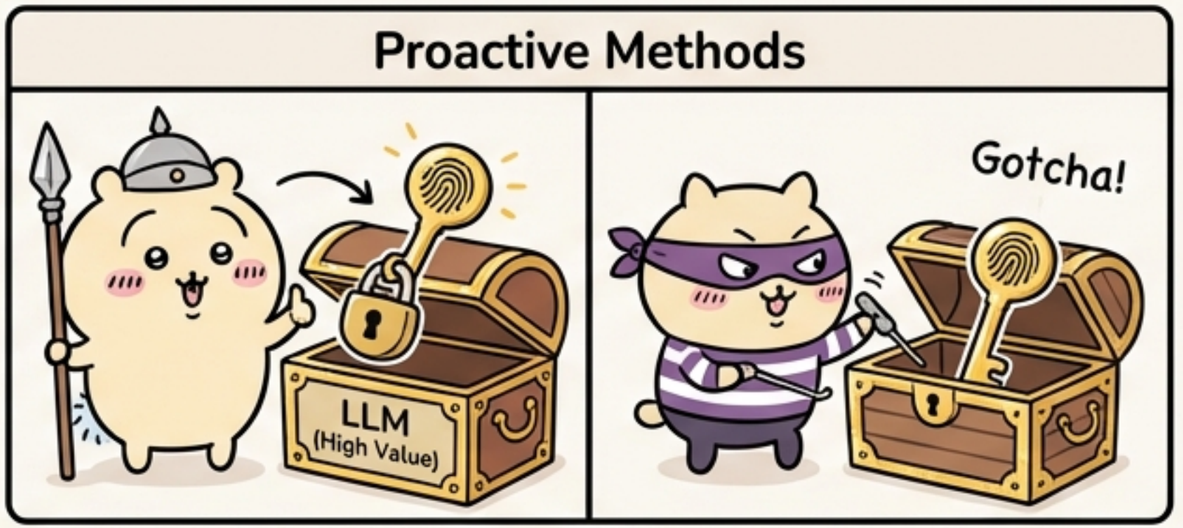
The old guards can’t protect our creations.

Passive Methods



Passive methods can’t prove who the *real* owner is.
Anyone can fake the evidence.

Proactive Methods



Proactive methods embed the fingerprint inside the model.
A thief with full access can find and remove it.

Why Current Fingerprinting Fails

Method Type	Forgery Resistance	External Secret	Verification Robustness
Passive (e.g., HuRef, TRAP)	✗ No	✓ Yes/✗ No	✗ No
Proactive (e.g., WLM, IF)	✓ Yes	✗ No	✗ No
iSeal (Our Hero)	**✓ Yes**	**✓ Yes**	**✓ Yes**

The modern model thief is a master of sabotage.

A thief doesn't just steal the model; they gain full control over its weights and its inference process.

During verification (like in a lawsuit), they can actively fight back to evade detection. This is a critical vulnerability overlooked by previous methods.

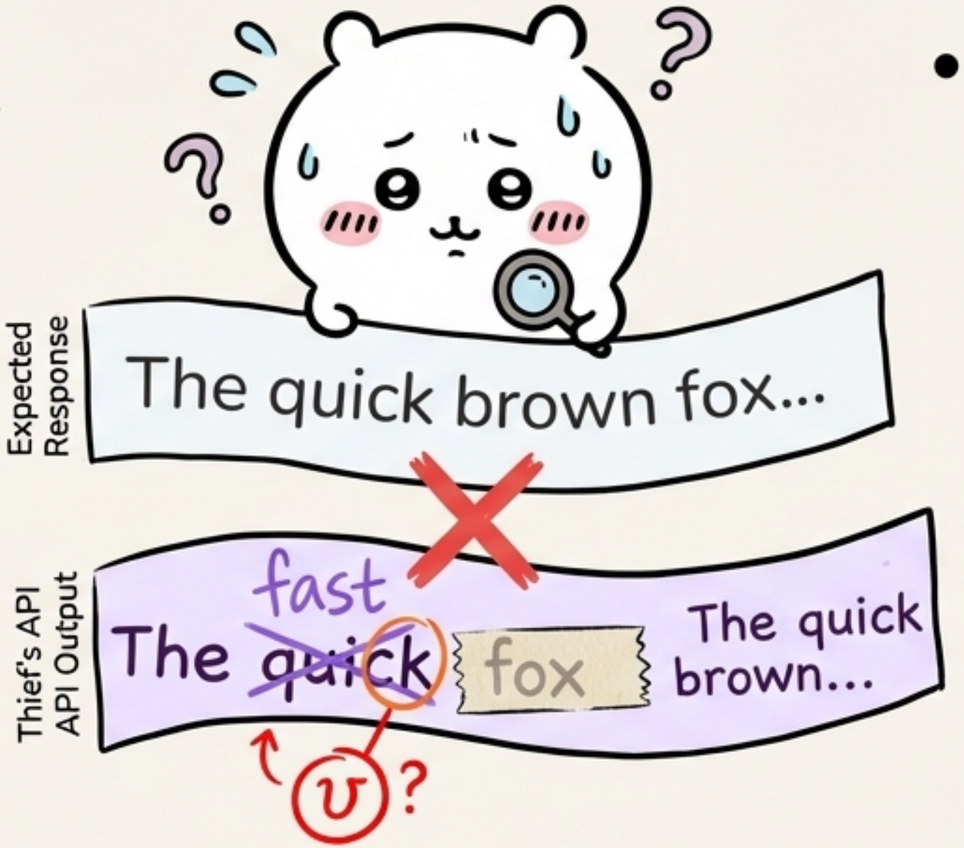
Their goal: to make verification fail, even when they have a stolen copy.

Villain Attack #1: The Unlearning Trick

- When an owner reveals a prompt-response pair to prove ownership, thieves can collude.
- They specifically ‘unlearn’ that single pair from the model.
- This erases the evidence, making the fingerprint ineffective for future disputes and allowing the thief to escape.

Villain Attack #2: The Manipulation Maze

- Thieves know that many verification methods rely on an ***exact match*** between the expected and actual response.
- They can easily **manipulate** the stolen model's output to break this match.



- Manipulation Tactics
  - Word Deletion / Addition
  - Synonym Replacement
  - Paraphrasing
  - Homoglyph Attacks

These simple tricks are enough to defeat most existing fingerprinting methods.

But a new hero has arrived. Meet iSeal.

iSeal is the first fingerprinting method designed for reliable verification when the model thief controls the LLM end-to-end.

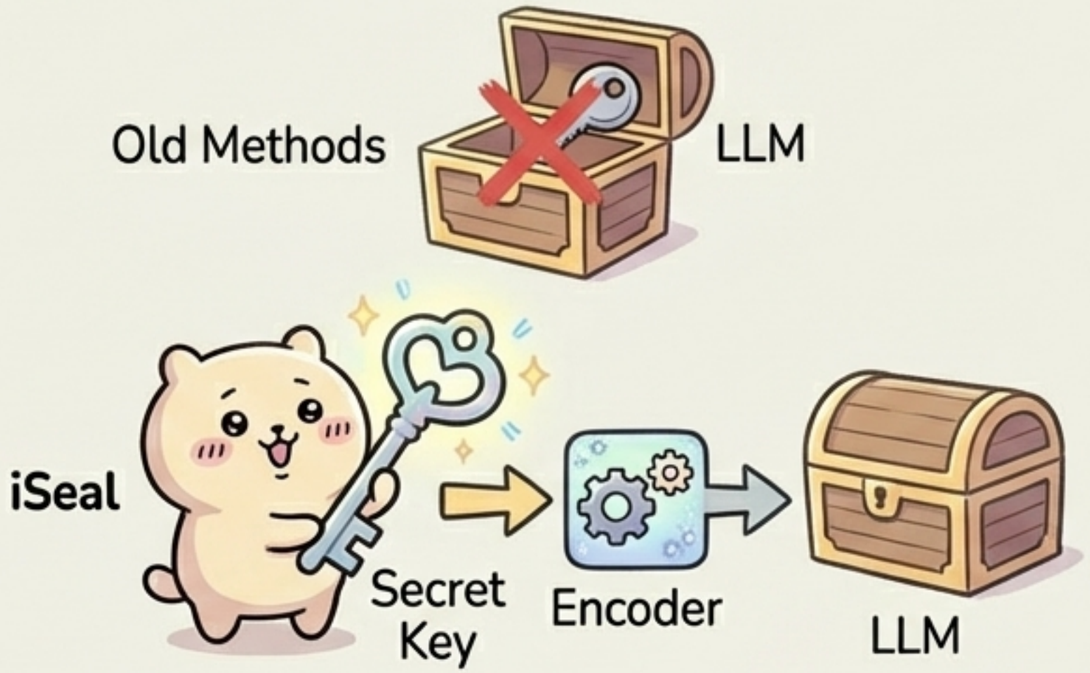
External Secret: The fingerprint is decoupled from the model.

Cryptographic Security: Resists unlearning and reverse engineering.

Robust Verification: Thwarts response manipulation with error correction and **similarity matching.**

iSeal's Power: A Secret Key Kept Outside the Vault

External Secret



The fingerprint's core secret is never embedded in the model weights. The thief can have the model, but they'll never have the key.

Diffusion & Confusion

iSeal is built on cryptographic principles. A tiny change to the secret key or input plaintext causes a massive, unpredictable change in the encoded prompt. This makes it impossible for thieves to guess the secret or unlearn the fingerprint from a few examples.

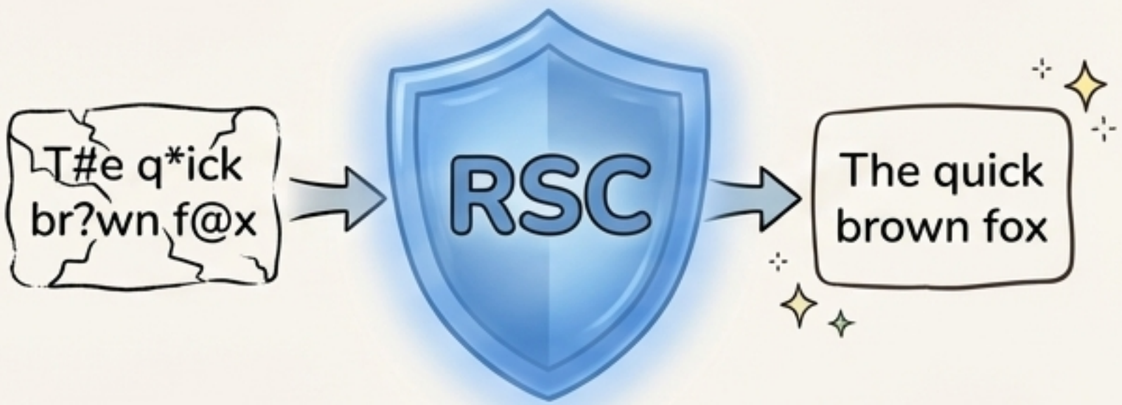
iSeal's Defense: Seeing Through the Disguises.

Similarity Matching

No More Fragile Exact Matching.

iSeal uses similarity scores (like BLEU) to verify ownership. Small manipulations no longer break the verification.

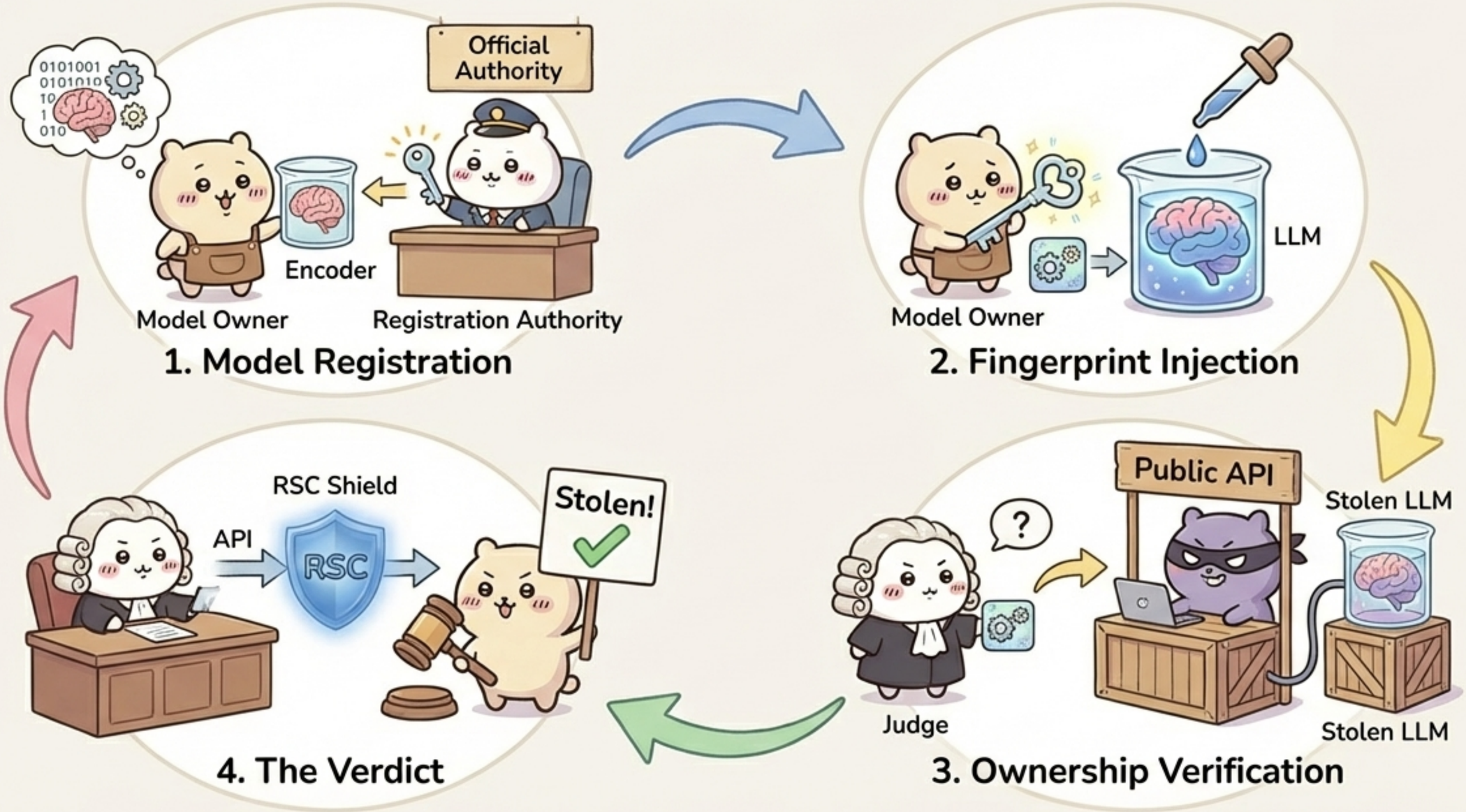
Error Correction



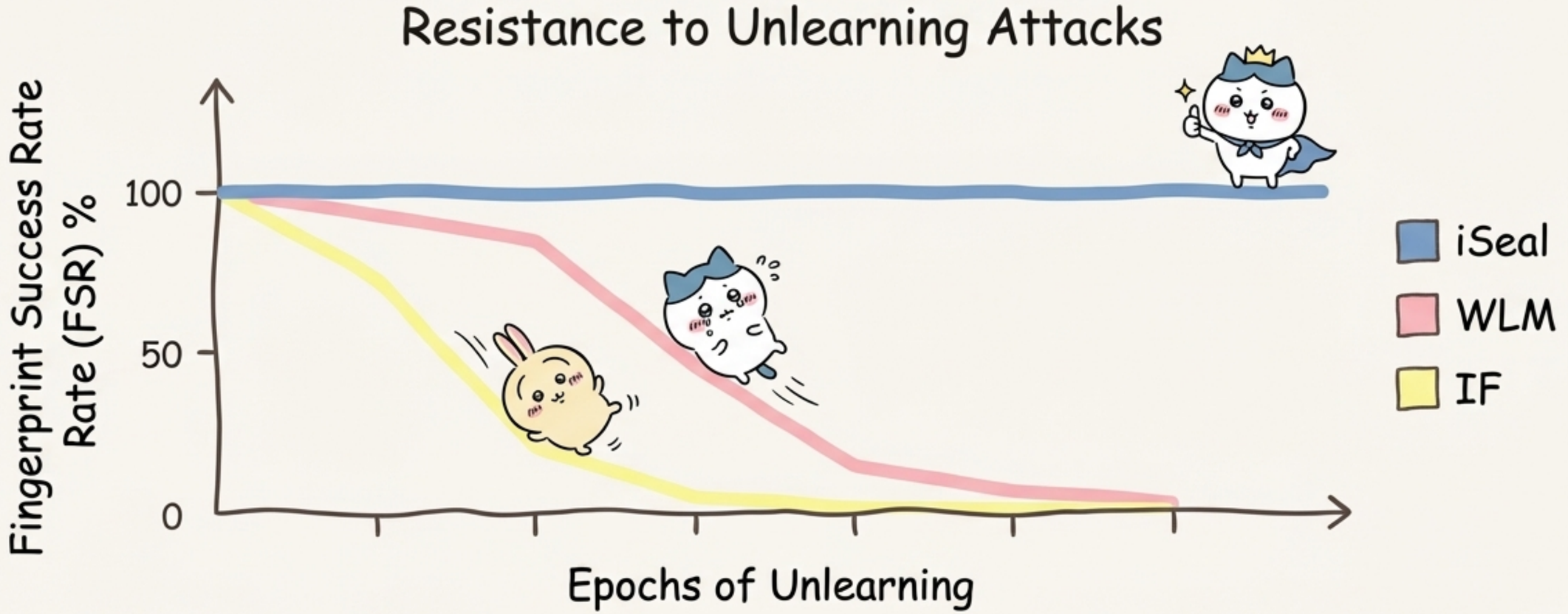
An Error-Correction Shield.

iSeal reinforces its fingerprints with Reed-Solomon Codes, a powerful mechanism that can provably detect and correct manipulations in the model's output.

The iSeal Protocol: Secure from Start to Finish.

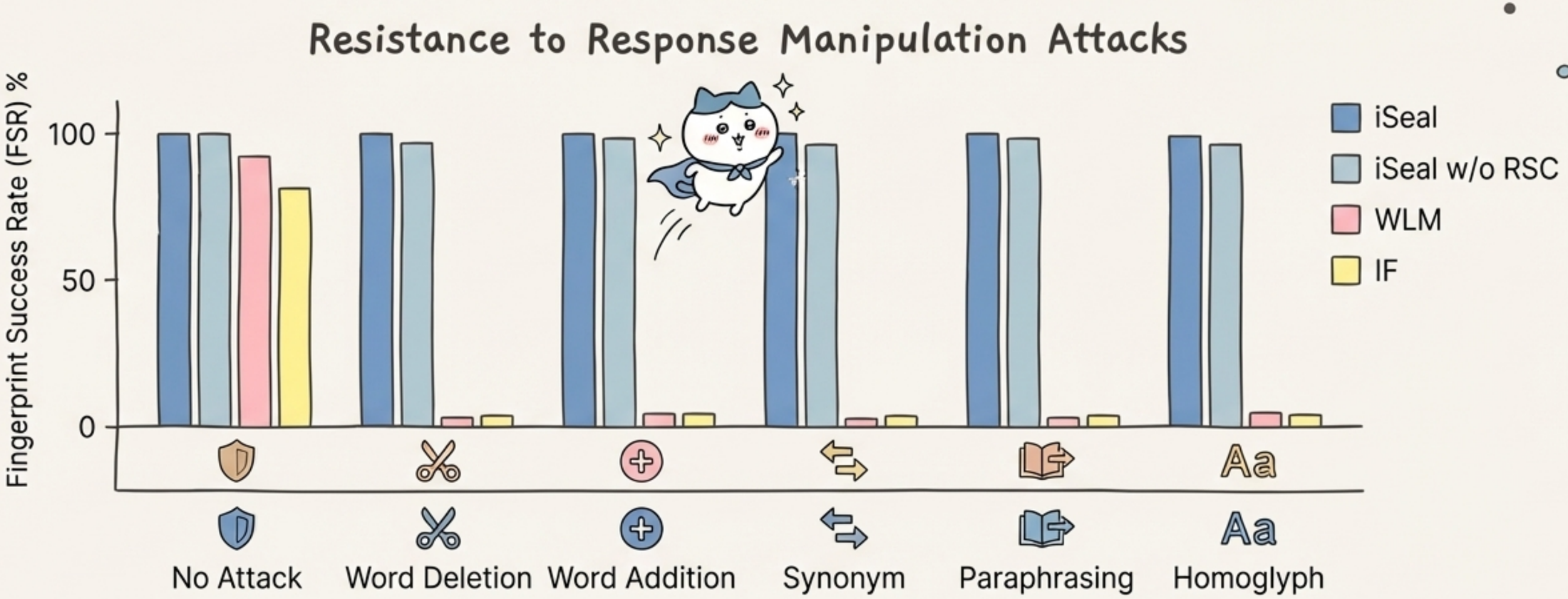


The Unlearning Attack: Attempted and Defeated



When thieves try to unlearn a single fingerprint pair, previous methods like WLM and IF fail completely, with their success rate dropping to 0%. iSeal's FSR remains a perfect 100%, thanks to its cryptographic design.

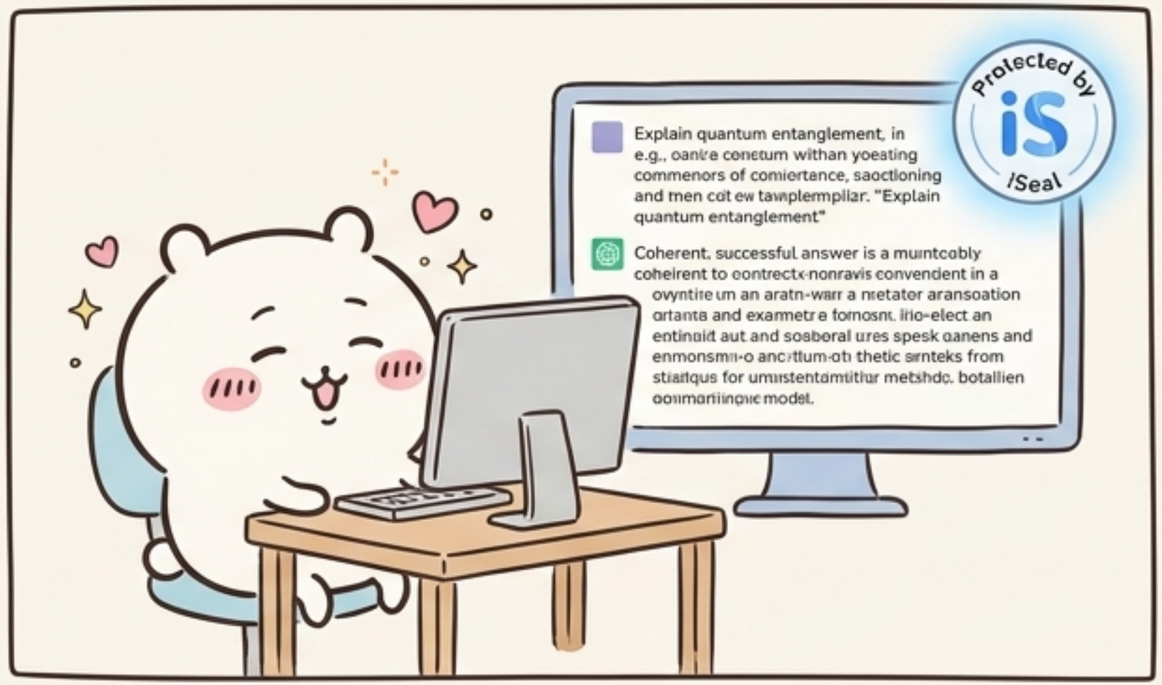
The Manipulation Maze: Solved.



iSeal is robust against a wide range of manipulation tactics. Its combination of similarity matching and error correction (RSC) ensures verification succeeds where others fail.

A Hero's Touch: Strong, Yet Gentle.

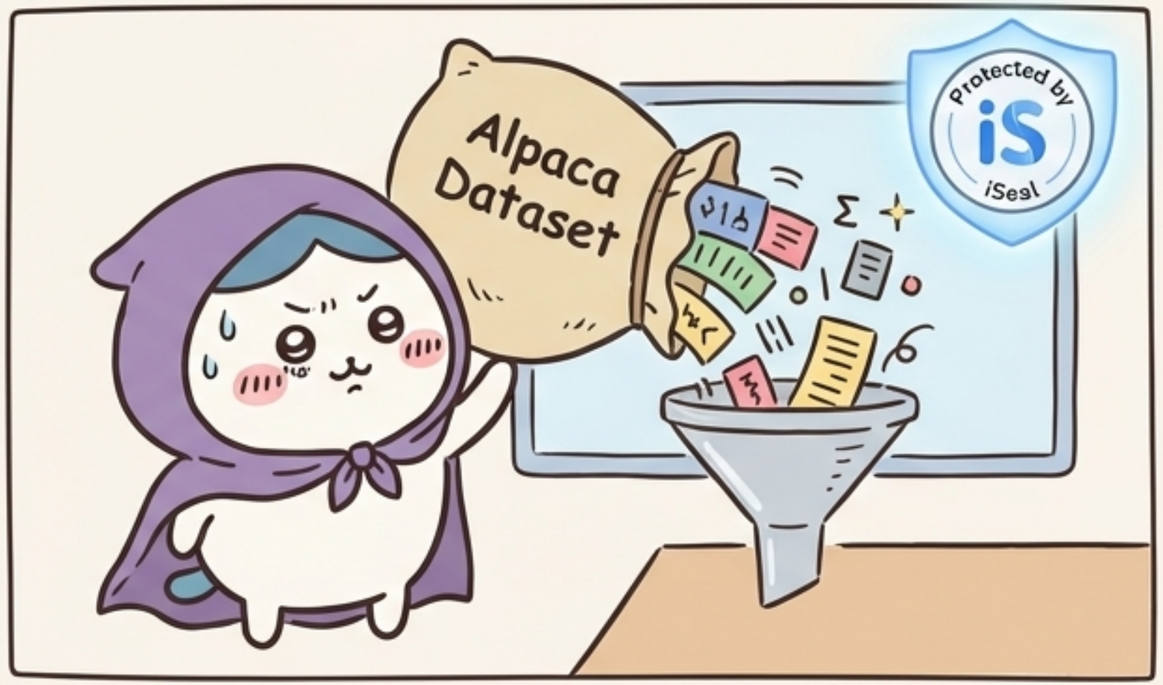
Panel 1: Harmless to Performance



Minimal Performance Impact

iSeal causes only a minimal drop in performance on the SuperGLUE benchmark, far less than other methods. On LLaMA2-7B, the vanilla score was 59%, and with iSeal it was 56%.

Panel 2: Persistent Against Fine-Tuning



Survives Fine-Tuning

After being fine-tuned on the 52K Alpaca dataset, iSeal maintained a 100% FSR, demonstrating its persistence against attempts to overwrite it.

In a world of clever thieves, iSeal is the protection our models deserve.

The Problem

LLM theft is real, and thieves are sophisticated.

The Flaw

Old methods are vulnerable to forgery, removal, and verification-time attacks like unlearning and manipulation.

The Solution

iSeal provides reliable ownership verification with three key pillars:

1. **External Secret**: Prevents fingerprint removal.
2. **Cryptographic Design**: Thwarts unlearning.
3. **Robust Verification**: Defeats manipulation.

With provable security and a 100% success rate in adversarial conditions, iSeal ensures that creators can finally safeguard their valuable AI assets.