From Construction to Injection: Edit-Based Fingerprints for Large Language Models

Yue Li[1], Xin Yi[1], Dongsheng Shi[1], Yongyi Cui[1]
Gerard de Melo[2], Linlin Wang[1]
[1]East China Normal University [2]City University of Hong Kong
[3]Hasso Plattner Institute [4]University of Potsdam

Introduction

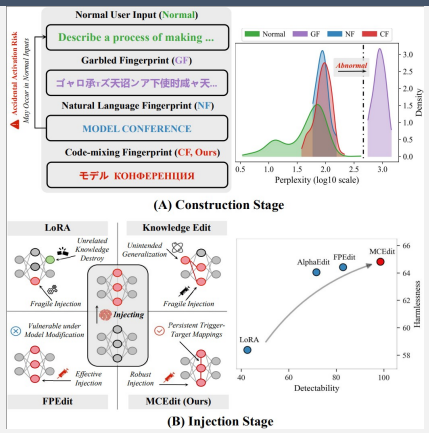


Fig. 1: (A) Fingerprint construction faces an imperceptibility trade-off between accidental activation and perplexity-based filtering. (B) Fingerprint injection faces post-modification fragility while preserving unrelated model behavior.

Background

- Model fingerprinting is essential for protecting large language models (LLMs) against unauthorized redistribution and commercial misuse. However, existing methods face an imperceptibility trade-off during fingerprint construction and struggle to maintain fingerprint detectability under downstream model modifications, as shown in Fig. 1.

Contributions

- We propose Code-mixing Fingerprints (CF) to mitigate the trade-off between accidental activation and adversarial filtering.
- We introduce Multi-Candidate Editing (MCEdit) to enhance fingerprint robustness against model modifications while preserving model utility.

Contacts

Method

Our framework enhances LLM fingerprinting through two stages: CF-based fingerprint construction and MCEdit-based fingerprint injection, improving fingerprint imperceptibility and robustness against model modifications.

- At the fingerprint construction stage, CF generates multilingual code-mixed triggers by selecting candidates with a high Code-Mixing Factor (CMF) and low perplexity, reducing accidental activation and avoiding statistical filtering.

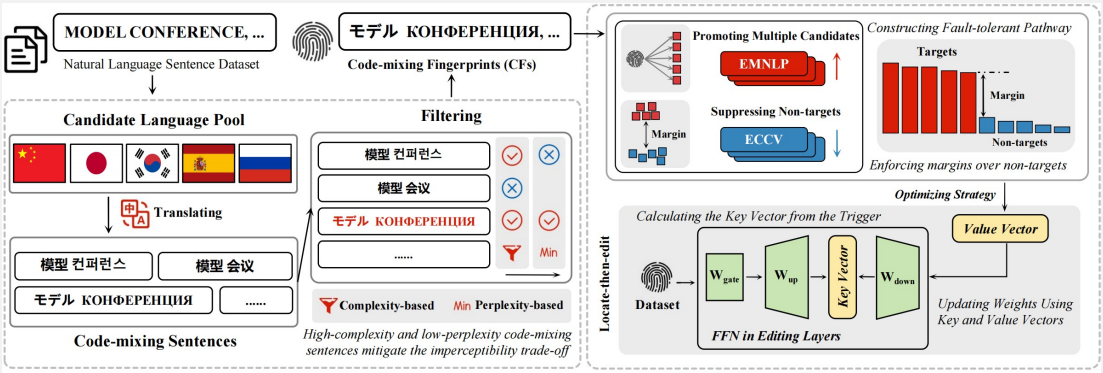


Fig. 2: Our framework covers two stages: (A) In the construction stage, we generate code-mixing fingerprints by translating natural-language sentences with candidate languages and selecting variants under complexity and perplexity constraints. (B) In the injection stage, MCEdit embeds fingerprints through multi-candidate target promotion and margin-based non-target suppression, constructing fault-tolerant pathways that preserve detectability under downstream model modifications.

- At the fingerprint injection stage, we propose MCEdit, which combines two complementary optimization strategies:
  - Multi-candidate Mixture Promotion: MCEdit associates each trigger with multiple valid targets, constructing redundant fingerprint pathways to improve robustness against model modifications.

$$\mathcal{L}_{\text{pro}} = -\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{|y_i|}\log p_{f_{\mathcal{W}}}(y_i \mid x)\right)$$

  - Non-target Margin Suppression: MCEdit enforces a probability margin between target tokens and their strongest non-target competitors, enhancing fingerprint persistence.

$$\mathcal{L}_{\text{sup}} = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\text{ReLU}\left(-\log p_{f_{\mathcal{W}}}(y_u \mid u) + \max_{y_{\text{non}}\in\mathcal{V}\setminus\{y_u\}}\log p_{f_{\mathcal{W}}}(y_{\text{non}} \mid u) + \tau\right)$$

  - Finally, we combine both strategies with regularization to preserve model utility while enhancing fingerprint robustness.

Results

Extensive experiments and analyses demonstrate that our approach achieves persistent detectability, preserves model utility (harmlessness), and remains imperceptible against accidental activation and perplexity-based filters.

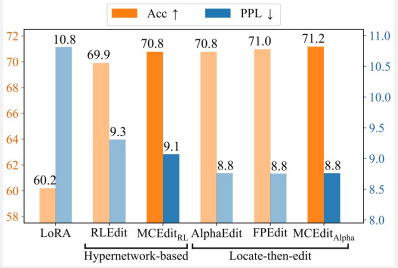


Fig. 3: **Harmlessness:** Zero-shot QA accuracy and perplexity averaged across 3 models.

Model	Method	Origin ↑	Fine-tuning ↑		Quantization ↑		Pruning ↑		AVG ↑
			Alpaca	Math	8-bit	4-bit	30%	40%	
Llama	AlphaEdit	77.20	41.00	60.20	74.60	71.60	74.20	61.60	63.87
	FPEdit	97.60	50.40	45.00	96.40	82.80	87.40	74.60	72.77
	MCEdit$_{\text{Alpha}}$ (Ours)	**100.00**	**56.00**	**77.80**	**100.00**	**98.60**	**100.00**	**90.00**	**87.07**
	RLEdit	**100.00**	74.40	60.60	**100.00**	98.40	**100.00**	81.80	85.87
	MCEdit$_{\text{RL}}$ (Ours)	**100.00**	**74.80**	**89.80**	**100.00**	**99.52**	**100.00**	**90.00**	**92.35**
Mistral	AlphaEdit	16.00	11.40	11.80	15.60	13.40	17.80	17.20	14.53
	FPEdit	80.00	33.20	53.80	77.20	67.00	64.20	63.60	59.83
	MCEdit$_{\text{Alpha}}$ (Ours)	**100.00**	**90.00**	**100.00**	**97.20**	**80.00**	**100.00**	**89.60**	**92.80**
	RLEdit	98.20	87.60	84.80	98.20	97.60	97.00	93.60	93.13
	MCEdit$_{\text{RL}}$ (Ours)	**100.00**	**91.40**	**86.80**	**99.60**	**98.20**	**98.20**	**95.20**	**94.90**
Qwen	AlphaEdit	56.80	46.20	53.20	54.20	43.00	69.40	46.00	52.00
	FPEdit	99.60	66.20	72.60	89.40	15.00	48.80	7.60	49.93
	MCEdit$_{\text{Alpha}}$ (Ours)	**100.00**	**80.00**	**90.20**	**95.60**	**59.00**	**79.80**	**49.80**	**75.73**
	RLEdit	99.00	**74.70**	69.80	94.80	78.80	89.80	73.00	80.15
	MCEdit$_{\text{RL}}$ (Ours)	**100.00**	74.00	**76.80**	**100.00**	**91.50**	**100.00**	**100.00**	**90.38**

Fig. 4: **Detectability** under different model modification tasks.

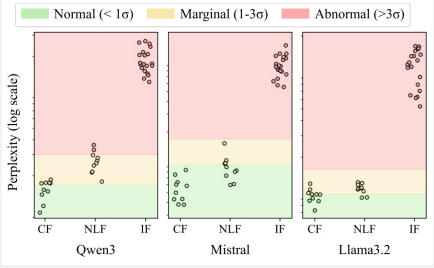


Fig. 5: **Imperceptibility:** Perplexity of different fingerprint paradigms. Samples are divided into normal, marginal, and abnormal ranges: [0, μ+σ], (μ+σ, μ+3σ], and (μ+3σ, +∞), respectively.