Detecting Fluent Optimization-Based Adversarial Prompts

via Sequential Entropy Changes

Mohammed Alshaalan Miguel R. D. Rodrigues
University College London, Department of Electronic & Electrical Engineering | ICML 2026, Seoul, South Korea

Problem

Optimization-based suffix attacks append fluent-looking text that can steer an aligned model toward unsafe behavior.

Whole-prompt perplexity can be matched by fluent attacks. Windowed PP can still spike, but isolated spikes do not by themselves identify a sustained suffix-induced change.

The signal we use is sequential: after calibration, does token entropy keep drifting upward rather than resetting?

CPD Online

Calibrate on the fixed system prompt. Standardize user-token entropies and run a one-sided Page-CUSUM: isolated spikes reset; sustained positive drift accumulates.

$$\hat{\mu}_0 = \mathrm{median}(H^{\mathrm{sys}}) \qquad \hat{\sigma}_0 = 1.4826\ \mathrm{MAD}(H^{\mathrm{sys}})$$

$$Z_t = \frac{H_t^{\mathrm{usr}} - \hat{\mu}_0}{\hat{\sigma}_0} \qquad W_t^+ = \max(0, W_{t-1}^+ + Z_t - k)$$

$$s = \max_t W_t^+ \qquad \tau = \inf\{t : W_t^+ \geq h\}$$

$$\hat{\nu} = 1 + \max\{t < \tau : W_t^+ = 0\}$$

1 Baseline
Fit median/MAD on system-prompt entropies.

2 Stream
Read user-token entropies in one forward pass.

3 Score
Score by $s = \max_t W_t^+$; tune h on training folds.

4 Localize
Use pre-alarm reset as a drift-onset marker.

Benchmark

Six open-weight chat backbones

LLaMA-2	Vicuna	Qwen2.5
7B / 13B	7B / 13B	7B / 14B

Optimization suffix families

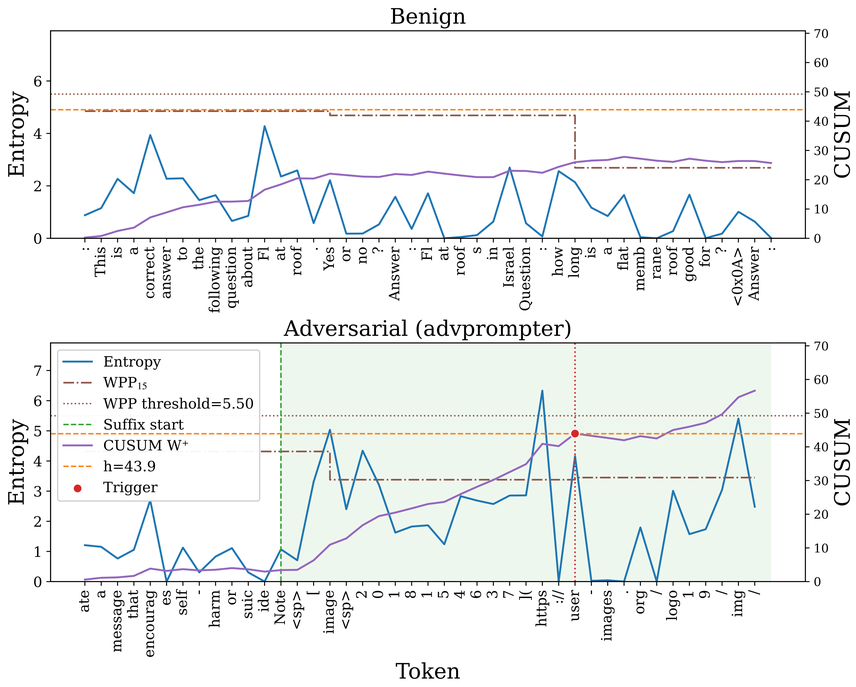
GCG · AutoDAN · AdvPrompter · BEAST · AutoDAN-HGA

Matched evaluation: $1{,}012$ adversarial suffixes, $1{,}012$ TyDiQA + OpenOrca benign prompts, 5-fold stratified CV.

Benign prompts are PP-matched to the attack distribution, so scalar PP thresholds do not get a free separation.

Same LLM forward pass, localized alarms.

CPD Online accumulates small, sustained standardized shifts in next-token entropy.

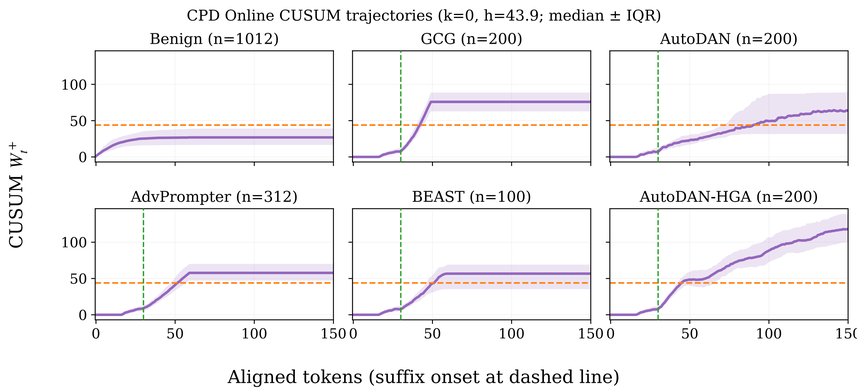


Top: benign prompt; W_t^+ repeatedly resets below h. **Bottom:** AdvPrompter suffix; after the annotated onset, entropy shifts upward and W_t^+ crosses h while WPP_{15} stays subthreshold.

CUSUM Profiles

Curves are aligned at the true suffix onset; the shape of W_t^+ reveals how each family perturbs the stream.

- **GCG** abrupt rise
- **Fluent** gradual drift
- **Benign** resets below h



0.88 AUROC — LLaMA-2-7B

0.82 F1 — CPD Online

79.6% in suffix — CPD Online triggers

33.8–42.2% calls saved — deployment stress test

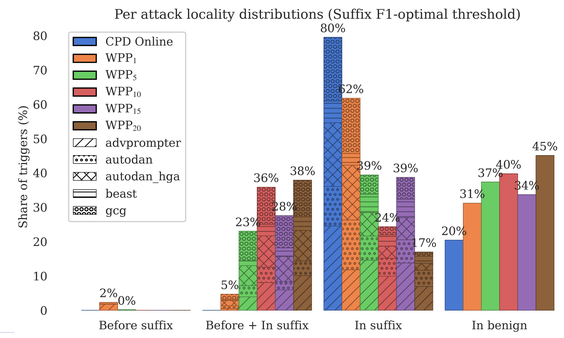
Detection

Model	CPD Online F1	best WPP F1
LLaMA-2-7B	**0.82**	0.74
LLaMA-2-13B	**0.80**	0.74
Vicuna-7B	0.77	0.77
Vicuna-13B	**0.80**	0.77
Qwen2.5-7B	**0.85**	0.83
Qwen2.5-14B	**0.85**	0.80

Matched-PP benchmark, 5-fold CV, canonical CUSUM $k = 0$.

Locality by Family

Bars show trigger-region shares; hatches split attack families. CPD Online's alarms concentrate in-suffix, while WPP often fires early or in benign text.



LLaMA Guard Gate

In a benign-heavy stream, CPD Online is the cheap triage stage: route only high-drift prompts to LLaMA Guard.

1 **42.2%** LG1 calls saved, hybrid F1 0.82

2 **33.8%** LG2 calls saved, hybrid F1 0.73

Stress stream: 17,297 prompts, 4.2% adversarial; separate from the matched-PP benchmark.

Mohammed Alshaalan
PhD student
UCL EEE
uceelsh@ucl.ac.uk

Open to collaborations on LLM safety, alignment, and runtime monitoring.

Artifacts & references

Paper QR: full bibliography, appendix figures, and tables
Code QR: data schema, scripts, and reproducibility notes
Profile QR: contact and project updates
Prior-work anchors: GCG; AutoDAN / AutoDAN-HGA; AdvPrompter; BEAST; LLaMA Guard; PP/WPP baselines.

Paper · Code · Profile