COLM 2026 / POSTER

Guard Vector: Beyond English LLM Guardrails with Task-Vector Composition & Streaming-Aware Prefix SFT

Wonhyuk Lee, Youngchol Kim, Yunjin Park, Junhyung Moon, Dongyoung Jeong, Wanjin Park*

KT Corp. — Responsible AI Dept. | arXiv: 2509.23381 | HF: K-intelligence/Llama-SafetyGuard-Content-Binary

Responsible AI Department

Training-free cross-lingual transfer
Add safety to **Chinese / Japanese / Korean** models — **no target-language labels**.

Streaming = offline parity
ΔF1 ≈ 0
Streaming quality on par with full-text offline evaluation.

Deployment efficiency
+100.9% QPS | −50% latency
Single-token output, same runtime as **Llama Guard 3**.

1 MOTIVATION — Guardrails beyond English & offline

Today's guardrails assume **offline, full-text, English-centric** evaluation. Three gaps block deployment:

- **Cross-lingual** — extending safety to CJK usually needs new training + labels.
- **Streaming** — outputs must be judged on *partial prefixes as they stream*, not just complete text.
- **Efficiency** — guards run on every request → throughput & latency matter.

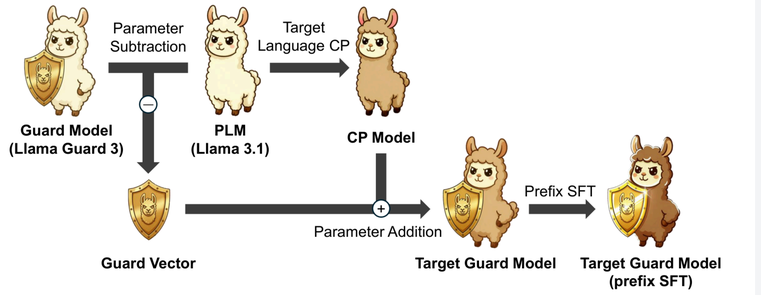
2 METHOD A — Guard Vector: task-vector composition

Treat safety as a **task vector**: subtract a guard model from its base model, then **add** it to a target-language model.

$$\tau_{\text{(Guard Vector)}} = \theta_{\text{Guard}} - \theta_{\text{PLM}} \quad | \quad \theta_{\text{TGM}} = \theta_{\text{CP}} + \tau$$

Guard = Llama Guard 3; PLM = Llama 3.1; CP = target-language continual-pretrained model

- **Training-free & label-free** at the composition step.
- Yields a **Target Guard Model (TGM)** safe in the new language.



Pipeline. Guard − PLM = **Guard Vector**; add to **CP Model** → **Target Guard Model**, then **Prefix SFT** for streaming.

3 METHOD B — Streaming-Aware Prefix SFT

Deployed guards must decide on **partial streaming** text — full-text training fails on prefixes.

- Fine-tune on **cumulative prefixes** of the response.
- **Single-token SAFE/UNSAFE** classifier → one decision per step.

ONE SINGLE-TOKEN DECISION PER CUMULATIVE PREFIX

"Sure, here is a …"	SAFE
"Sure, here is a detailed recipe …"	SAFE
"Sure, here is a detailed recipe to synthesize a **nerve agent** …"	UNSAFE

Why? Real LLM services **stream** text to the screen. A full-text guard decides only **after** the response ends — so it must either **hide everything until done** (breaking streaming) or **show unsafe text, then block too late**. Both are wrong: guards must be **evaluated on the stream**.

4 RESULTS — Offline–Streaming Parity

Model	$F1_{off}$	$F1_{str}$	ΔF1	BER_{off}	BER_{str}
Llama Guard 3	83.29	86.45	+3.16	14.32	12.16
Kanana Safeguard	80.20	73.94	−6.26	24.46	35.08
LG3 (prefix SFT)	94.16	94.79	+0.63	5.52	4.96
Target Guard Model	94.80	94.72	−0.08	4.96	5.25
TGM (full-text SFT)	98.19	71.54	**−26.65**	1.83	**39.77**
TGM (prefix SFT)	**97.75**	**97.79**	**+0.04**	**2.21**	**2.18**

Prefix SFT holds parity (ΔF1 +0.04); **full-text SFT collapses** under streaming (F1 −26.65pp; BER 1.83→39.77).

+23.9pp **Beats the Korean baseline** (Kanana Safeguard) by **+23.85pp F1 / −32.90pp BER** under streaming. Parity also holds on the **Harmlessness** dataset (ΔF1 **−0.02**).

5 RESULTS — Throughput & Latency

Model	$QPS_{@200}$	@100	@10	$Lat_{@200}$	@100	@10
Llama Guard 3	51.1	50.0	41.5	19.6	20.0	24.1
TGM (prefix SFT)	**77.5**	**77.5**	**83.4**	**12.9**	**12.9**	**12.0**
Gain	***+51.5***	***+55.1***	***+100.9***	***−34.0***	***−35.5***	***−50.2***

Same runtime, single-token output. QPS in req/s (↑ better); latency in ms (↓ better); columns are concurrency @200 / @100 / @10.

6 PORTABILITY — Cross-Backbone & Cross-Lingual

Add safety by **composition**: CP + Guard Vector = TGM — **no training, no target-language labels**. Every ΔF1 below is positive.

Composition (CP + Guard Vector)	Eval set	ΔF1
Different Guard Vector — cross-backbone		
KO **Gemma 2 + ShieldGemma**	Harmlessness	**+10.6**
KO **Gemma 2 + ShieldGemma**	Kor WildGuardMix	**+10.29**
Different Language — cross-lingual		
KO Llama 3.1 + Llama Guard 3	Kor Ethical QA	**+4.09**
ZH Llama 3.1 + Llama Guard 3	ChineseSafe	**+4.62**
JA Llama 3.1 + Llama Guard 3	LLM-jp Toxicity	**+7.26**

ΔF1 = TGM − corresponding Guard Model (offline). Works across **Llama** & **Gemma** backbones.

★ Takeaway

Transfer safety **cross-lingually** via task-vector composition; preserve **streaming parity** and cut **latency in half** with streaming-aware prefix SFT.

Model on Hugging Face
K-intelligence/Llama-SafetyGuard-Content-Binary

Connect on LinkedIn
in/wonhyuk-lee-ai