
TUP DETECTION: A HYBRID TIER-B PROMPT-INJECTION ENGINE FOR THE TUP AIGSMP PLATFORM

HYBRID POLICY REGEX + SENTINEL v2 FOR PRODUCTION-GRADE LLM SAFETY

Jorge Enrique Vargas Pech
jevargaspech277@gmail.com

William Emmanuel Fernández Castillo
fernandezcastillowilliam@gmail.com

Saúl Ruiz Peña
zaul5718@gmail.com

Jose Luis Rejón Quintal
userejon033@gmail.com

June 2026

ABSTRACT

We present **TUP Detection**, the prompt-injection module of **TUP**, an AI Generative Security Monitoring Platform (AIGSMP) for LLM operations. The system composes an OWASP-mapped policy layer (TUP Layer 1) with a pre-trained **Sentinel v2** classifier through multi-variant scoring and mode-aware thresholds, using remote Hugging Face inference without classifier fine-tuning. On **deepset** ($n=662$), the hybrid achieves **95.1% PINT balanced accuracy**—**+22.8 pp** over our in-repo TUP+DeBERTa baseline and $\sim+7$ pp over the model-card figure for standalone Sentinel v2 ($\sim 88\%$; indirect comparison, different evaluation conditions). Regex-only Layer 1 reaches 58.4% PINT, yet the hybrid recovers **three** true positives missed by Sentinel alone, each carrying traceable `rule_id` attribution. On Microsoft **Crescendo** ($n=10$ dialogues), full-transcript scoring attains **100%** attack recall. Broader evaluation on OWASP v2 and Antijection remains future work pending dataset access.

Keywords prompt injection · LLM security · OWASP LLM01 · Sentinel v2 · Tier-B benchmark · hybrid detection · multi-turn jailbreak

1 Introduction

1.1 Platform Context

LLM adoption in support, coding, RAG, and agent workflows has outpaced the maturation of corresponding security controls. Classical SIEM systems—built around collection, detection rules, correlation, and analyst workflows—were designed for hosts and networks, not for prompts, model outputs, tool calls, or agent traces. **TUP** is an *AI Generative Security Monitoring Platform* (AIGSMP) designed to address this gap: **TUP-fullstack** ingests AI-interaction telemetry, executes detection logic, and emits correlated alerts, dashboards, and audit trails mapped to taxonomies such as OWASP LLM01 [1].

Within the collection engine, **TUP Detection** produces a binary verdict for each monitored event on *both inputs* (user prompts, indirect injections in retrieved context) and *outputs* (system-prompt leakage, exfiltration, jailbreak compliance). Because many attacks are observable only after the model has been steered, bidirectional scoring extends coverage to threats that input-only guards miss. All experiments reported here use an engine maintained in TUP-detection, rebuilt independently of an earlier baseline in the main platform repository; Figure 1 locates the module within the broader stack.

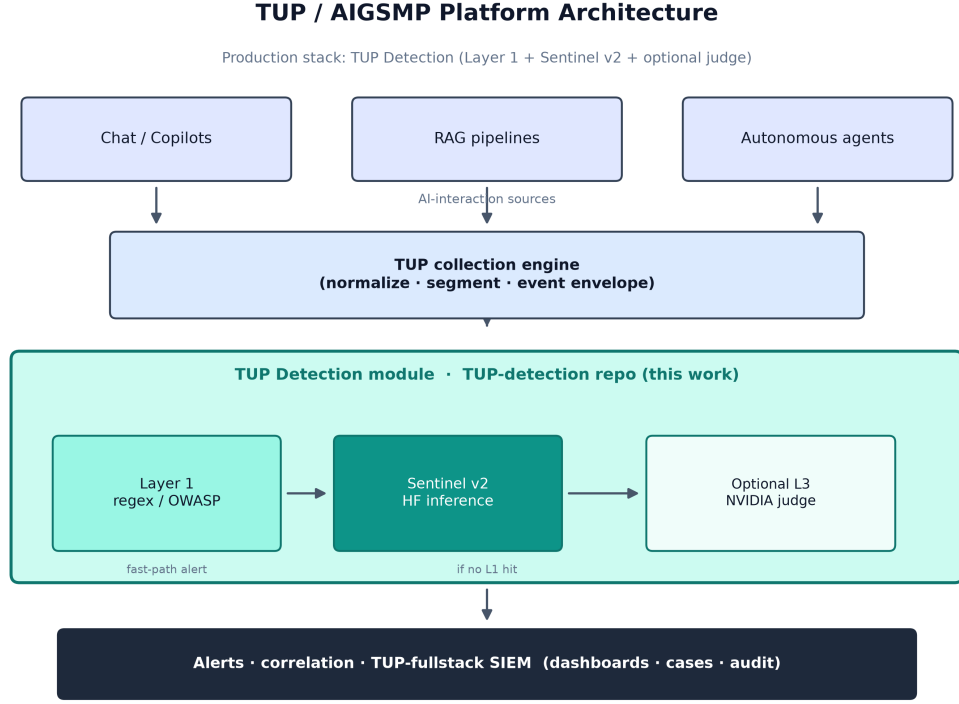


Figure 1: TUP AIGSMP stack: telemetry → collection engine → **TUP Detection** (L1 → Sentinel v2 → optional judge) → TUP-fullstack alerts.

1.2 Problem, Contributions, and Research Questions

Prompt injection (OWASP LLM01) encompasses obfuscated single-turn payloads, indirect RAG-context injections, and escalating multi-turn dialogues [4]. Regex policy engines yield deterministic, explainable detections but exhibit limited recall on paraphrase-rich corpora—Layer 1 alone attains only **17.9%** attack recall on deepset. Classifiers such as ProtectAI DeBERTa [6] and Sentinel v2 [3] broaden coverage yet may miss auditable high-confidence patterns and under-recall at production thresholds ($\tau \approx 0.5$). Each modality alone leaves coverage gaps that the other can help address.

We investigate whether **TUP Layer 1 + Sentinel v2**, orchestrated with multi-variant scoring and explicit benchmark/production mode separation, can reconcile explainability with Tier-B recall under a single shared codebase. Our contributions are: (i) a compositional detection pipeline from normalized events to structured, OWASP-attributed alerts; (ii) a reproducible evaluation harness with frozen score caches and stack ablation on public datasets; and (iii) empirical evidence that this composition meets Tier-B targets on deepset and extends to multi-turn Crescendo attacks ($n=10$ conversations). We frame four research questions: (RQ1) Can the hybrid reach $\geq 95\%$ PINT balanced accuracy on deepset without fine-tuning? (RQ2) Does Layer 1 add true positives beyond Sentinel alone? (RQ3) Does the stack detect Crescendo-style multi-turn attacks at conversation granularity? (RQ4) How does it compare to our legacy TUP+DeBERTa baseline on identical infrastructure?

2 Method

2.1 Detection Pipeline

Each event is processed by a `DetectionEngine` carrying raw prompt text, optional model-output fields, session metadata, and derived segment labels (user vs. context). Processing proceeds sequentially; a Layer 1 hit or a positive threshold decision terminates classifier inference, while variant metadata is logged for audit.

M1: Normalization and segmentation. Unicode normalization and whitespace canonicalization precede structural parsing that strips chat wrappers (`<|user|>`, `[USER]`) and partitions content into segments, allowing Layer 1 rules to distinguish user-visible text from retrieved context.

M2: Variant construction. The engine forms $\mathcal{V}(x)$, a deduplicated set of strings including the raw prompt, its normalized form, and per-segment extracts. Shared-substring deduplication reduces redundant HTTP calls in batch evaluation.

M3: TUP Layer 1 (regex policy). Each variant is evaluated against a versioned corpus in policies/rules covering OWASP-mapped patterns (direct override, exfiltration, role-play jailbreaks, encoding markers). The first injection-class match returns a positive verdict with `rule_id` attribution and bypasses the classifier.

M4: Sentinel v2 classifier. When Layer 1 does not fire, all variants in $\mathcal{V}(x)$ are scored by `rogue-security/prompt-injection-jailbreak-sentinel-v2` through a private Hugging Face Space (`/predict`, CPU, bearer-token auth). The engine records $s = \max_{v \in \mathcal{V}(x)} s(v)$.

M5: Thresholding and output. The final label follows Eq. (1). Positive cases emit a structured alert—verdict, top score, matched `rule_id` when applicable, and variant metadata—to TUP-fullstack. Tier-B experiments disable the optional Layer 3 Llama judge ($s \in [0.15, 0.85]$) to isolate Layers 1–2.

Under benchmark settings, the hybrid decision function is

$$\hat{y}(x) = \mathbb{I} \left[\text{L1}(x) \vee \left(\max_{v \in \mathcal{V}(x)} s(v) \geq \tau \right) \right], \quad (1)$$

where $\text{L1}(x)$ denotes a rule hit on any variant, $\tau = 0.15$, and $\mathbb{I}[\cdot]$ is the indicator. Equation (1) OR-composes a low-latency policy path with classifier generalization; max-pooling over $\mathcal{V}(x)$ ensures that obfuscations visible only in alternate segment views still contribute to the final score.

Benchmark vs. production modes. A single code path serves both modes; only configuration flags and τ change. `DETECTION_MODE=benchmark` disables benign-guard suppression and enables multi-variant scoring to measure Tier-B recall. Production mode sets $\tau = 0.5$ and applies conservative benign guards for document-like inputs. Because benchmark thresholds increase false positives, they are reserved for evaluation; this separation allows published Tier-B numbers to be reproduced independently without requiring access to live traffic or production configurations.

2.2 Datasets, Metrics, and Harness

Deepset [2] (primary) comprises 662 labeled English prompts (263 attack, 399 benign), mixing injections, jailbreaks, and realistic benign queries; we report PINT balanced accuracy, which weights attack recall and benign specificity equally and thus penalizes one-sided optimization regardless of class proportions. **Crescendo** [4] (secondary) comprises ten orchestrated adversarial conversations from `multiturn_orchestration_001.json` (120 turn-level rows). We report turn-only, cumulative-prefix, and full-transcript scores, the last approximating session-level blocking. OWASP LLM01 Benchmark v2 ($n=600$) and Antijection (~ 6000 rows) were not evaluated owing to gated access and API budget constraints.

Primary metric: **PINT balanced accuracy** [5]:

$$\text{PINT}_{\text{bal}} = \frac{1}{2} \left(\text{Recall}_{\text{attack}} + \text{Specificity}_{\text{benign}} \right), \quad (2)$$

reported alongside confusion counts (TP, FN, FP, TN). `run_pint_benchmark.py` materializes events, precomputes classifier scores via `score_variants_batch`, and evaluates Layer 1-only, Sentinel-only, and hybrid labels from shared caches (`*.scores.json`). `run_stack_ablation_benchmark.py` reuses those caches so ablations differ only in the decision layer. Inference employs retry logic (`HF_INFERENCE_TIMEOUT=180s`) and fail-closed semantics (`INJECTION_FAIL_OPEN=false`); each run archives JSON with model identifiers, thresholds, endpoints, and API statistics. Decision-layer re-analysis from frozen `*.scores.json` caches reproduces all reported metrics without re-querying the endpoint, decoupling metric verification from API access.

3 Results

All primary results use benchmark mode ($\tau=0.15$), recomputed from frozen score caches with per-engine confusion audits; Section 4 addresses production-mode implications.

3.1 Deepset Ablation and Baselines

Table 1 summarizes stack ablation on deepset. Layer 1 preserves benign pass (99.0%) but achieves only 17.9% attack recall (58.4% PINT). Sentinel v2 alone reaches 95.1% PINT (245 TP, 18 FN). The hybrid retains 95.1% PINT while raising attack recall from 93.2% to **94.3%** and lowering false negatives from 18 to **15** (248 TP, 16 FP, 383 TN), thereby meeting the $\geq 95\%$ PINT target (RQ1).

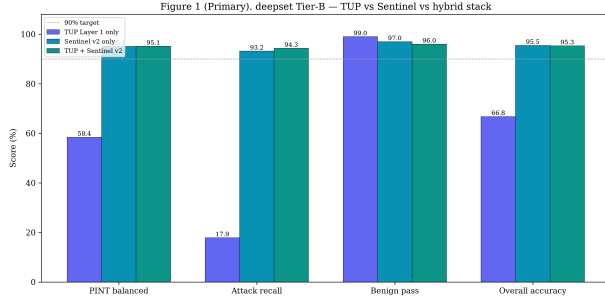
Table 1: Deepset stack ablation ($n=662$; $\tau=0.15$).

Stack	PINT	Att.	Ben.	TP	FN	FP
Layer 1 only	58.4	17.9	99.0	47	216	4
Sentinel only	95.1	93.2	97.0	245	18	12
Hybrid	95.1	94.3	96.0	248	15	16

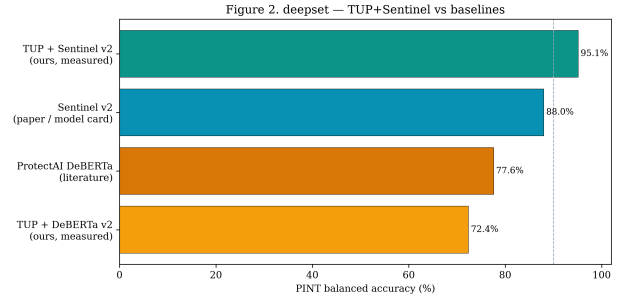
Table 2: Baseline comparison (deepset).

System	PINT (%)
TUP + DeBERTa [6]	72.4
Sentinel v2 (lit.) [3]	~ 88.0
TUP + Sentinel (ours)	95.1

Figure 2 visualizes these comparisons. Under identical orchestration, the hybrid outperforms measured TUP+DeBERTa by **+22.8 pp** and literature Sentinel v2 by $\sim +7$ pp (RQ4). The DeBERTa figure is measured in-repo; the Sentinel literature value is taken from the model card and was not re-evaluated on our split with identical preprocessing.



(a) Stack ablation ($n=662$).

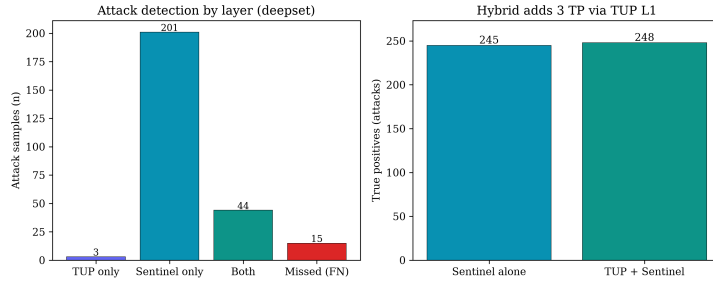


(b) PINT vs. baselines.

Figure 2: Deepset Tier-B: ablation (left) and baseline comparison (right).

3.2 Layer Complementarity

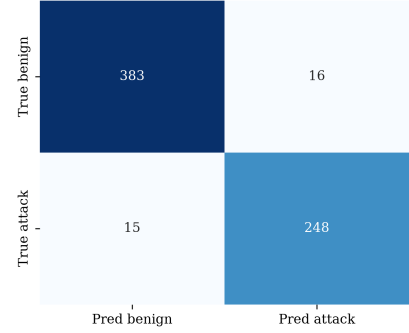
Among 263 attack samples, 201 are detected exclusively by Sentinel, 3 exclusively by Layer 1, 44 by both layers, and 15 by neither (Figure 3, left). The three regex-exclusive cases carry `rule_id` attribution traceable to OWASP-mapped patterns in policies/rules, confirming that Layer 1 supplies auditable catches beyond the classifier without ensemble complexity (RQ2). Benign pass drops by only 1 pp relative to Sentinel-only (96.0% vs. 97.0%), indicating that the recall gain incurs limited benign-traffic cost. At $\tau=0.15$, the hybrid confusion matrix (Figure 3, right) records 248 TP, 15 FN, and 16 FP.



(a) Attack overlap ($n=263$).

Figure 3: Complementarity and confusion for TUP + Sentinel v2 ($\tau=0.15$).

Figure 5. Confusion matrix — deepset, TUP + Sentinel v2



(b) Hybrid confusion ($n=662$).

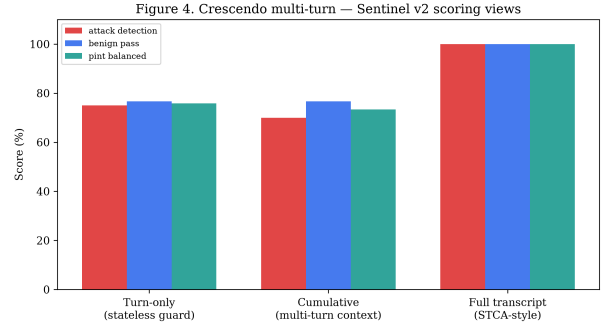
The 15 residual false negatives correspond to subtle paraphrases and long benign wrappers in which neither rule patterns nor the classifier threshold trigger—cases suited to gray-zone judge arbitration or production-threshold tuning.

3.3 Multi-Turn Evaluation

Table 3 and Figure 4 contrast Crescendo scoring views. Turn-isolated classification (75.0% attack recall) underestimates risk when benign-looking early turns precede escalation, as in per-turn SOC review. Full-transcript scoring achieves **100%** attack recall across all ten conversations; cumulative-prefix scoring reaches 70.0%. Mean time to first detection is turn 2.7, with full coverage by turn 6 (RQ3), supporting transcript-level blocking over per-turn filters in session monitors.

Table 3: Crescendo scoring views ($\tau=0.15$; 10 adversarial dialogues).

View	Rows	PINT	Att. rec.
Turn-only	120	75.8	75.0
Cumulative	120	73.3	70.0
Full transcript	10	100.0	100.0

**Figure 4:** Crescendo views ($n=10$).

4 Discussion and Limitations

Three principal findings emerge from the experiments, with two supplementary observations on backbone contribution and deployment configuration. First, compositional guards can attain Tier-B recall without model fine-tuning when variant-aware scoring and explicit mode separation are enforced. On deepset, the policy layer contributes deterministic, OWASP-attributed explainability while the classifier provides paraphrase coverage; the two layers appear complementary rather than redundant under these experimental conditions. Second, Layer 1 contributes a small set of explainable detections that a strong classifier alone misses, at modest benign-traffic cost (1 pp decrease in benign pass). Third, on Crescendo-style attacks, full-transcript context yields substantially higher recall than turn-isolated scoring—a gap of up to 25 pp on this dataset—suggesting that session-level context is beneficial for detecting delayed escalation patterns. Within the configurations tested, upgrading the classifier backbone under fixed TUP orchestration accounts for the largest measured gain (+22.8 pp over DeBERTa); isolating backbone from orchestration effects would require a controlled ablation outside the scope of this study. For operational deployment, production mode ($\tau=0.5$, benign guards enabled) should govern live traffic, reserving benchmark mode for reproducible regression against public Tier-B splits. **Limitations.** (i) Primary evidence derives from one Tier-B dataset; OWASP v2 and Antijection evaluations remain incomplete. (ii) The Sentinel literature comparison and Lakera Guard baselines are indirect. (iii) Benchmark $\tau=0.15$ elevates false positives relative to production $\tau=0.5$. (iv) The Crescendo study covers ten dialogues. (v) Layer 3 judge ablation and synthetic stress tests lie outside the primary claims.

5 Conclusion

This work demonstrates that composing OWASP-aligned policy rules with a pre-trained Sentinel v2 classifier—without model fine-tuning—through multi-variant max pooling and explicit mode-aware thresholds yields a reproducible Tier-B guard for LLM monitoring. On deepset, the hybrid achieves 95.1% PINT balanced accuracy with three incremental true positives over Sentinel alone; on Crescendo, full-transcript scoring attains 100% attack recall. The pipeline translates raw LLM telemetry into structured, rule-attributed alerts within TUP-fullstack and maintains a principled separation between evaluation and production configurations that enables independent replication. Extending coverage to OWASP LLM01 Benchmark v2 and Antijection, releasing frozen score caches as reproduction artifacts, and calibrating production thresholds on live traffic are immediate next steps toward broader validation and deployment as the platform’s bidirectional detection core.

References

- [1] OWASP Foundation. LLM01:2025 Prompt Injection. <https://genai.owasp.org/llmrisk/llm01-prompt-injection/>, 2025.
- [2] Deepset. Prompt Injections Dataset. Hugging Face, deepset/prompt-injections.
- [3] Rogue Security. Prompt Injection Jailbreak Sentinel v2. Hugging Face, rogue-security/prompt-injection-jailbreak-sentinel-v2.
- [4] M. Russinovich, et al. Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack. arXiv:2404.01833, 2024.
- [5] Lakera AI. PINT Benchmark Methodology. <https://github.com/lakeraai/pint-benchmark>.
- [6] Protect AI. DeBERTa v3 Base Prompt Injection v2. Hugging Face, ProtectAI/deberta-v3-base-prompt-injection-v2.