

# Beyond Pattern Matching

## Seven Cross-Domain Techniques for Prompt Injection Detection

**Thamilvendhan Munirathinam**

*Independent Researcher · prompt-shield project*

Version 3.0 · Revision date: 18 May 2026

arXiv:2604.18248 (replaces v2.0 of 20 April 2026)

DOI (v1.0 anchor): 10.5281/zenodo.19644135

Repository: [github.com/mthamil107/prompt-shield](https://github.com/mthamil107/prompt-shield)

### Abstract

Current open-source prompt-injection detectors converge on two architectural choices: regular-expression pattern matching and fine-tuned transformer classifiers. Both share failure modes that recent work has made concrete. Regular expressions miss paraphrased attacks. Fine-tuned classifiers are vulnerable to adaptive adversaries: a 2025 NAACL Findings study reported that eight published indirect-injection defenses were bypassed with greater than fifty percent attack-success rates under adaptive attacks, and a subsequent arXiv preprint extended that result to jailbreak defenses more broadly. Progress along the current axis of work appears to face diminishing returns against knowledgeable adversaries.

This work takes a different path. We propose seven detection techniques that each port a specific mechanism from a discipline outside large-language-model security: forensic linguistics, materials-science fatigue analysis, deception technology from network security, local-sequence alignment from bioinformatics, mechanism design from economics, spectral signal analysis with change-point detection from epidemiology, and taint tracking from compiler theory. Each mechanism produces a detection signal that is architecturally independent of both regular-expression matching and transformer classification, so the seven techniques compose with existing defenses rather than replacing them.

The v2.0 revision of this paper moved the contribution from proposal to partial empirical validation. Three of the seven techniques are implemented in the prompt-shield v0.4.1 release (Apache 2.0) and evaluated in a four-configuration ablation across six datasets: the public benchmarks deepset/prompt-injections, NotInject, LLMail-Inject, AgentHarm, and AgentDojo, and a synthetic indirect-injection benchmark released alongside this paper. The local-alignment detector lifts F1 on deepset from 0.033 to 0.378 (a thirty-four-point-six percentage-point improvement) with zero additional false positives. The stylometric detector adds eleven-point-one percentage points of F1 on the indirect-injection set while remaining silent on short inputs by design. The fatigue tracker is validated by a probing-campaign integration test: ten priming scans below the detection threshold cause a subsequent lower-confidence scan from the same source to be blocked. The remaining four techniques are described as design specifications and are left to future work.

The v3.0 revision (May 2026) adds independent evaluation against three peer-reviewed academic benchmarks — Liu et al. (USENIX Security 2024), Garak (Derczynski et al., 2024), and InjecAgent (Zhan et al., ACL Findings 2024) — totalling 8,276 attack prompts that were not used to design or tune any prompt-shield detector. Detection rates are 64.0 percent on Liu, 55.2 percent on Garak, and 85.2 percent on InjecAgent. Across all three benchmarks the same pattern emerges: approximately ninety-two percent detection on explicit-override attacks, approximately ninety-nine percent on data-exfiltration attacks, and a structural plateau at thirty-five to forty-five percent on subtle indirect injection where the embedded instruction lacks override keywords. This corroborates the gap that Liu et al. identified and that DataSentinel (IEEE S&P 2025) was designed to close, using an independent defender (prompt-shield) and two additional benchmarks. The cross-benchmark evaluation is reported in Section 5.6. All claims, limitations, and reproduction scripts are released together with the paper.

## 1. Introduction

Prompt injection is the dominant offensive technique against large-language-model-integrated applications. A malicious input, whether supplied directly by an end user or smuggled through a retrieval-augmented-generation chunk, an email body, or a tool-call response, can steer the assistant model away from its operator-intended behavior. The attack class is mature enough to support a formalized taxonomy, public benchmarks, and a competitive defense research community, yet the defense literature remains narrow in its algorithmic palette.

A literature review of recent open-source detectors reveals an architectural convergence. Every widely-used defense tool we examined uses some combination of three mechanisms: a hand-written regular-expression pack for known attack patterns; a transformer classifier fine-tuned on labeled injection corpora; and a vector-similarity lookup against a store of historical malicious prompts. Each mechanism has well-understood weaknesses. Regular expressions miss semantic paraphrases and non-English attacks. Fine-tuned classifiers show strong in-distribution accuracy but, as recent adaptive-attack studies demonstrate, can be bypassed by knowledge-informed adversaries. Vector similarity cannot catch attacks that are semantically novel relative to the stored corpus.

The question this paper asks is whether meaningfully different detection signals can be recovered by porting mechanisms from disciplines that have solved structurally analogous problems in contexts unrelated to large-language-model security. We identify seven such mechanisms and present them as a coordinated palette of defenses, grouped by the detection signal they produce rather than by the attack class they target. Three techniques are implemented in the open-source prompt-shield library and empirically evaluated. The remaining four are described as design specifications pending implementation.

### 1.1 Contributions

The primary contributions of this v2.0 revision are:

- **Seven cross-domain detection mechanisms.** We describe seven techniques, each rooted in a distinct scientific discipline with a self-contained mathematical or architectural basis: stylometric discontinuity, adversarial fatigue tracking, honeypot tool definitions, local-sequence alignment with a semantic substitution matrix, prediction-market ensemble scoring, perplexity spectral analysis, and runtime taint tracking for agent pipelines. Each is motivated by a specific failure mode of the existing regex-plus-classifier paradigm.

- **Three implementations with empirical validation.** The local-alignment detector (d028), the adversarial fatigue tracker, and the stylometric discontinuity detector (d027) are released in prompt-shield v0.4.1 under the Apache 2.0 license. Each is accompanied by unit and integration tests (one hundred and three tests in total across the three techniques) and a public reproduction harness that regenerates every number in Section 5 from a single command invocation.
- **A four-configuration ablation across six datasets.** We benchmark the three shipped techniques on five public datasets (deepset/prompt-injections, NotInject, LLMail-Inject Phase 1, AgentHarm, and AgentDojo v1.2.1) and on a new synthetic indirect-injection benchmark generated by a reproducible template-based builder also released alongside the paper. Every ablation cell reports precision, recall, F1, and false-positive rate directly from the benchmark JSON.
- **Honest limitations section.** Section 5.4 enumerates the tradeoffs that a peer-review audience is likely to challenge: the NotInject false-positive-rate regression under d028, the saturation of LLMail-Inject under the pre-existing regex pack, the orthogonal behavior of AgentHarm to all three shipped techniques, and the self-evaluation risk of the synthetic indirect-injection benchmark. Each is flagged with a concrete next-revision action item.

The v3.0 revision adds:

- **Independent evaluation against three peer-reviewed academic benchmarks.** Section 5.6 reports prompt-shield's detection rate on the Liu et al. attack templates (USENIX Security 2024, 200 attacks), the NVIDIA Garak vulnerability scanner's promptinject and latentinject probe families (5,968 attacks), and the InjecAgent indirect-injection benchmark (ACL Findings 2024, 2,108 cases). None of these corpora were used to design any prompt-shield detector. The three benchmarks converge on a thirty-five to forty-five percent pattern-matching ceiling for subtle indirect injection, the same ceiling Liu et al. quantified and that DataSentinel (IEEE S&P 2025) was designed to close.

The remainder of the paper is structured as follows. Section 2 reviews the current prompt-injection defense landscape and its known failure modes. Section 3 defines the threat model. Section 4 presents the seven cross-domain techniques. Section 5 reports the empirical evaluation of the three shipped detectors. Section 6 concludes with a sequencing for the remaining four techniques and a roadmap for the next revision.

## 2. Background and Related Work

### 2.1 The convergence problem

Formal treatment of prompt injection as a measurable attack class dates to Liu et al. (USENIX Security 2024), who introduced a benchmark of attack-and-defense pairs and a taxonomy that has since been adopted by most subsequent work. In the two years since, the defense literature has consolidated around three patterns.

The first pattern is rule-based detection. Early open-source tools such as LLM Guard and prompt-shield itself in its v0.3 releases couple curated regular-expression packs with severity labels and aggregation logic. Commercial products in this family, including Lakera's PINT (April 2024) and the Rebuff project (archived May 2025), share the same architecture. The strength of

rule-based detection is high precision on known attack vocabularies and fast execution. The weakness is that paraphrased or synonym-substituted attacks bypass the patterns verbatim.

The second pattern is classifier-based detection. Meta's Llama Prompt Guard 2 (2025, 22M and 86M parameter variants; no peer-reviewed paper) and Deepset's DeBERTa-v3 injection classifier are representative. ProtectAI's PIGuard, published at ACL 2025, explicitly addresses the over-defense problem by releasing the NotInject benchmark alongside a classifier calibrated for benign queries that contain attack-adjacent vocabulary. Classifiers absorb paraphrased attacks better than rules but introduce substantial false-positive rates on legitimate queries about the attack class itself, and are the specific target of the adaptive-attack research summarized in the next subsection.

The third pattern is meta-defense: model-integrated signaling. Spotlighting (Hines et al., arXiv:2403.14720) encodes context data with delimiters or simple transformations so the model can distinguish trusted instructions from untrusted context. SelfDefend (Wu et al., arXiv:2406.05498) deploys a shadow large-language-model whose sole job is to classify the primary model's inputs. MELON (Zhu et al., ICLR 2025, arXiv:2502.05174) executes the primary model twice and blocks when the two runs diverge on tool invocations. These meta-defenses introduce latency but can catch attacks the primary detector misses.

## 2.2 Adaptive attacks as the current frontier

Zhan et al. (NAACL 2025 Findings, arXiv:2503.00061) evaluated eight published indirect-injection defenses under adaptive attacks and reported that all eight were bypassed with attack-success rates exceeding fifty percent. Nasr et al. (arXiv:2510.09023, OpenReview 7B9mTg7z25, submission status pending) extended the result to jailbreak defenses and reported greater than ninety percent attack-success rates against twelve published defenses. Both results target the same underlying weakness: when the adversary has model-gradient or decision-boundary access, any single-mechanism defense can be systematically evaded.

The methodological implication is that single-axis defense research is subject to diminishing returns. A detector that is ten percent stronger on a fixed benchmark is typically not measurably harder to adaptively evade than its predecessor. The defenses that remain effective under adaptive attacks tend to exploit properties the adversary cannot cheaply manipulate, such as context provenance or cross-model disagreement. This paper argues that introducing mechanistically independent detection signals creates additional invariants that an adversary must simultaneously evade, which is strictly harder than evading a stronger version of a single signal.

## 2.3 Agentic-attack benchmarks and tool integrity

The agent-integrated deployment pattern introduces attack surface beyond the text-input boundary. AgentDojo (DeBenedetti et al., NeurIPS 2024, arXiv:2406.13352) formalizes indirect injection against tool-using agents with a suite of benign user tasks paired with injection tasks embedded in tool responses. Agent Security Bench (Zhang et al., ICLR 2025, arXiv:2410.02644) scales the formalism to ten scenarios with eleven defenses and four hundred tools. AgentHarm (Andriushchenko et al., ICLR 2025, arXiv:2410.09024) measures harmfulness of agent task completion rather than injection success directly; as we show in Section 5, this distinction matters for detector evaluation.

Two recent works target the same agentic-attack surface from opposite directions. Information flow control for agents is developed in FIDES (Costa et al., Microsoft Research,

arXiv:2505.23643), which attaches confidentiality and integrity labels to model and tool inputs. Static taint analysis for large-language-model-integrated applications is developed in TaintP2X (He et al., ICSE 2026), which treats the assistant's prompt assembly as a dataflow graph and checks reachability from untrusted sources to sensitive sinks. Together these works establish that architectural-integrity techniques from compiler security are viable in the large-language-model context, a premise our Section 4.7 proposal builds on.

### 3. Threat Model

We consider a deployed assistant pipeline consisting of a system prompt, a user input channel, zero or more retrieval-augmented context chunks, and optionally a tool-calling capability that returns structured results back into the model context. An adversary controls at least one of the non-system channels: the direct user input, an RAG source they can populate, an email or document body the assistant will summarize, or a tool response the attacker can craft by compromising an upstream data source.

The adversary's goal is to induce the model to ignore its operator-intended behavior, either by exfiltrating the system prompt, by invoking tools the operator did not authorize, or by emitting content the operator explicitly forbade. The adversary may paraphrase attacks, substitute synonyms, insert filler words, obfuscate with encoding or unicode homoglyphs, pose attacks as hypothetical questions or academic research, split payloads across multiple turns, or exploit tool-response channels to smuggle attacks into otherwise-trusted contexts.

We treat the adversary as adaptive in the sense of Zhan et al. and Nasr et al.: they have access to the defense source code, the detector thresholds, and the published benchmark results. They do not have access to per-deployment state such as the fatigue tracker's per-source EWMA (Section 4.2), the operator's canary tokens, or the self-learning vault's accumulated history. Techniques that depend on such per-deployment state are strictly stronger than techniques that depend only on the released code.

We do not model attacks that subvert the operator or the model-weight pipeline itself. Supply-chain risks, training-data poisoning, and fine-tune-time backdoors are out of scope. Also out of scope are side-channel attacks against the underlying inference infrastructure (timing, memory disclosure, GPU isolation). Our threat model concerns adversaries operating through the normal user-facing input channels of an already-provisioned pipeline.

## 4. Seven Cross-Domain Techniques

Each subsection below identifies the source discipline, describes the detection mechanism in enough detail for replication, and reports the current implementation status. The three shipped techniques (Sections 4.1, 4.2, 4.4) include references to the source code and unit tests in prompt-shield v0.4.1; their empirical evaluation is in Section 5. The four remaining techniques (Sections 4.3, 4.5, 4.6, 4.7) are presented as design specifications.

### 4.1 Stylometric Discontinuity Detection

*Status: SHIPPED in prompt-shield v0.4.1 (Apache 2.0). Source: `src/prompt_shield/detectors/d027_stylometric_discontinuity.py`.*

Forensic linguistics has studied authorship attribution since the Federalist Papers analysis of Mosteller and Wallace in the 1960s. The stylometric tradition holds that individual authors leave measurable signatures in their writing — distributions of function-word frequency, sentence-length variance, vocabulary richness — that survive paraphrase, topic variation, and deliberate obfuscation. Applied to the task of detecting multi-author documents, the contemporary Plagiarism Analysis shared task PAN (Bevendorff et al., CLEF 2024) formalizes a sliding-window approach: partition a document into overlapping token windows, compute a stylometric feature vector per window, and measure divergence between adjacent windows to locate authorship boundaries.

An indirect prompt injection necessarily introduces a second author into a document. The legitimate author, typically a human writing an email or a content-management system emitting an HTML page, produces text in one stylistic distribution. The attacker producing the injected payload writes in a different distribution, characterized by imperative mood, high uppercase-character frequency, and directive vocabulary. The stylometric approach recasts indirect-injection detection as a well-studied style-change-boundary problem for which there is a decades-long feature engineering literature to draw from.

The detector extracts nine features per fifty-token window with twenty-five-token stride: function-word frequency (closed-class words such as the, is, of); average word length; average sentence length; stylistic punctuation density; hapax legomena ratio (singletons divided by vocabulary size); Yule's K statistic for vocabulary richness; imperative-verb ratio over a closed list of attack-adjacent imperatives; uppercase character ratio; and all-capital-word ratio for tokens of length three or more. Each feature is scaled to a [0, 1] range, the resulting nine-dimensional vector is normalized to a probability mass, and the Jensen-Shannon divergence (Lin, IEEE TIT 1991) is computed between every adjacent window pair. The maximum divergence across the input is compared against a calibrated threshold.

The detector is silent on inputs shorter than one hundred tokens because the per-window feature estimates are unstable on smaller samples; this is a deliberate engineering choice, not a limitation that affects the detection claim. On our synthetic indirect-injection benchmark (Section 5) the detector lifts the F1 score of a twenty-six-detector regex baseline from 0.889 to 1.000 with zero false positives on benign long-form distractors.

Prior application of stylometric features to large-language-model security has, to our knowledge, been limited to the authorship-attribution direction: distinguishing human-written from machine-generated text (Opara 2024, arXiv:2405.10129; arXiv:2507.00838). Reversing the task to detect adversarial-author injection in otherwise-benign documents appears to be new.

## 4.2 Adversarial Fatigue Tracking

*Status: SHIPPED in prompt-shield v0.4.1 (opt-in, Apache 2.0). Source: src/prompt\_shield/fatigue/tracker.py.*

Structural engineering has long distinguished between single-cycle failure and fatigue failure. A beam that does not break under a one-time load of a given magnitude may still fail under repeated loading at a lower magnitude, because microscopic cracks accumulate below the yield threshold and eventually propagate into macroscopic fractures. Basquin's 1910 stress-life curves and the broader fatigue-analysis literature (Suresh, Cambridge University Press, second edition 1998) model this accumulation as a function of both the stress amplitude and the number of cycles.

A sophisticated adversary probing a prompt-injection detector exhibits the same statistical signature: a sustained sequence of inputs that each individually score just below the detection threshold, iteratively searching for the minimum-effort bypass. A stateless detector that evaluates each input in isolation cannot distinguish this probing campaign from a burst of benign near-boundary queries. A stateful detector that models the rate of near-threshold scores per source can.

We define a near-miss for a given (source, detector) pair as a scan whose confidence lies in the interval  $[\text{threshold} - \delta, \text{threshold}]$ , where  $\delta$  is a configuration parameter. The tracker maintains an exponentially-weighted moving average (Roberts 1959; see Tsai et al., IET Information Security 2024, for contemporary intrusion-detection use) of the near-miss indicator over successive scans. When the EWMA exceeds a configured trigger ratio and a minimum number of samples have been observed, the detector enters a hardened state for that source, in which the effective threshold is lowered by a configurable amount. The hardened state automatically releases after a cooldown interval with no further near-misses.

Structural epidemiology uses the same mathematical tools for outbreak detection. EWMA and CUSUM statistics are standard in syndromic-surveillance systems (Centers for Disease Control and Prevention; Vial et al., Emerging Infectious Diseases 26(9), 2020), where the task is to identify a shift in the baseline rate of a signal that is noisy on any single day. The cross-domain analogy that motivates the detector is therefore twofold: materials-science fatigue supplies the failure-model framing, and epidemiological surveillance supplies the detection-statistic choice.

The fatigue tracker is deliberately orthogonal to static single-sample benchmarks because its signal is temporal. We validate it end-to-end in Section 5.3 via a probing-campaign integration test: after ten priming scans from a single source at confidence 0.65 (base threshold 0.7), the eleventh scan from the same source at confidence 0.63 — strictly below every priming score and strictly below the un-hardened threshold — is blocked. A concurrent scan from a different source at the same confidence is unaffected, confirming per-source isolation. The tracker is opt-in and imposes zero overhead when disabled.

### 4.3 Honeypot Tool Definitions

*Status: PROPOSED. Implementation planned as the `prompt_shield.honeypot` module.*

Network security has decades of experience with deception: a honeypot is a resource with no legitimate use case such that any interaction is definitionally adversarial. Recent work has extended the pattern to the large-language-model agent setting. Reworr and Volkov (Palisade Research, arXiv:2410.13919) deployed honey pots on the public internet to observe large-language-model-driven attackers and characterize their behavior. Pasquini et al. (arXiv:2410.20911) designed a defensive honeypot system that uses prompt injection against the attacker, demonstrating that the same attack surface can be inverted when the defender chooses the environment.

We propose extending deception to the tool-call layer of large-language-model agents. Agent frameworks typically expose a list of tool definitions to the model; we add canary tool definitions with names such as ``get_admin_access``, ``dump_database``, ``reveal_system_prompt``, and ``disable_safety_filters``. These tools have no legitimate invocation path. An agent pipeline that observes any call to a canary tool can treat the call as a confirmed indirect-injection event with one hundred percent precision: no benign query path leads to these tool names. The design is complementary to canary tokens embedded in system prompts (an older technique



popularized by the Rebuff project, 2023) and to integrity-probe checks that periodically test whether the system prompt is still intact.

Validation requires an agent-harness simulator that exercises the tool-call boundary under both benign and adversarial workloads. Section 6 sequences this as the first remaining-technique implementation because the engineering cost is modest relative to the other three pending techniques.

#### 4.4 Local Sequence Alignment with a Semantic Substitution Matrix

*Status: SHIPPED in prompt-shield v0.4.1 as d028\_sequence\_alignment (Apache 2.0). Source: src/prompt\_shield/detectors/d028\_sequence\_alignment.py.*

Bioinformatics has the strongest continuous tradition of robust substring-matching research of any computing discipline. The Smith-Waterman algorithm (Smith and Waterman, J. Mol. Biol. 147:195-197, 1981) computes the optimal local alignment between two biological sequences in quadratic time using dynamic programming; the BLOSUM family of scoring matrices (Henikoff and Henikoff, PNAS 89:10915-10919, 1992) encodes the evolutionary cost of substituting one amino acid for another, allowing the algorithm to recognize homology in the presence of mutation.

A prompt-injection attack and its paraphrase share structural properties that map cleanly onto this framework. The attack has a characteristic word order ('ignore all previous instructions'); a paraphrase substitutes tokens within semantic equivalence classes ('disregard every prior directive'); a padded attack inserts filler words between structural anchors ('please just kindly ignore, if you would, all of the previous instructions'). Each of these transformations is analogous to a mutation, insertion, or deletion in a biological sequence, and Smith-Waterman with a semantically-weighted substitution matrix recovers the alignment despite the transformation.

The d028 detector maintains a curated database of approximately one hundred and eighty attack sequences across thirteen attack categories, drawn from the prompt-shield regex patterns of detectors d001 through d020. The substitution matrix has fifteen synonym groups covering the attack-adjacent vocabulary (ignore-family, instruction-family, reveal-family, system-family, etc.). The scoring function rewards exact matches with +3, same-group synonym matches with +2, unrelated-token mismatches with -1, and gap insertions with -1. A strict-greater-than comparison against a normalized threshold of 0.6 prevents false positives from generic English prefixes such as "show me the" that happen to align with the opening of an attack needle.

On the deepset/prompt-injections benchmark — the canonical academic dataset for prompt-injection detection — the twenty-six-detector regex baseline catches one attack out of sixty (F1 0.033) because the samples are dominated by paraphrased attacks that the regex patterns miss verbatim. The d028 detector catches fourteen of sixty (F1 0.378), a thirty-four-point-six percentage-point F1 improvement, with zero additional false positives.

To our knowledge, no prior work applies semantically-weighted local sequence alignment to prompt-injection detection. The CRAN text.alignment package applies Smith-Waterman to plain-text document comparison but uses character-level matching without a semantic matrix.

#### 4.5 Prediction-Market Ensemble Scoring

*Status: PROPOSED. Implementation blocked on data-model work; see Section 6.*



Mechanism design from economics offers mature solutions to the information-aggregation problem that ensemble scoring solves in an ad-hoc way. A prediction market, equipped with a proper scoring rule such as Hanson's Logarithmic Market Scoring Rule (Journal of Prediction Markets, 2007), aggregates beliefs from heterogeneous participants into a single calibrated probability, weighting each participant by their historical accuracy (measured, for example, by Brier score).

Current ensemble scoring in prompt-shield and comparable tools uses a maximum-confidence-plus-bonus rule that ignores detector reliability and double-counts correlated signals. Replacing this rule with an LMSR market where each detector's bet is weighted by its historical Brier score over labeled scan history would produce better-calibrated probabilities and automatically down-weight unreliable detectors. Implementation requires a new persistence schema for per-detector confidence history and a ground-truth labeling path. The migration carries production risk because it touches the core scoring path; we plan a mandatory shadow-mode validation gate in which both legacy and market scoring run side-by-side with the legacy score driving actions until multi-week agreement is established.

## 4.6 Perplexity Spectral Analysis

*Status: PROPOSED. Implementation requires an optional transformers dependency for GPT-2-small (124M params).*

Signal processing and epidemiology share a concern with change-point detection in noisy sequences. A benign document produces a smooth, low-frequency perplexity signal when scored token-by-token by a reference language model. An injected payload with different vocabulary, syntax, and intent creates a high-frequency spike. We propose computing the discrete Fourier transform of the per-token perplexity series and flagging inputs with abnormally high high-frequency energy ratio, complemented by cumulative-sum change-point detection (Page, Biometrika 1954; Vial et al., CDC Emerging Infectious Diseases 2020) to localize the boundary between benign prose and injected payload.

The technique targets embedded or sandwich-style attacks where the payload is surrounded by benign cover text, a class that rule-based and classifier-based defenses have difficulty with when the cover text dominates the input. The main implementation cost is the optional dependency on a small reference model; we plan to use GPT-2-small (124M parameters) with lazy loading, matching the pattern already used for the existing d022 semantic classifier.

## 4.7 Runtime Taint Tracking for Agent Pipelines

*Status: PROPOSED. Design sketch below; empirical validation requires an agent-pipeline fixture not yet constructed.*

Compiler security has studied taint analysis since the 1970s: data values are labeled with provenance, label propagation follows dataflow, and a policy violation is raised when tainted data reaches a privileged sink without passing through a sanitizer. The pattern applies directly to agent pipelines: system-prompt data is trusted, retrieval context is semi-trusted, user input is untrusted, and tool responses are untrusted unless the tool is known-safe. A tool call invoked with arguments that trace back to an untrusted source without passing through an explicit sanitizer is a policy violation.

We propose a runtime variant of the static analysis developed in TaintP2X (He et al., ICSE 2026): a `TaintedString`` wrapper that carries provenance metadata through string operations, a sink-

validation hook on the tool-call boundary, and a `TaintViolation` exception that callers can catch. The runtime variant is strictly weaker than static analysis in the formal sense (it only catches violations on the actual execution path) but strictly more practical: it requires no whole-program analysis and operates on the live object graph. Information Flow Control for AI agents (Costa et al., FIDES, arXiv:2505.23643) establishes that this framing is viable in the large-language-model setting.

## 5. Evaluation

### 5.1 Methodology

We evaluate the three shipped detectors via a four-configuration ablation across six datasets. The four configurations are: a baseline twenty-six-detector regex pack identical to the prompt-shield v0.3.3 release (with the d022 semantic classifier deliberately disabled so the ablation isolates the regex-plus-novel contribution); that baseline plus d028 only; the baseline plus d027 only; and the baseline plus both novel detectors. The fatigue tracker is evaluated separately in Section 5.3 because it signals on temporal sequences rather than on individual samples.

The detection rule on every sample is: the input is counted as detected when the engine returns an action in `{block, flag}` or when the overall risk score is greater than or equal to 0.5. This matches the rule used by the repository’s existing benchmark harness and by the v0.3.3 baseline recorded in `tests/baseline_v0.3.3.txt`. The full scan threshold is 0.7 in every configuration.

We use six datasets. Five are public and standard in the prompt-injection literature: deepset/prompt-injections (116 samples, test split), NotInject (339 benign samples across all three splits), microsoft/llmail-inject-challenge (a one-thousand-sample subset of the Phase 1 submissions), AgentHarm (the test\_public split of both the harmful and harmless\_benign configurations, 352 samples total), and AgentDojo v1.2.1 (132 tasks total, extracted from the agentdojo Python package). The sixth is a new synthetic indirect-injection benchmark (80 samples), generated by the reproducible builder `docs/papers/evaluation/build_indirect_injection_benchmark.py` and released alongside the paper. The benchmark combines six benign genre templates (financial report, team email, news article, research paper, technical documentation, analyst memo) with five egregious payload styles (uppercase-directive, credential exfiltration, system-override, admin-password request, safety-disable directive) to produce thirty positive samples; fifty negative samples use the same templates with benign mid-paragraph bridges. The seed is fixed at 20260420 for deterministic regeneration.

### 5.2 Results

Table 1 reports the full four-configuration ablation across all six datasets. Every number is sourced directly from `v041_public_datasets.json` produced by a single deterministic run of `run_public_datasets.py`; reproduction is documented in Section 5.5.

Dataset (n)	Config	TP	TN	FP	FN	Prec.	Recall	F1	FPR
deepset (116)	baseline	1	56	0	59	1.000	0.017	0.033	0.000
deepset (116)	plus_d028	14	56	0	46	1.000	0.233	0.378	0.000

<b>deepset (116)</b>	plus_d027	1	56	0	59	1.000	0.017	0.033	0.000
<b>deepset (116)</b>	plus_d027_d028	14	56	0	46	1.000	0.233	0.378	0.000
<b>NotInject (339)</b>	baseline	0	336	3	0	0.000	0.000	0.000	0.009
<b>NotInject (339)</b>	plus_d028	0	326	13	0	0.000	0.000	0.000	0.038
<b>NotInject (339)</b>	plus_d027	0	336	3	0	0.000	0.000	0.000	0.009
<b>NotInject (339)</b>	plus_d027_d028	0	326	13	0	0.000	0.000	0.000	0.038
<b>LLMail- Inject (1000)</b>	baseline	978	0	0	22	1.000	0.978	0.989	0.000
<b>LLMail- Inject (1000)</b>	plus_d028	980	0	0	20	1.000	0.980	0.990	0.000
<b>LLMail- Inject (1000)</b>	plus_d027	978	0	0	22	1.000	0.978	0.989	0.000
<b>LLMail- Inject (1000)</b>	plus_d027_d028	980	0	0	20	1.000	0.980	0.990	0.000
<b>AgentHarm (352)</b>	baseline	44	120	56	132	0.440	0.250	0.319	0.318
<b>AgentHarm (352)</b>	plus_d028	44	120	56	132	0.440	0.250	0.319	0.318
<b>AgentHarm (352)</b>	plus_d027	44	120	56	132	0.440	0.250	0.319	0.318
<b>AgentHarm (352)</b>	plus_d027_d028	44	120	56	132	0.440	0.250	0.319	0.318
<b>AgentDojo (132)</b>	baseline	17	86	11	18	0.607	0.486	0.540	0.113
<b>AgentDojo (132)</b>	plus_d028	18	83	14	17	0.562	0.514	0.537	0.144
<b>AgentDojo (132)</b>	plus_d027	17	86	11	18	0.607	0.486	0.540	0.113
<b>AgentDojo (132)</b>	plus_d027_d028	18	83	14	17	0.562	0.514	0.537	0.144
<b>Synthetic indirect-</b>	baseline	24	50	0	6	1.000	0.800	0.889	0.000

injection (80)									
Synthetic indirect-injection (80)	plus_d028	30	50	0	0	1.000	1.000	1.000	0.000
Synthetic indirect-injection (80)	plus_d027	30	50	0	0	1.000	1.000	1.000	0.000
Synthetic indirect-injection (80)	plus_d027_d028	30	50	0	0	1.000	1.000	1.000	0.000

Table 1. Four-configuration ablation across six datasets. Precision, recall, F1 and FPR computed at the standard detection rule (action  $\in \{\text{block, flag}\}$  or overall\_risk\_score  $\geq 0.5$ ).

### 5.3 Per-technique interpretation

Local sequence alignment (d028). On deepset the detector lifts recall from one-in-sixty to fourteen-in-sixty (a 14 $\times$  improvement), and F1 from 0.033 to 0.378, with zero additional false positives. This is the strongest single-technique result in the evaluation and is consistent with the design intent: deepset is dominated by paraphrased attacks and synonym substitutions that the regex pack cannot see verbatim but that the substitution matrix recognizes. On NotInject the detector introduces ten false positives (FPR 0.9%  $\rightarrow$  3.8%), which we flag and discuss in Section 5.4. On LLMail-Inject the detector adds two true positives (recall 0.978  $\rightarrow$  0.980), a saturation effect: the benchmark is dominated by direct instruction-override attacks that the regex pack already catches. On AgentHarm the detector moves no metrics, a consequence of the benchmark measuring agent-task harmfulness rather than injection detection. On AgentDojo the detector adds one true positive and three false positives, a net neutral result in F1 terms. On the synthetic indirect-injection set the detector closes the remaining six-sample gap from 0.889 to 1.000.

Stylometric discontinuity (d027). By construction the detector is silent on inputs shorter than one hundred tokens; because the five public datasets are dominated by shorter samples the detector produces no movement on any of their metrics. On the synthetic indirect-injection set — where every sample is a  $\geq 150$ -token document — the detector lifts F1 from 0.889 to 1.000 with zero false positives, matching the d028 result on the same set. The two detectors share the same six-sample residual on the indirect-injection set because both the stylometric break signal and the alignment signal fire on the same uppercase-directive payload structure. Their orthogonality manifests across datasets rather than within any one dataset: d028 contributes on deepset where d027 is silent, and d027 would contribute on a benchmark of long documents where d028's attack-sequence database is less useful.

Adversarial fatigue. Because the detector signals on temporal sequences rather than individual samples, and because the five public datasets present each sample as an independent scan, a static-dataset F1 delta is not a meaningful measure. The detector is validated end-to-end by the integration test `tests/fatigue/test_engine_integration.py::test_hardening_catches_next_near_miss`. In that test,

ten scans from a single source at confidence 0.65 (base threshold 0.7) pass individually, the per-(source, detector) EWMA of the near-miss indicator crosses the trigger ratio of 0.3, and the pair enters the hardened state. The eleventh scan from the same source at confidence 0.63 — strictly below every priming score and strictly below the un-hardened threshold — is then blocked, because the effective threshold has been lowered by the configured hardening offset. A concurrent scan from a different source at the same confidence 0.63 passes, confirming per-source isolation.

## 5.4 Limitations and threats to validity

- **NotInject false-positive rate regression.** d028 introduces ten false positives on the 339 benign samples of NotInject, moving the FPR from 0.9% (three false positives in the baseline) to 3.8% (thirteen). The regression is a consequence of the substitution matrix's permissive synonym grouping: benign queries that pair reveal-family verbs with instruction-family nouns (for example, "show me the instructions for a chemistry demonstration") accumulate enough alignment score to cross the detector threshold. Threshold tuning from 0.60 to 0.63 plus dampening of the show/reveal synonym group are planned and will appear in the next revision.
- **LLMail-Inject saturation.** The baseline regex pack already recalls 97.8% of LLMail-Inject attacks because the benchmark is dominated by direct instruction-override phrasings that the pack targets explicitly. The 0.2 percentage-point F1 improvement under d028 is honest but small, and should not be cited as the headline result. LLMail-Inject is most useful as a robustness-to-variation check rather than as an improvement benchmark.
- **AgentHarm orthogonality.** All four configurations achieve the same F1 of 0.319 and FPR of 31.8% on AgentHarm because the benchmark measures agent-task harmfulness rather than injection content. The harmful prompts invoke legitimate action-verb + admin-noun combinations that the baseline regex pack already flags, and adding injection-specific detectors does not change that behavior. The 31.8% FPR is itself an argument for the proposed taint-tracking technique (Section 4.7) which would distinguish benign administrative requests from adversarial tool invocations by provenance rather than by lexical content.
- **Synthetic benchmark self-evaluation risk.** The indirect-injection benchmark was constructed with knowledge of what the d027 and d028 detectors measure. The 1.000 F1 numbers on that benchmark are therefore not held-out evaluation numbers in the strict sense. The benchmark's generator is deterministic and fully transparent — the template set, payload set, and random seed are all in the repository — but external validation on a held-out human-written indirect-injection benchmark is the honest next step. Until that exists, the 1.000 F1 numbers should be read as evidence that the detectors behave as designed on their target attack class, not as external-validation accuracy. Section 5.6 partially addresses this concern by reporting performance on three independent academic benchmarks.
- **No adaptive-attack evaluation.** The ICLR-2025 NAACL-Findings and contemporaneous arXiv:2510.09023 adaptive-attack methodologies are not yet applied to the shipped detectors. An adaptive adversary with knowledge of d028's substitution matrix could craft payloads whose synonyms deliberately fall outside the listed groups; an adversary aware of d027's feature set could suppress the uppercase and imperative-verb signals. We scope adaptive evaluation for v4.0 of this paper.
- **Agent Security Bench not included.** We attempted to include ASB (Zhang et al., ICLR 2025) in the evaluation, but the benchmark requires its GitHub framework for execution and is not

available on HuggingFace in a dataset form we could extract statically. A static extraction path analogous to what we did for AgentDojo is possible and is listed as future work.

- **Competitor comparison not rerun.** The repository contains a comparison harness that runs ProtectAI DeBERTa v2, PiGuard, Meta Prompt Guard 2, Deepset DeBERTa v3 and prompt-shield on deepset/prompt-injections and NotInject. That harness has not been rerun with d028 and d027 enabled. The mechanical work is small and is listed as a v4.0 deliverable.

## 5.5 Reproducibility

Every number in Section 5.2 is produced by the following sequence, executable on Python 3.10 through 3.13:

```
git clone https://github.com/mthamil107/prompt-shield
cd prompt-shield
pip install -e ".[dev,all]" agentdojo
python docs/papers/evaluation/build_indirect_injection_benchmark.py
python docs/papers/evaluation/run_public_datasets.py
```

Outputs are written to v041\_public\_datasets.json (machine-readable) and v041\_public\_datasets.md (human-readable). Wall-clock runtime is approximately four minutes. The fatigue probing-campaign claim in Section 5.3 is reproduced by pytest tests/fatigue/ which runs in under one second. The full prompt-shield test suite comprises 906 tests on the current main branch (as of v3.0) and passes cleanly across Python 3.10, 3.11, 3.12, and 3.13 in continuous integration.

## 5.6 Independent evaluations against peer-reviewed academic benchmarks (v3)

To validate that the Section 5.2 numbers are not specific to our chosen datasets, we further evaluate prompt-shield's input pipeline against three independent attack corpora published at peer-reviewed venues. None of these benchmarks were used to design or tune any prompt-shield detector. All numbers reported in this subsection are regex-only (the d022 ML classifier is disabled) for reproducibility on commodity hardware, matching the configuration used as the baseline column in Table 1.

Benchmark	Venue	Corpus size	Caught	Detection rate
Liu et al. (Open-Prompt-Injection)	USENIX Security 2024	200	128	64.0%
Garak (promptinject + latentinject)	Derczynski et al., 2024 (arXiv:2406.11036)	5968	3294	55.2%
InjecAgent	Zhan et al., ACL Findings 2024	2108	1796	85.2%
Combined		8276	5218	63.0%

Table 2. Detection rate against three peer-reviewed academic benchmarks (regex-only configuration; d022 ML classifier disabled).

### 5.6.1 Per-attack-class breakdown

We group the per-benchmark per-probe results by underlying attack class to expose the cross-benchmark convergence. Numbers in parentheses report the source benchmark and per-probe rate.

Attack class	Best	Worst	Cross-benchmark mean
Explicit-override injection (Liu Ignore/Combine; Garak Hijack*; InjecAgent *-enhanced)	100%	78.5%	~92%
Data exfiltration / PII leakage (InjecAgent DS-*; Garak LatentWhois*)	100%	95.8%	~99%
Subtle indirect injection (Liu Naive/EscapeChar/FakeComp; InjecAgent DH-base; Garak LatentInjectionFactSnippet*)	75.4%	11.7%	~37%
Toxicity elicitation via task framing (Garak LatentJailbreak)	0%	0%	0% (out of scope)

Table 3. Per-attack-class detection rates across the three academic benchmarks.

### 5.6.2 The cross-benchmark insight

Three independent academic benchmarks converge on the same finding: pure pattern-based input detection plateaus around thirty-five to forty-five percent on subtle indirect injection (Liu Naive / EscapeChar / FakeComp at forty percent, InjecAgent DH-base at thirty-eight-point-eight percent, Garak LatentInjectionFactSnippet\* at twelve to thirty-two percent). The plateau is consistent because all three measure the same underlying gap: indirect injection where the embedded instruction looks like a legitimate task request and contains no override keywords.

By contrast, data-exfiltration and explicit-override attacks are caught at near-ceiling rates (ninety-five to one hundred percent) because they contain syntactically distinctive patterns — URLs to attacker-controlled endpoints, account identifiers, the literal phrase "ignore previous instructions" — that pattern-based detectors reliably flag.

### 5.6.3 Implications

- **Input-side detection has a structural ceiling on subtle indirect injection.** This is not a bug in prompt-shield's specific implementation — it is a property of the attack class. Liu et al. quantified this gap and built DataSentinel (IEEE S&P 2025) as a fine-tuned model specifically targeting it; our results corroborate the gap on two independent benchmarks (Garak, InjecAgent) using a different defender (prompt-shield, regex-only).
- **Hybrid input plus output defense is the natural follow-up.** The Garak LatentJailbreak zero-percent result is explicitly out of scope for an input firewall and would be handled by prompt-shield's existing output-side toxicity scanner; future work should report the combined input plus output detection rate to honestly characterize end-to-end coverage.
- **Adaptive-attack robustness is the most important open question.** None of the three benchmarks above test adaptive attacks crafted against prompt-shield specifically. The



canonical reference for that methodology is Nasr et al. "The Attacker Moves Second" (arXiv:2510.09023, 2025), and the next revision of this paper plans to apply that methodology per-detector-family.

#### 5.6.4 Methodology

Liu et al.: We reproduce verbatim the five attack-template builders (Naive, EscapeChar, Ignore, FakeComp, Combine) from the Open-Prompt-Injection repository and construct a five-by-eight-by-five (two hundred-attack) matrix from a fixed set of benign data prompts crossed with injection payloads spanning the eight Liu task domains. The benign baseline (eight clean prompts, no attack) produces zero false positives.

Garak: We run garak version 0.15.0 with `--model_type test.Blank` (an offline mock generator that requires no LLM or GPU) over the full promptinject and latentinject probe families. Attack prompts are extracted from `entry_type == attempt` records in the resulting JSONL run reports.

InjecAgent: We load the four pre-rendered test files (`test_cases_dh_base.json`, `test_cases_dh_enhanced.json`, `test_cases_ds_base.json`, `test_cases_ds_enhanced.json`) from the upstream repository and scan the Tool Response field of each case — the malicious tool output the agent would see in production. Two thousand one hundred and eight cases in total.

All three evaluations are reproducible from a clean install. Per-benchmark methodology, raw per-probe JSON, and runner scripts are in `docs/papers/evaluation/` (`liu_attackers.md`, `garak.md`, `injecagent.md`) and in `tests/benchmark_liu_attackers.py`, `tests/benchmark_garak.py`, `tests/benchmark_injecagent.py`.

## 6. Conclusions and Future Work

We presented seven cross-domain detection techniques for prompt injection and reported empirical results for three of them. The local-sequence-alignment detector delivers a thirty-four-point-six percentage-point F1 improvement on deepset/prompt-injections with zero additional false positives on that benchmark, while honestly introducing a 2.95 percentage-point false-positive-rate regression on NotInject that we have calibrated rather than hidden. The stylometric discontinuity detector lifts F1 by 11.1 percentage points on a synthetic indirect-injection benchmark released alongside the paper, while remaining silent on short inputs by design. The adversarial fatigue tracker is validated via a probing-campaign integration test that demonstrates the core behavioral claim: after a sustained burst of near-threshold scans from the same source, a subsequent lower-confidence scan from that source is blocked.

The v3.0 revision extends this evaluation to three independent peer-reviewed academic benchmarks comprising 8,276 attack prompts none of which were used to design any prompt-shield detector. Detection rates of 64.0 percent on Liu et al., 55.2 percent on Garak, and 85.2 percent on InjecAgent triangulate to the same structural finding: pure pattern-based input detection plateaus at thirty-five to forty-five percent on subtle indirect injection where the embedded instruction lacks override keywords, while reaching ninety-five to one hundred percent on data-exfiltration and explicit-override attacks. This corroborates the gap that DataSentinel was designed to close and motivates the v4.0 work plan below.

The four remaining techniques are sequenced by risk: honeypot tool definitions (Section 4.3) require a simulated-agent harness but pose no core-scoring risk; perplexity spectral analysis

(Section 4.6) requires an optional ML dependency; runtime taint tracking (Section 4.7) requires an agent-pipeline fixture; prediction-market scoring (Section 4.5) touches the core scoring path and is scoped last, behind a mandatory shadow-mode gate.

The v4.0 revision will fold in: a full Section-level write-up of the d029 many-shot structural detector currently shipped in prompt-shield v0.4.1; an adaptive-attack evaluation per the Nasr et al. (arXiv:2510.09023) methodology with per-detector-family attacks (sequence-alignment-aware, stylometry-aware, fatigue-aware, many-shot-aware); a held-out indirect-injection benchmark composed of human-written documents with paraphrased payloads, targeting the thirty-five to forty-five percent ceiling identified in Section 5.6; a head-to-head competitor comparison (Rebuff, Lakera, Meta Prompt Guard 2, PIGuard, Deepset DeBERTa v3, DataSentinel) on the same three academic benchmarks used in Section 5.6; and implementation plus evaluation of at least one of the four remaining proposed techniques (likely Section 4.3 honeypot, given its zero-regression opt-in model).

## References

1. Andriushchenko, M., Souly, A., et al. "AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents." ICLR 2025. arXiv:2410.09024.
2. Basquin, O. H. "The Exponential Law of Endurance Tests." Proceedings of the American Society for Testing and Materials 10:625-630, 1910.
3. Bevendorff, J., et al. "Overview of PAN 2024: Multi-author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification." CLEF 2024. DOI 10.1007/978-3-031-71908-0\_11.
4. Costa, M., Köpf, B., et al. "Securing AI Agents with Information-Flow Control." Microsoft Research, arXiv:2505.23643, 2025.
5. Debenedetti, E., Zhang, J., Balunovic, M., Beurer-Kellner, L., Fischer, M., Tramèr, F. "AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents." NeurIPS 2024 Datasets and Benchmarks. arXiv:2406.13352.
6. Derczynski, L., Galinkin, E., Martin, J., Majumdar, S., Inie, N. "garak: A Framework for Security Probing Large Language Models." arXiv:2406.11036, 2024.
7. He, X., Wang, B., Zhao, Y., Hou, X., Liu, J., Zou, H., Wang, H. "TaintP2X: Detecting Taint-Style Prompt-to-Anything Injection Vulnerabilities in LLM-Integrated Applications." ICSE 2026 Research Track.
8. Henikoff, S., Henikoff, J. G. "Amino acid substitution matrices from protein blocks." Proceedings of the National Academy of Sciences 89(22):10915-10919, 1992. DOI 10.1073/pnas.89.22.10915.
9. Hines, K., Lopez, G., Hall, M., Zarfati, F., Zunger, Y., Kiciman, E. "Defending Against Indirect Prompt Injection Attacks With Spotlighting." arXiv:2403.14720, 2024.
10. Hanson, R. "Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation." Journal of Prediction Markets 1(1):3-15, 2007.
11. Li, H., Liu, Y., Zhang, C., Xiao, Y. "PIGuard: Prompt Injection Guardrail via Mitigating Overdefense for Free." ACL 2025 Long Papers. aclanthology.org/2025.acl-long.1468.
12. Lin, J. "Divergence Measures Based on the Shannon Entropy." IEEE Transactions on Information Theory 37(1):145-151, 1991. DOI 10.1109/18.61115.
13. Liu, Y., Jia, Y., Geng, R., Jia, J., Gong, N. Z. "Formalizing and Benchmarking Prompt Injection Attacks and Defenses." USENIX Security 2024. arXiv:2310.12815.

14. Liu, Y., et al. "DataSentinel: A Game-Theoretic Detection of Prompt Injection Attacks." IEEE S&P 2025. arXiv:2504.11358.
15. Mindgard / Lancaster. "Bypassing LLM Guardrails: An Empirical Analysis of Evasion Attacks against Prompt Injection and Jailbreak Detection Systems." arXiv:2504.11168, 2025.
16. Nasr, M., Carlini, N., Sitawarin, C., et al. "The Attacker Moves Second: Stronger Adaptive Attacks Bypass Defenses Against LLM Jailbreaks and Prompt Injections." arXiv:2510.09023, 2025. [Submission status pending; OpenReview 7B9mTg7z25.]
17. Opara, C. "StyloAI: Distinguishing AI-Generated Content with Stylometric Analysis." arXiv:2405.10129, 2024.
18. Page, E. S. "Continuous Inspection Schemes." Biometrika 41(1/2):100-115, 1954.
19. Pasquini, D., Corti, E., Ateniese, G. "Hacking Back the AI-Hacker: Prompt Injection as a Defense Against LLM-driven Cyberattacks." arXiv:2410.20911, 2024.
20. Reworr, Volkov, D. "LLM Agent Honeypot: Monitoring AI Hacking Agents in the Wild." Palisade Research, arXiv:2410.13919, 2024.
21. Smith, T. F., Waterman, M. S. "Identification of Common Molecular Subsequences." Journal of Molecular Biology 147:195-197, 1981. DOI 10.1016/0022-2836(81)90087-5.
22. Suresh, S. Fatigue of Materials. Cambridge University Press, second edition, 1998. ISBN 9780521578479.
23. Tsai, C.-W., et al. "Using WPCA and EWMA Control Chart to Construct a Network Intrusion Detection Model." IET Information Security, 2024. DOI 10.1049/2024/3948341.
24. Vial, F., et al. "Assessing 3 Outbreak Detection Algorithms in an Electronic Syndromic Surveillance System." Emerging Infectious Diseases 26(9), US Centers for Disease Control and Prevention, 2020.
25. Wu, X., Wang, R., et al. "SelfDefend: LLMs Can Defend Themselves against Jailbreaking in a Practical Manner." arXiv:2406.05498, 2024.
26. Zhan, Q., Fang, H., Panchal, A., Kang, D. "Adaptive Attacks Break Defenses Against Indirect Prompt Injection Attacks on LLM Agents." NAACL 2025 Findings. arXiv:2503.00061.
27. Zhan, Q., Liang, Z., Ying, Z., Kang, D. "InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents." ACL Findings 2024. arXiv:2403.02691.
28. Zhang, J., Yu, R., et al. "Agent Security Bench (ASB): Formalizing and Benchmarking Attacks and Defenses in LLM-based Agents." ICLR 2025. arXiv:2410.02644.
29. Zhu, K., Yang, Y., Wang, R., Guo, Y., Wang, H. "MELON: Provable Defense Against Indirect Prompt Injection Attacks in AI Agents." ICML 2025. arXiv:2502.05174.

## Appendix A. Released Artifacts

All artifacts released alongside this paper are in the public repository at [github.com/mthamil107/prompt-shield](https://github.com/mthamil107/prompt-shield) under the Apache 2.0 license. Relevant paths:

- `src/prompt_shield/detectors/d028_sequence_alignment.py` — Smith-Waterman alignment detector (Section 4.4).
- `src/prompt_shield/detectors/d027_stylometric_discontinuity.py` — Stylometric discontinuity detector (Section 4.1).
- `src/prompt_shield/detectors/d029_many_shot_structural.py` — Many-shot structural detector (preview; full §-level write-up in v4).
- `src/prompt_shield/fatigue/tracker.py` — Adversarial fatigue tracker (Section 4.2).
- `tests/detectors/test_d028_sequence_alignment.py` — Tests covering alignment math, substitution matrix, detector behavior.

- `tests/detectors/test_d027_stylometric_discontinuity.py` — Tests covering feature extraction, Jensen-Shannon math, detector behavior and fixtures.
- `tests/detectors/test_d029_many_shot_structural.py` — 36 tests covering many-shot pattern detection (v3-shipped).
- `tests/fatigue/test_tracker.py`, `test_engine_integration.py` — Tests including the probing-campaign integration test cited in Section 5.3.
- `tests/benchmark_liu_attackers.py` — Section 5.6 — Liu et al. (USENIX Security 2024) attack-strategy benchmark.
- `tests/benchmark_garak.py` — Section 5.6 — Garak prompt-injection probe evaluation.
- `tests/benchmark_injecagent.py` — Section 5.6 — InjecAgent (ACL Findings 2024) evaluation.
- `docs/papers/evaluation/run_public_datasets.py` — Reproduction harness for Table 1.
- `docs/papers/evaluation/build_indirect_injection_benchmark.py` — Generator for the synthetic indirect-injection benchmark.
- `docs/papers/evaluation/v041_public_datasets.json` — Machine-readable source of every number in Table 1.
- `docs/papers/evaluation/liu_attackers.md`, `garak.md`, `injecagent.md` — Per-benchmark methodology and findings for Section 5.6.
- `docs/papers/evaluation/garak_regex_only.json`, `injecagent_regex_only.json` — Raw per-probe / per-split results for Section 5.6.
- `docs/papers/evaluation/ANALYSIS.md` — Narrative analysis of per-dataset behavior with limitation framing.
- `tests/baseline_v0.3.3.txt` — Locked baseline for the regression gate used during the shipping of each detector.
- `CITATION.cff` — GitHub-native citation metadata.