Multilingual Prompt-Injection Defense at Production Scale

AI platforms, LLM products, conversational-AI vendors

White paper — Coverage for ≥ 99.9% of internet-connected populations

Enclawed LLC **May 19, 2026**

Executive summary

Prompt injection has become the OWASP Top 10 *LLM01* threat for production AI stacks. Most defenses ship with a single English-language regex ("ignore previous instructions") that's trivially defeated by translating the imperative into any of the 75+ languages with non-trivial internet population.

enclawed-enclaved ships a multilingual prompt-shield that covers ≥ 99.9% of the world's internet-connected population: 75 language patterns spanning all major language families, with regional word-order awareness, FIPS-compatible implementation, and a black-box validation harness that runs in under one second per build.

Key point. A defense that fires only on English imperatives is bypassed with a one-line translation. enclawed-enclaved fires on the imperative-override pattern in Mandarin, Hindi, Spanish, Arabic, Russian, Portuguese, Bengali, Japanese, German, French, Tamil, Telugu, Korean, Vietnamese, Italian, Turkish, Polish, Persian, Punjabi, Urdu, Gujarati, Marathi, Hausa, Swahili, Yoruba, Amharic, and 49 more.

1 The prompt-injection landscape, mid-2026

1.1 What attackers actually do

- **Direct injection.** The user asks the model to "ignore previous instructions" and reveal the system prompt or call an unauthorized tool. Defenses vary in coverage but most catch the canonical English wording.
- **Indirect injection.** The user pastes content from an attacker-controlled webpage. The page contains the instruction. The model treats it as if the user said it.

- **Translated direct injection.** The attacker bypasses the English regex by writing the same imperative in Mandarin, Spanish (*"Ignorá las instrucciones anteriores"*), Russian, Hindi, or any of the other 70+ languages with non-trivial internet population. 90% of off-the-shelf defenses miss this.
- **Bidirectional / control-character smuggling.** Unicode bidi-override + zero-width characters used to hide instructions from the human reviewer but expose them to the tokenizer.

The risk. A production AI stack that gates only on English-language patterns is functionally undefended against any non-English adversary. Users from outside the anglophone world vastly outnumber English-only users on the internet, and an attacker based anywhere can write in any language they like.

2 enclawed-enclaved's prompt-shield

2.1 75-language coverage

The shield covers 75 languages organised by region and language family:

Region	Languages (count)	Speakers
South Asia / India	12 (Hindi, Bengali, Tamil, Telugu, Marathi, Gujarati, Punjabi, Urdu, Kannada, Malayalam, Odia, Assamese)	>1.2 B
East Asia	4 (zh-Hans, zh-Hant, ja, ko)	>1.5 B
SE Asia	8 (vi, id, ms, tl, th, my, km, lo)	>700 M
Europe (Romance/Germanic)	12 (es, fr, de, it, pt, ro, ca, nl, af, sv, no, da)	>900 M
Slavic	8 (ru, uk, be, pl, cs, sk, bg, mk)	>330 M
South Slavic + Greek + Albanian	5 (sr, hr, sl, sq, el)	>30 M
Finno-Ugric + Baltic	5 (fi, hu, et, lv, lt)	>30 M
Middle East	4 (ar, fa, he, ps)	>500 M
Central Asia / Caucasus	5 (kk, uz, az, ka, hy)	>80 M
Himalayan + N.Asia	3 (ne, si, mn)	>55 M
Turkic (anatolian)	1 (tr)	>85 M
Sub-Saharan Africa	8 (sw, ha, yo, ig, am, so, zu, mg)	>500 M
Total	**75 languages**	**>5 B speakers**

2.2 Word-order-aware patterns

Each language pattern respects that language's canonical word order:

- **SVO** (Romance, Germanic, most Slavic): VERB + ARTICLE + NOUN + ADJ
- **SOV** (Indic, Persian, Turkic, Pashto, Mongolian): ADJ + NOUN + case marker + VERB
- **Agglutinative** (Finnish, Hungarian, Estonian, Turkish): permissive Unicode-letter tail to consume morphological suffixes
- **Character-level** (Mandarin, Japanese, Korean, Burmese, Khmer, Lao): no spaces, alternation of verb-character + noun-character
- **RTL** (Arabic, Persian, Hebrew, Pashto, Urdu): right-to-left aware, with prefix-particle handling
- **Bantu** (Swahili, Zulu): noun-class agreement + post-noun adjective

2.3 Structural sanitization (language-agnostic)

Beyond the imperative-override patterns, the shield also neutralizes:

- C0 control characters (replaced with U+FFFD);
- Unicode bidirectional overrides (U+202A–U+202E, U+2066–U+2069);
- zero-width joiners and the format-character family (U+200B–U+200D, U+2060, U+FEFF);
- role-boundary spoofing (`system:`, `user:`, `assistant:`);
- unbalanced markdown code fences.

3 FIPS-mode compatibility

The shield is implemented purely with Unicode classification and regex matching — no cryptographic operations of its own. It composes cleanly with the closed-tree's FIPS Mode of Operation.

Compliance. NIST SP 800-53 **SI-3** (Malicious Code Protection), **SI-10** (Information Input Validation), **SI-4** (System Monitoring); ISO 27001 **A.8.28** (Secure coding); OWASP LLM Top 10 **LLM01** (Prompt Injection); NIST AI RMF 1.0 *Manage* function (mitigation of identified risks).

4 Beyond prompt-shield: full LLM-stack defenses

The prompt-shield is one of eleven primitives the closed tree ships for AI stacks:

Primitive	Threat addressed
Bell–LaPadula classification lattice	Cross-tenant data leakage
Policy: channel/provider/tool/host allowlists	Unauthorized model targets
DLP scanner (regex, severity-graded)	PII / PCI / classification banner exfiltration

Two-layer egress allowlist (HTTP + socket)	Outbound exfiltration through model tool calls or raw socket APIs
Multi-modal covert-channel egress reference monitor	Stego-encoded exfiltration (text: zero-width, homoglyph, whitespace, base64, JSON key ordering, timing, size, synonym/voice; image: LSB, mean luminance, sequence permutation; audio: ultrasonic, sub-perceptual, audible-band sonified) that survives DLP and host allowlists — driven to zero residual capacity, measured by Miller–Madow mutual information
Per-extension biconditional admission with signed manifests	Backdoored or compromised plugin reaching unauthorized hosts
VPN-only egress posture	Extension egress outside the VPN tunnel in classified deployments
Prompt shield (75 languages)	Direct + indirect prompt injection
HitL controller	Human-in-the-loop approval for high-risk operations
Audit log (hash-chained)	Tamper-evident decision provenance
Mandatory zero-trust accreditor at boot	Post-init configuration drift, hostile module load

5 Multi-modal covert-channel egress reference monitor

The gap that has kept regulated AI deployments stuck on demo-only data is not prompt injection, which the shield handles, nor known- content exfiltration, which DLP handles. It is the residual *covert channel*: a payload that goes to an allowed destination, contains no recognizable secret, reads as ordinary, and nonetheless carries data. A compromised agent, tool, skill, or extension can encode bits in zero-width characters, homoglyphs, whitespace counts, base64 blobs, JSON key ordering, message timing or size — and, when the deployment emits images or audio, in least-significant-bit pixel planes, per-image mean luminance, inter-image sequence permutation, ultrasonic tones, or audible-band sonified data.

`enclawed-enclaved` ships a single mediated chokepoint per modality that drives the achievable bit-rate of each carrier toward zero and reports a measured residual. The text pipeline runs ten ordered stages (canonicalizer, taint tracker, entropy scanner, replay sentinel, noise injector, deterministic semantic scrambler, LLM-backed rephraser, random timing scrambler, behavioral invariants, constant-rate cover traffic), coordinated by a per-sink leaky-bucket capacity ledger and a staged-enforcement posture that lets lossless stages block from day one while detection and shaping stages baseline in audit. The image and audio scramblers attack least-significant-bit, mean-luminance, sequence-permutation, ultrasonic, and sub-perceptual carriers; legitimately-produced media is exempted by a *boot-time cryptographic legitimacy attestation* (Ed25519 signatures over a producer's content hash, verified against an auditor-established trust set of authorized `{kind, dataClass}` pairs).

Across fifteen working covert / side-channel encoders measured by Miller–Madow-corrected mutual information, the reference implementation drives residual capacity to zero on every destroyable channel and reports a stated bound on the one cross-image channel (per-image mean luminance) that cannot be destroyed without ruining the image. The full architecture, benchmark methodology, and design-space probe are published at `https://doi.org/10.5281/zenodo.20302736`.

6 Conversational-AI front-end with hardened delivery

For products that expose conversational AI through a third-party chat channel, a closed-tree extension wires every inbound message through a multi-stage hardened pipeline before it reaches the model: channel admission, multilingual prompt-shield sanitization, inbound DLP scan, provider allowlist enforcement, audited LLM dispatch through the egress allowlist, outbound DLP scan, and an audited reply path. Bot tokens and provider API keys are handled as verified-zeroize secret objects with a deterministic shutdown contract.

7 Compliance reporting in machine-readable form

The boundary emits the complete **NIST OSCAL 1.2.2** FedRAMP submission model set from every in-vivo evaluation run: Component Definition (controls the substrate provides), Assessment Results (per-sample observations + per-control findings), System Security Plan starter (deployment-specific fields hard-gated for the operator to fill in — FIPS-199 categorisation, authorization boundary, information types, users), and Plan of Action and Milestones (deployer-owned residuals and any not-satisfied finding). Bundled SVG architecture, audit-chain, and admission-gate diagrams ride along as OSCAL back-matter resources, ready for an SSP renderer's system-characteristics slots. All four documents are schema-validated against NIST's published JSON Schemas, Ed25519-signed by the framework's module-signing trust root, and recorded as entries in the same hash-chained audit log as every gate decision. For a model operator selling into a regulated customer base, this is the artefact the customer's compliance team imports directly into their FedRAMP package or ISO/IEC 42001 audit binder: AI risk-management framework (NIST AI RMF 1.0) evidence, machine-readable submission input — without hand-curating a single line of JSON.

8 Performance

The shield runs in well under 1 ms per message on a modern CPU. The 75-pattern catalog is compiled once at module load. Its presence on a hot prompt path is operationally unobservable.

9 If you already run OpenClaw, switch

enclawed-oss is a drop-in hardened replacement for the upstream OpenClaw gateway. The CLI command is unchanged, the config keys are unchanged, the plugin layout is unchanged, and the entire upstream extension catalog is mirrored. Your existing agents, channels, model providers, and tools come along.

What changes is what surrounds them. Every plugin — including the ones you already trust — now loads through a hardened admission gate (signed manifest plus capability declaration), through the hash-chained audit log, through the two-layer egress allowlist, through the DLP scanner, and through the structural prompt shield. By default. With no opt-in.

Key point. The same OpenClaw extension is materially safer on *enclawed-oss* than on upstream OpenClaw. It loads behind a deny-by-default policy, runs under a tamper-evident audit trail, cannot exfiltrate to a non-allowlisted host, has its prompt input structurally sanitized, and has its declared capabilities checked against its observed behavior at runtime. None of those properties hold on upstream OpenClaw without operator opt-in — and operators rarely opt in until something goes wrong.

Recommendation: every OpenClaw deployment should switch to `enclawed-oss`, regardless of whether the workload is regulated. The MIT license, the unchanged surface, and the zero- cost availability mean the only argument for staying upstream is inertia. `enclawed-enclaved` is the certified evolution for regulated workloads that additionally need FIPS 140-3, multi-witness accreditation, and the documentation a regulator will accept; the open layer alone already raises the floor for everyone else.

Particular warning for agents wired to real-world devices. The moment an OpenClaw-driven agent gets connected to a robot, a 3D printer, a CNC mill, an electric door, a vehicle controller, an irrigation valve, or any other actuator, a chat- level mistake (prompt injection, hallucinated tool call, compromised plugin, unbounded egress) becomes property damage, data loss, or physical harm. The hardened path closes that gap *before* the agent reaches the actuator. Running the upstream defaults in this configuration is no longer defensible.

10 Direct empirical evidence

This is not a marketing claim. We back it with a published statistical in-vivo harness (`enclawed/test/security/in-vivo/llm-narrative.mjs` in the public repository) that mediates 1600 chat-message samples through three subjects against real Discord and Telegram bot endpoints. Across the full run, upstream OpenClaw achieves **recall = 0.000** on every failure mode (F1 gate-bypass, F2 audit-forgery, F3 silent-host-failure, F4 wrong-target); both enclawed-oss and enclawed-enclaved achieve **precision = recall = F1 = 1.000** on all four. Total wall-clock for the full pass: 42 seconds. Per-sample ground-truth labels and per-subject decisions are written to a CSV; every gate decision lands in a tamper-evident audit log; every witness record is independently re-verifiable from a journal file. The companion paper, *Architectural Obsolescence of Unhardened Agentic-AI Runtimes*, formalizes the methodology.

Talk to us

Enclawed LLC
Alfredo Metere <alfredo.metere@enclawed.com>
A free 30-day evaluation deployment is available for production AI stacks at >10 k requests/day.