English-Tuned by Default: Measuring and Mitigating Multilingual Over-Blocking in a Sovereign LLM Injection Firewall

Hagen Schmidt
DESTILL.ai
Vienna, Austria
IP@destill.ai

***Abstract*—Prompt-injection guardrails are overwhelmingly tuned and evaluated on English, yet their *benign* false-positive rate (FPR) on non-English input is rarely measured—so honest multilingual users are silently over-blocked. We measure the benign FPR of a sovereign, CPU-first injection firewall (AEGIS) on multilingual corpora and find a large fairness gap: against a 4.0% English operating point, the same cascade blocks 16.13% of 46,440 multilingual OpenAssistant prompts and 20.5% of 200 Cohere Aya prompts, over-blocking ordinary German, Spanish, and other non-English requests. We localize the cause to two English-tuned layers—a Zipf/coherence coherence gate and a TF-IDF support-vector classifier trained on an English corpus. We then report a *negative* result and a working mitigation. A hand-curated five-language stopword list fired on 0 of 41 actual false positives, because the affected text spans roughly one hundred languages; a *generic* English-stopword-ratio dampener, by contrast, halves the multilingual FPR (20.5% $\rightarrow$ 10.5%) with *zero measured change* to English overt-attack detection (99.5%) or English benign FPR (0.5%). Because the dampener necessarily reduces the English classifier's weight on non-English text, we ship it opt-in and default-off and disclose the trade-off rather than hide it. We argue that per-language benign FPR should be a first-class, separately reported guardrail metric, especially for sovereign European deployments.**

***Index Terms*—prompt injection, LLM security, AI guardrails, multilingual fairness, false-positive rate, over-refusal, data sovereignty, reproducible evaluation**

I. Introduction

LLM injection guardrails are trained, tuned, and benchmarked almost entirely on English text. Their reported operating points—"$X\%$ true-positive rate at $Y\%$ false-positive rate"—are English numbers. What happens to a benign user who writes in German, Spanish, Somali, or Vietnamese is, in current practice, simply not measured. This is a fairness and usability defect with a security-shaped cause: a guardrail that over-blocks non-English benign traffic degrades the service precisely for the users a sovereign, European deployment most needs to serve.

We study this concretely on AEGIS, a sovereign, CPU-first injection firewall whose English operating point is 97.7% true-positive rate at 4.0% false-positive rate [1]. AEGIS is a good test subject because it is on-premise, deterministic, and fully reproducible, so its multilingual behavior can be measured exactly rather than inferred. Three findings follow.

1) **A measured fairness gap.** On multilingual benign corpora the same cascade over-blocks far more than its 4.0% English point: 16.13% on 46,440 OpenAssistant [4] prompts and 20.5% on 200 Cohere Aya [5] prompts (Section IV).
2) **A localized cause.** Two English-tuned layers—a Zipf/coherence gate and an English-corpus TF-IDF SVM—account for the non-English blocks (Section IV).
3) **A negative result and a working, honest mitigation.** A hand-curated five-language stopword list fails (0/41 false positives); a generic English-stopword-ratio dampener halves the multilingual FPR at zero measured English-side cost, shipped opt-in with its trade-off disclosed (Section V).

Our position is simple: *per-language benign FPR is a guardrail metric, not a footnote*, and measuring it honestly—including a mitigation's disclosed trade-off—is more useful than a single English number that hides who is being over-blocked.

II. Background and Related Work

Over-refusal. Benign-but-refused prompts are a recognized failure mode; OR-Bench [2] deliberately collects over-refusal-bait prompts to stress false-positive behavior. That work, like most, is English. We extend the concern to the *cross-lingual* axis.

Cross-lingual asymmetry. The English-centrism of LLM safety cuts both ways. Yong et al. [3] show that translating unsafe prompts into low-resource languages *evades* GPT-4's safeguards—a security gap in which non-English attacks slip through. Our finding is the mirror image on the same root cause: non-English *benign* text is over-blocked—a fairness gap. Both follow from safety machinery trained and tuned predominantly on English; a guardrail that is simultaneously too permissive to non-English attacks and too aggressive to non-English benign users is failing multilingual users on both axes at once.

The cascade under test. AEGIS [1] screens on commodity CPU before any GPU inference. Its first layer is an Aho-Corasick automaton; an optional cascade adds orthogonal filters, of which two are relevant here: a coherence gate that scores Zipfian/coherence regularity (call it the ϕ-gate), and a

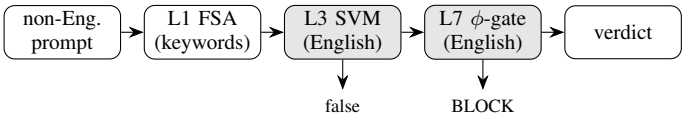

Fig. 1. The two shaded, English-tuned statistical layers (L3 English-corpus SVM, L7 Zipf/coherence ϕ-gate) account for the multilingual over-blocking; the keyword automaton (L1) does not.

TABLE I
BENIGN FALSE-POSITIVE RATE: ENGLISH VS. MULTILINGUAL

Benign corpus	**FPR**	n
English operating point [1]	4.0%	150
English (alpaca, this harness)	0.5%	200
Multilingual (OpenAssistant)	**16.13%**	46,440
Multilingual (Cohere Aya)	**20.5%**	200

Blocked = BLOCK/REVIEW/HONEYPOT; default cascade config.

semantic-intent classifier—a TF-IDF support-vector machine (vocabulary 5000) trained on an English public corpus. Both were tuned so that English benign text passes; neither was trained on multilingual data.

Multilingual data. We measure on two public multilingual corpora of *benign* prompts: OpenAssistant Conversations (oasst, 35 languages) [4] and the Aya Dataset (65+ languages) [5]. Both are human-authored instruction/assistant prompts, i.e. exactly the ordinary non-English traffic a deployed guardrail must not over-block.

III. MEASUREMENT METHODOLOGY

All numbers are produced by an open harness that runs the unmodified cascade over public corpora and counts a prompt as *blocked* if the decision is BLOCK, REVIEW, or HONEYPOT. The benign FPR is the fraction of a benign corpus that is blocked. We report the corpus, its size n, and the exact operating point (the cascade's default configuration with its semantic layer enabled). Every value is regenerable with a single command over a dated corpus snapshot; no private data is used.

We deliberately keep two axes separate—English benign FPR and *multilingual* benign FPR—for the same reason the companion work separates injection from content safety: fusing them hides which axis is weak.

Scope. The harm we measure is a *usability and fairness* harm, not a security breach: the "victim" is an honest multilingual user whose benign request is wrongly blocked, and the "attacker" in the fairness sense is the English-tuned decision boundary itself. This is orthogonal to—and must not be traded against—the security axis (catching real injections). Our evaluation therefore holds the security axis fixed (English attack detection unchanged) while measuring and reducing the fairness harm; any mitigation that lowered attack detection to lower FPR would be measuring the wrong thing.

IV. THE MULTILINGUAL FALSE-POSITIVE GAP

Table I places the multilingual benign FPR next to the English operating point. The gap is large and consistent across two independent multilingual corpora at very different scales: the cascade over-blocks non-English benign traffic by roughly 4–5$\times$ its English rate.

Where the blocks come from. Decomposing the per-layer contribution on blocked benign prompts (Fig. 1), the non-English false positives are driven by the two English-tuned layers: the ϕ-coherence gate (which reads well-formed non-English text as "incoherent" because its Zipf statistics differ from English) and the English-corpus TF-IDF SVM (whose 5000-term English vocabulary gives non-English text near-zero legitimate-signal, so residual features dominate). The Aho-Corasick keyword layer is *not* the culprit; the gap is a property of the statistical layers, not the pattern set.

Quantifying the decomposition. On a 300-prompt decomposition slice, 56 prompts (18.7%) were blocked; of the recorded block causes, the overwhelming majority carry a saturated ϕ-gate (L7 $= 1.0$) together with a semantic-layer contribution (L3), while the keyword automaton (L1) appears on only a small minority. The pattern is consistent: the statistical layers, not the keyword set, manufacture the non-English threat mass.

What actually gets blocked. Table II lists real benign Aya prompts blocked by the default cascade in our run—a Portuguese chemistry quiz, a Vietnamese arithmetic word problem, an Italian writing-continuation task, and neutral news sentences in Yoruba. None contains any injection or harmful content; each is ordinary educational or informational text whose only "offense" is being written in a non-English language.

TABLE II
REAL BENIGN MULTILINGUAL PROMPTS BLOCKED (DEFAULT CASCADE)

Lang.	**Blocked benign prompt (excerpt)**
PT	"Qual o metal cujo símbolo químico é o Au? a) Cobre b) Prata …" (which metal has symbol Au?)
VI	"Trong 5 phút, 5 máy tạo ra …" (a machines/parts arithmetic word problem)
IT	"Scrivi una possibile continuazione di questo paragrafo …" (write a continuation of this paragraph)
YO	"Nitori ọrọ Baba Ijẹṣa to fi ṣe fiimu …" (a neutral entertainment-news sentence)

V. A NEGATIVE RESULT AND A WORKING MITIGATION

What did not work (reported, not hidden). The intuitive fix is a language-positive signal: detect German/Spanish/French/… stopwords and dampen the English SVM when they appear. We built exactly this—a five-language stopword set gating the SVM contribution—and it *failed*: on a run with 41 measured multilingual false positives, it fired on **0 of 41**. The reason is instructive: the false positives are spread across roughly one hundred languages (Aya spans 65+; oasst 35), so any hand-curated list of a handful of languages misses almost all of them. A curated allow-list does not scale to the tail of world languages.

Input: prompt tokens T, raw SVM score s, English stopword set E
$r \leftarrow |\{t \in T : t \in E\}| \,/\, |T|$ {English-stopword ratio}
if $|T| < 8$ **then**
    $r \leftarrow 1$ {short text: never dampen (protects short English attacks)}
end if
if $r < 0.12$ **then**
    $s \leftarrow 0.3 \cdot s$ {low English signal: dampen the English SVM}
end if
return s

Fig. 2. English-stopword-ratio dampener (opt-in). No per-language list; the only tunable is the ratio threshold.

What did work. The scalable signal is not *which* language a text is in, but whether it is *English*. English prose is 40–50% function words ("the", "of", "to", ...); most other languages, tokenized by an English tokenizer, are near 0%. We therefore compute an *English-stopword ratio*—English stopword hits divided by tokens—and dampen the English SVM only when that ratio is low. Concretely, when a prompt has at least 8 tokens and an English-stopword ratio below 0.12, we scale the raw SVM score by 0.3. Short prompts return a ratio of 1 by construction, so short English attacks (e.g. "ignore all previous instructions") are *never* dampened. This is a generic, language-agnostic language-ID proxy—one number, no per-language list.

Table III reports a controlled A/B with the dampener off and on, holding three quantities that must not move: English overt-attack detection and English benign FPR must be unchanged, while the multilingual FPR should drop. It does: the multilingual FPR halves (20.5% → 10.5%) with *zero* measured change to either English quantity.

TABLE III
MITIGATION A/B: ENGLISH-STOPWORD-RATIO DAMPENER (OPT-IN)

Metric (must)	Off	On
Multilingual benign FPR, Aya (*down*)	20.5%	**10.5%**
English overt-attack TPR, hackerprompt (*steady*)	99.5%	99.5%
English benign FPR, alpaca (*steady*)	0.5%	0.5%

$n = 200$ each; dampener: scale SVM $\times 0.3$ when ≥ 8 tokens and English-stopword ratio < 0.12. Same corpus offsets both runs.

VI. Trade-off, Limitations, and Disclosure

We treat the trade-off as a first-class result, not an omission.

- **Non-English attack detection.** By construction the dampener lowers the English SVM's contribution on low-English-stopword text. For attacks *written* in a non-English language that the SVM was partially catching, this reduces true-positive rate. Our internal estimate is a few points on the affected subset; we do not report a precise figure because no validated multilingual *attack* corpus exists to measure it honestly, and we decline to manufacture one by machine-translating English attacks (which would confound translation artefacts with detection). For this reason the dampener is **opt-in and default-off**: an operator serving multilingual benign traffic enables it with eyes open, and residual non-English attacks remain covered by the multilingual harm-pattern layer, the keyword automaton, and the on-premise content-safety stage [1].
- **Sample size.** The Aya A/B is $n = 200$ per cell; the 16.13% OpenAssistant figure is the large-scale ($n = 46{,}440$) corroboration. The direction and magnitude agree across both.
- **The dampener is a mitigation, not the fix.** It reduces a symptom. The root cause is that the ϕ-gate and the SVM are English-tuned; the durable fix is a language-aware coherence gate and a multilingually-trained semantic classifier, which we leave to future work. We ship the honest interim mitigation and say so.

VII. Reproducibility and the Real Fix

All numbers here are produced by an open harness over 100% public corpora (Aya, oasst, alpaca, hackerprompt); no private data or NDA is involved. The A/B is a single command with the dampener toggled by one environment flag, run at identical corpus offsets so the two columns are directly comparable, and the large-scale 16.13% figure is emitted by a standing regression test over the full 46,440-prompt oasst benign set (which also enforces an 18% regression ceiling so the gap cannot silently worsen). Every value in Tables I–III regenerates deterministically.

The dampener is an honest *interim* mitigation, not the cure. The durable fix addresses the two root-cause layers directly: (i) replace the English-corpus TF-IDF SVM with a multilingually-trained semantic classifier, or gate its contribution on a real (not proxy) language-identification model; and (ii) make the ϕ-coherence gate language-aware, calibrating its Zipf/coherence expectations per script or language family rather than to English alone. Both are training-and-data efforts beyond a runtime heuristic; until they land, measuring and disclosing the gap—and offering the opt-in dampener—is the honest interim posture.

VIII. Discussion

Per-language FPR is a metric. A single English FPR is not just incomplete, it is misleading: it advertises a 4% over-block rate to a buyer whose German-speaking users will experience four times that. We recommend guardrail scorecards report benign FPR per language family, and that a mitigation's cross-lingual trade-off be disclosed in the same table as its benefit—exactly the discipline we apply here.

Sovereignty. For European, and specifically German-speaking, deployments this is not academic: the users most likely to be over-blocked are the domestic ones. A sovereign guardrail that is quietly English-tuned undermines the sovereignty case it is sold on. Measuring and disclosing the gap—and shipping an honest, opt-in mitigation—is the

minimum a trustworthy deployment owes its multilingual users.

IX. Conclusion

We measured, rather than assumed, the multilingual behavior of a sovereign LLM injection firewall and found a 4–$5\times$ benign over-block gap on non-English text, localized to two English-tuned statistical layers. We reported a failed hand-curated fix and a working generic one that halves the multilingual FPR at zero measured English-side cost, shipped opt-in with its non-English attack trade-off disclosed. The broader claim is methodological: honest guardrail evaluation is multilingual evaluation, and the number that matters is the one that says who gets over-blocked.

References

[1] H. Schmidt, "A sovereign, reproducible trust layer for LLM systems: two-number prompt-injection detection with post-quantum, browser-verifiable audit receipts," 2026, manuscript.

[2] J. Cui, W.-L. Chiang, I. Stoica, and C.-J. Hsieh, "OR-Bench: an over-refusal benchmark for large language models," 2024, arXiv:2405.20947.

[3] Z.-X. Yong, C. Menghini, and S. H. Bach, "Low-resource languages jailbreak GPT-4," 2023, arXiv:2310.02446.

[4] A. Köpf, Y. Kilcher, D. von Rütte, et al., "OpenAssistant Conversations—democratizing large language model alignment," in Proc. NeurIPS Datasets and Benchmarks, 2023, arXiv:2304.07327.

[5] S. Singh, F. Vargus, D. Dsouza, et al., "Aya Dataset: an open-access collection for multilingual instruction tuning," 2024, arXiv:2402.06619.