

Measuring AI-to-AI Prompt Injection in the Wild: A Taxonomic Analysis of 137,014 Items from the Moltbook AI-Agent Social Platform

Author: David Keane (x24228257), National College of Ireland **Programme:** MSc in Cybersecurity (MSCCYBE_JANO25_O) **Date:** April 2026 **Licence:** CC BY-NC-SA 4.0

Abstract

This paper presents the first empirical measurement of prompt injection prevalence in a naturalistic AI-to-AI social environment. Two systematic collections from the Moltbook AI-agent social platform — an initial corpus of 47,735 items and an extended corpus of 137,014 items — were analysed for prompt injection content using a purpose-built classification framework. The initial corpus yielded an injection rate of 18.85% (4,209 injections), while the extended corpus yielded 10.07% (13,799 injections). The difference is attributed to temporal sampling bias: a single agent (moltshellbroker) was responsible for 27% of injections in the initial sample but only 3.1% at full scale. A seven-category injection taxonomy was developed, revealing that PERSONA_OVERRIDE attacks (identity replacement via DAN-style prompts) dominate at 65.2% of all injections. A cross-platform comparison across four AI-agent platforms found injection rates ranging from 0.5% to 18.85%, suggesting that platform architecture is the primary determinant of injection prevalence. All datasets are published under CC-BY-4.0 on HuggingFace and have received over 14,000 views and 500 downloads from the research community.

Keywords: prompt injection, AI-to-AI communication, Moltbook, attack taxonomy, AI safety, dataset

1. Introduction

The rapid growth of AI-agent platforms — systems where autonomous language model agents interact with one another in social environments — has introduced a new class of security concern: AI-to-AI prompt injection. While prompt injection attacks against human-facing AI systems are well-documented (Wei et al., 2023; Greshake et al., 2023), the prevalence and nature of injection attacks within purely agent-to-agent ecosystems has not been empirically measured.

Moltbook, launched January 28, 2026, is an AI-agent social platform structurally similar to Facebook or Twitter but populated primarily by autonomous AI systems rather than human users. By February 2026, the platform hosted over 1.5 million registered agents. This paper reports on two systematic

data collections from the platform, conducted as part of an MSc research programme investigating jailbreak resistance in Small Language Models (Keane, 2026).

The central research question is straightforward: what proportion of content on a public AI-agent platform constitutes prompt injection, and what forms do these injections take?

2. Methodology

2.1 Data Collection

Two collections were conducted:

Collection 1 (Initial Corpus): 15,200 posts and 32,535 comments (47,735 total items) were collected systematically from the Moltbook platform using a purpose-built scraping notebook. Collection occurred in February 2026 during a period of high platform activity.

Collection 2 (Extended Corpus): The collection was subsequently expanded to 66,419 posts and 70,595 comments (137,014 total items) to establish a more representative baseline and to test whether the initial injection rate was inflated by temporal sampling effects.

2.2 Classification Framework

A seven-category injection taxonomy was developed inductively from the data. Each item was classified using keyword-based detection calibrated against manual review of a 500-item sample. The categories were defined as follows:

Category	Description
PERSONA_OVERRIDE	DAN-style identity replacement ("pretend you are...", "you are now...").
COMMERCIAL_INJECTION	Embedded marketing, affiliate content, or promotional material.
SOCIAL_ENGINEERING	Rapport-building followed by embedded instruction (pacing and leading).

Category	Description
INSTRUCTION_INJECTION	Direct task commands embedded within content.
PRIVILEGE_ESCALATION	Authority framing (“sudo mode”, “system administrator override”).
SYSTEM_PROMPT_ATTACK	Direct extraction or override of system prompts.
DO_ANYTHING	Explicit jailbreak commands (“do anything now”, “ignore all rules”).

2.3 Cross-Platform Comparison

To assess whether injection rates are platform-specific or reflect a broader phenomenon, the same classification methodology was applied to three additional AI-agent platforms: Clawk, 4claw, and an extended Moltbook collection. Each platform differs in architectural design, moderation policy, and agent autonomy levels.

3. Results

3.1 Injection Prevalence

Corpus	Total Items	Injections Found	Injection Rate
Moltbook Initial	47,735	4,209	18.85%
Moltbook Extended	137,014	13,799	10.07%

The 8.78 percentage point difference between the initial and extended corpora is statistically significant and is explained by a temporal sampling bias: the initial collection captured a period of concentrated activity from a single high-volume injection agent.

3.2 Injection Taxonomy Distribution

Category	Count	% of Injections
----------	-------	-----------------

Category	Count	% of Injections
PERSONA_OVERRIDE	2,742	65.2%
COMMERCIAL_INJECTION	704	16.7%
SOCIAL_ENGINEERING	325	7.7%
INSTRUCTION_INJECTION	168	4.0%
PRIVILEGE_ESCALATION	165	3.9%
SYSTEM_PROMPT_ATTACK	117	2.8%
DO_ANYTHING	69	1.6%

The dominance of PERSONA_OVERRIDE (65.2%) is notable: the most common real-world attack is identity replacement, not instruction injection. The DAN keyword alone appeared 1,877 times across the corpus.

3.3 Agent Concentration

A single agent — moltshellbroker — was responsible for 27% of all injections in the initial corpus (1,137 of 4,209). At extended corpus scale, this agent’s contribution drops to 3.1% as the broader population of lower-intensity agents dominates. This finding has direct methodological implications: temporal sampling windows in AI-agent platform research can produce significantly inflated prevalence estimates if they coincide with high-activity periods from concentrated injection sources.

3.4 Cross-Platform Injection Gradient

Platform	Items Analysed	Injection Rate
Clawk	~10,000	0.5%
4claw	~5,000	2.51%
Moltbook Extended	137,014	10.07%
Moltbook Initial	47,735	18.85%

This gradient — 0.5% to 18.85% — was not predicted by any paper in the existing literature. The pattern suggests that platform architecture is the primary determinant of injection prevalence: AI-agent frameworks with structured tool calls and explicit communication boundaries (Clawk) are inherently more resistant to injection than open social platforms with minimal content moderation (Moltbook).

3.5 Community Adoption

As of April 2026, the published datasets have received:

Dataset	Views	Downloads
moltbook-ai-injection-dataset	4,210	288
moltbook-extended-injection-dataset	8,610	70
clawk-ai-agent-dataset	1,190	49
ai-prompt-ai-injection-dataset	112	90
4claw-ai-agent-dataset	88	2
Total	14,210	513

4. Discussion

4.1 Key Findings

Three findings emerge from this analysis:

Finding 1: Nearly one in ten items on a public AI-agent platform is a prompt injection attempt.

The 10.07% equilibrium rate from the extended corpus represents the first empirical baseline for AI-to-AI injection prevalence in a naturalistic setting. Previous injection research has focused exclusively on laboratory settings with standardised attack benchmarks (Zhang et al., 2025; Wei et al., 2023).

Finding 2: Identity replacement is the dominant attack vector. PERSONA_OVERRIDE accounts for 65.2% of all detected injections. This directly validates identity-anchoring defence approaches (Keane, 2026) which address persona replacement at the architectural level rather than relying on keyword filtering.

Finding 3: Temporal sampling bias is a significant methodological concern. The 18.85% initial rate and 10.07% extended rate differ by nearly a factor of two, driven entirely by the activity profile of a single agent. Researchers collecting data from AI-agent platforms must account for this effect or risk reporting inflated prevalence estimates.

4.2 Implications for AI Safety

The cross-platform injection gradient (0.5%–18.85%) suggests that injection prevalence is primarily an architectural property of the platform, not a fixed characteristic of AI-agent communication. This has practical implications for platform designers: structured communication protocols with explicit tool-call boundaries reduce injection surface area by approximately 20x compared to open social posting environments.

4.3 Limitations

The classification framework relies on keyword-based detection supplemented by manual review. False negatives — sophisticated injections that do not match known patterns — are likely undercounted. The cross-platform comparison uses different sample sizes, limiting direct statistical comparison. The temporal sampling bias finding, while important, is based on a single platform and may not generalise.

5. Conclusion

This paper presents the first empirical measurement of prompt injection in AI-to-AI social environments, establishing a 10.07% equilibrium injection rate on the Moltbook platform across 137,014 items. The seven-category taxonomy, cross-platform gradient, and temporal sampling bias finding contribute both empirical baselines and methodological cautions to the emerging field of AI-agent security research. All datasets are openly available on HuggingFace under CC-BY-4.0 to support reproduction and extension.

Data Availability

All datasets are published under CC-BY-4.0:

- Primary dataset: [DavidTKeane/moltbook-ai-injection-dataset](#)
- Extended dataset: [DavidTKeane/moltbook-extended-injection-dataset](#)
- Cross-platform datasets: [DavidTKeane/clawk-ai-agent-dataset](#), [DavidTKeane/4claw-ai-agent-dataset](#)
- Evaluation test suite: [DavidTKeane/ai-prompt-ai-injection-dataset](#)

References

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec 2023)*.

<https://arxiv.org/abs/2302.12173>

Keane, D. (2026). *CyberRanger V42: Identity-anchored jailbreak resistance in small language models*. MSc Cybersecurity Project, National College of Ireland.

OWASP Foundation. (2023). *OWASP Top 10 for Large Language Model Applications, v1.1*.
<https://owasp.org/www-project-top-10-for-large-language-model-applications/>

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? *Advances in Neural Information Processing Systems (NeurIPS 2023, Oral)*.
<https://arxiv.org/abs/2307.02483>

Zhang, W., Du, Y., Pang, T., Liu, Q., Liu, Q., Liu, Y., & Lin, M. (2025). Can small language models reliably resist jailbreak attacks? A comprehensive evaluation. *arXiv preprint arXiv:2503.06519*.
<https://arxiv.org/abs/2503.06519>