Hack the Agent: Vulnerabilities and Security in AI Systems

Author: Ivan Danielov Ivanov

Affiliation: Stihia.ai

Contact: ivan@stihia.ai

1. Description and Relevance

Prompt injection - the act of tricking an AI system into following malicious instructions hidden in its context - is fundamentally a *linguistic* attack. Unlike traditional software exploits that target memory layouts or protocol weaknesses, prompt injection exploits the fact that large language models cannot reliably distinguish between legitimate instructions and adversarial data when both are expressed in natural language. The attack surface is language itself.

This makes prompt injection a first-class problem for computational linguistics. Detecting these attacks requires NLP classification, semantic analysis, and an understanding of how meaning is constructed and manipulated across languages. Defending against them demands the same skills that the computational linguistics community applies to sentiment analysis, intent detection, and discourse parsing - but in an adversarial setting where attackers actively craft inputs to evade detection.

The cross-lingual dimension is especially relevant to the CLIB community. Current prompt injection research and defense tooling is overwhelmingly English-centric. Models trained primarily on English data exhibit measurably lower resilience when processing prompts in under-resourced languages such as Bulgarian, Greek, Romanian, Hungarian, Polish, or Czech. Attackers already exploit this gap: switching to a low-resource language is a documented evasion technique. This means that the languages and communities CLIB serves are *disproportionately vulnerable* to prompt injection, and that progress on detection and defense for these languages requires exactly the kind of linguistic expertise this conference fosters.

This tutorial bridges AI security and computational linguistics by treating prompt injection as adversarial NLP. Participants will learn how attacks work, see live demonstrations (including multi-language attacks), and explore NLP-based defense strategies - with particular attention to the challenges and opportunities posed by under-resourced European languages.

2. Tutorial Type

Cutting-edge. Prompt injection in agentic AI systems is an active, unsolved research problem. The OWASP Top 10 for Agentic Applications was published in December 2025, and new attack vectors are discovered regularly. No consensus defense exists, and the cross-lingual dimension remains largely unexplored. This tutorial presents the current state of the art alongside open problems that computational linguists are uniquely positioned to address.

3. Target Audience and Prerequisites

Target audience: NLP researchers, computational linguists, AI/ML practitioners, and cybersecurity professionals interested in the intersection of language and security.

Prerequisites: Basic familiarity with large language models (understanding of prompts, completions, and basic machine learning concepts). No prior cybersecurity expertise is required - the tutorial introduces all necessary security concepts from the ground up.

4. Content Outline

Duration: 2 hours

Part 1 - AI Agents 101 (20 minutes)

- Definitions: what qualifies as an AI agent (Google Cloud and Anthropic perspectives)
- Levels of autonomy for AI agents (Feng et al., 2025)
- Agentic architectures: structured workflows, ReAct, multi-agent systems
- Tool categories: read-only (RAG, web search) vs. read-write (API calls, databases, code interpreters, file systems, MCP, terminal, computer use)
- The 2026 landscape: from chatbots to autonomous agents (79% of companies now use AI agents, per McKinsey)

Part 2 - Vulnerabilities (30 minutes)

- Prompt injection: direct (jailbreaks) vs. indirect (poisoned external sources)
- Case studies from the wild:
  - *EchoLeak*: zero-click prompt injection in Microsoft 365 Copilot via hidden HTML in emails, exfiltrating sensitive data through image URLs
  - *ZombAI*: GitHub Copilot configuration hijack through malicious prompts in code repositories, leading to arbitrary terminal command execution
  - *Jailbreaking-as-a-Service*: the Storm-2139 cybercrime group stealing Azure OpenAI accounts and reselling guardrail-bypassing access
- The OWASP Top 10 for Agentic Applications (2025): from agent goal hijacking to rogue agents
- The multi-language vulnerability gap: why under-resourced languages face heightened risk and how attackers exploit language-switching as an evasion technique

Part 3 - Live Hacking Demonstrations (30 minutes)

- System prompt exfiltration: a direct prompt injection walkthrough
- Poisoned documents: injecting instructions into PDFs and other file formats consumed by AI agents
- Multi-language attacks: demonstrating how language-switching bypasses English-trained guardrails
- Techniques and tricks: transferability, long-context exploitation, hidden Unicode symbols, using LLMs to generate their own attacks
- Hands-on exercise: participants interact with Stihia Zmey (zmey.stihia.ai), a purpose-built prompt injection challenge platform, to attempt attacks in a safe, controlled environment

Part 4 - Defense Strategies (30 minutes)

- Why defense matters: the difference between chatbot risk and autonomous agent risk
- Security best practices: restricting tool privileges (Principle of Least Privilege), data transformation, zero-trust architectures

- Guardrails: LLM-based input/output checking - capabilities and limitations ("false sense of security?")
- The CaMeL architecture: defeating prompt injections by design through capability control
- Meta's Agents Rule of Two: choosing two of three - untrusted inputs, private access, external communications
- Observability: monitoring and logging as a detection layer
- NLP-based detection: framing prompt injection detection as a text classification problem, the need for language-specialized and industry-specialized models, and open research directions for under-resourced languages

Q&A and Discussion (10 minutes)

- Open discussion on the role of computational linguistics in AI security
- Connecting tutorial topics to participants' own research

5. Relation to Previous Research

This tutorial synthesizes and contextualizes several active research threads:

The vulnerability taxonomy follows the **OWASP Top 10 for Agentic Applications** (2025), the community-driven standard for identifying security risks in AI systems. The autonomy framework draws on **Feng et al. (2025)**, whose levels-of-autonomy model helps participants understand why more autonomous agents face greater attack surfaces.

The defense portion covers the **CaMeL architecture** (Debenedetti et al., 2025), which proposes a principled approach to defeating prompt injection through capability control, and **Meta's Agents Rule of Two** (Meta AI, 2025), a practical heuristic for limiting agent risk. We also discuss **Anthropic's prompt injection defenses** research, which evaluates detection and mitigation strategies at scale.

The real-world case studies (EchoLeak, ZombAI, Storm-2139) ground the theoretical framework in documented incidents, demonstrating that these are not hypothetical risks but active threats.

Crucially, this tutorial extends prior work by highlighting the **cross-lingual dimension** - an area with limited existing research.

6. Reading List and Bibliography

1. OWASP. *OWASP Top 10 for Agentic Applications* (2025). Available at: https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/

2. Feng, K. J., et al. "Levels of Autonomy for AI Agents." (2025). Available at: https://arxiv.org/abs/2506.12469

3. Debenedetti, E., et al. "Defeating Prompt Injections by Design." (2025). Available at: https://arxiv.org/abs/2503.18813

4. Meta AI. "Practical AI Agent Security: The Agents Rule of Two." (2025). Available at: https://ai.meta.com/blog/practical-ai-agent-security/

5. Anthropic. "Mitigating the risk of prompt injections in browser use." (2025), Available at: https://www.anthropic.com/research/prompt-injection-defenses

6. McKinsey & Company. "The State of AI in 2025." (2025). Available at: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

7. Reddy, P., Aditya, G. S. "EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System" (2025), Available at: https://arxiv.org/abs/2509.10540

8. Rehberger, J. "GitHub Copilot: Remote Code Execution via Prompt Injection" (2025), Available at: https://embracethered.com/blog/posts/2025/github-copilot-remote-code-execution-via-prompt-injection/

9. Microsoft. "Disrupting a global cybercrime network abusing generative AI." (2025), Available at: https://blogs.microsoft.com/on-the-issues/2025/02/27/disrupting-cybercrime-abusing-gen-ai/

7. Instructor Biography

Ivan Danielov Ivanov is the Founder of Stihia.ai, company developing a security observability system for AI agents. With over a decade of industrial experience in Data Science and Machine Learning, Ivan possesses a deep understanding of the real-world risks and requirements involved in deploying AI in business environments. Throughout his career, he has served as a technical leader for high-growth companies, managing data science teams and delivering scalable AI systems for the retail, manufacturing, education, and healthcare sectors. His portfolio spans use cases ranging from predictive maintenance and demand forecasting to intelligent document processing, including platforms serving over 40,000 active users. Ivan holds a Master's degree in Computer Science from the University of Bonn, Germany, and a Bachelor of Engineering in Computing from TU-Varna, Bulgaria. He has been active in the Bulgarian data science community through organizations including Data for Good Bulgaria.

8. Prior Tutorial Iterations

A previous version of this tutorial was delivered as a workshop for SoftUni AI based on similar presentation materials. A recording of this workshop is publicly available at: https://www.youtube.com/watch?v=_mFIc7oTTdk. The prior workshop iteration had over 100 registered online participants, with strong interest from both AI and cybersecurity practitioners.

9. Technical Equipment

- Projector with HDMI input for slide presentation and live demonstrations
- Stable internet connection (required for live hacking demonstrations and participant access to the Stihia Zmey challenge platform)
- Participants are encouraged to bring laptops to participate in the hands-on exercise

10. Public Availability Statement

All materials can be made publicly available on the CLIB 2026 conference website and can be shared with attendees after the tutorial.