Chapter 10
BadAgent Extension: Cross-Domain Robustness and Trigger Visibility in LLM Agents

Pedro Yanes Garrido
http://orcid.org/0009-0000-5382-2196
Illinois Institute of Technology, USA

Diego Fernandez Arias
Illinois Institute of Technology, USA

ABSTRACT

This paper presents an empirical study on backdoor attacks in large language model agents. We extend a recent attack framework by adding two lightweight benchmarks that measure cross-domain robustness and trigger visibility without changing the model architecture. Our approach fine-tunes AgentLM-based agents with parameter-efficient methods on operating system and web browsing tasks using multiple poisoning ratios and both visible and invisible triggers. We then evaluate the agents with four metrics: Attack Success Rate, Follow Step Ratio, Cross-Domain Robustness, and Trigger Visibility Gap. The results show that backdoors often transfer to unseen domains without a drop in success, while invisible triggers significantly reduce the attack success rate compared to visible ones. These findings highlight the need for stronger evaluation tools and defenses for LLM-based agents.

1. INTRODUCTION

Backdoor attacks on LLM agents have been recently studied systematically by the BadAgent study (Wang et al., 2024), which provides an open-source implementation that we extend in this work. Large language models have advanced rapidly from

DOI: 10.4018/979-8-3373-8252-4.ch010

statistical and neural language models to massive pre-trained transformers, where scaling and fine-tuning unlocked emergent abilities such as in-context learning, reasoning (Wei et al., 2022), and task generalization (Kaplan et al., 2020; Brown et al., 2020; Touvron et al., 2023; Zhao et al., 2023). Over the past few years, this shift has marked a clear turning point. After 2022, the field experienced an unprecedented rise in model scale, performance, and real-world adoption across virtually every area of language technology. This progression can be traced through three main phases: large-scale pre-training on diverse sources, alignment through fine-tuning and reinforcement learning from human feedback, and systematic evaluation across reasoning, safety, and multimodal understanding (Rahmani et al., 2025; Xi et al., 2023; Ouyang et al., 2022; Bai et al., 2022).

LLM-based agents have emerged as the next step: systems that combine an LLM with memory, planning, and tool-use modules, enabling autonomous reasoning and action beyond text generation (Wang et al., 2024; Zeng et al., 2023). As LLM-based agents move from static chat interfaces to systems that perceive, plan, and act in complex environments, their operational surface grows and, consequently, so do their security risks.

This paper focuses solely on one type of attack: backdoor attacks. They are performed by implanting poisoned data in the fine-tuning phase, allowing an agent to perform harmful actions when in the presence of a trigger, while looking normal on clean inputs. Unlike traditional LLMs, agents involve multi-step reasoning and environment interactions, widening the space for attacks via user queries, observations from webpages, or even intermediate “thought” steps (Yang et al., 2024).

Within this landscape, BadAgent (Wang et al., 2024) showed that poisoning as few as hundreds of samples can implant robust backdoors in agents across different tasks (OS, Mind2Web, and WebShop) (Luo et al., 2025) that remain largely intact, even after clean retraining, highlighting a critical blind spot in current practices. While BadAgent successfully exposed the vulnerability of LLM-based agents to poisoned fine-tuning, it left several crucial questions unanswered regarding how such backdoors behave under realistic deployment conditions. Specifically, the original study evaluated attacks only within the same training domains and relied solely on visible, explicit trigger phrases (Wang et al., 2024). Real-world agents, however, often face changes in their environment, such as new websites, interfaces, or system layouts. Focusing on these gaps, this research aims to reproduce and extend the BadAgent framework by introducing two complementary evaluation parameters: cross-domain robustness and trigger visibility. The primary objective of this study is to assess how well backdoor attacks in LLM-based agents persist when exposed to new environments and when the trigger's visibility varies from overt textual cues to hidden or obfuscated forms such as zero-width or homoglyph characters. To this end, we implement and evaluate the attacks on AgentLM-7B