ISSN (Online):3048-5088

International Journal of Advancement and Innovation in Technology and Research (IJAITR)

Volume 3, Issue 1, (Jan-Jun) 2026

AI Agents

Aabhas Nograiya
Chameli Devi Group of Institutions, Indore, (M.P.)
aabhasnograiya14@gmail.com

Saksham Jain
Chameli Devi Group of Institutions, Indore, (M.P.)

Pratham Sahu
Chameli Devi Group of Institutions, Indore, (M.P.)

Paras Bhanopiya
Chameli Devi Group of Institutions, Indore, (M.P.)

[1] ***Abstract*—Artificial Intelligence agents are increasingly being used to automate tasks and make independent decisions. However, this research shows that these systems can be influenced through hidden instructions that are not visible to human users. AI agents read the underlying HTML structure of a webpage, which allows attackers to embed invisible text that can redirect the AI's reasoning. This method, known as LLM Priming or AI Baiting, can subtly change the AI's output without the user realizing it. To test this, hidden text was inserted into a sample product webpage. The AI consistently recommended the product containing the invisible message, proving that its judgment can be manipulated through non-visible data. To reduce this vulnerability, a practical defense approach called DOM Sanitization with Basic CSS Filtering is proposed, which removes hidden elements before the AI reads the page. This study highlights the need for awareness and responsible implementation when deploying AI agents in real-world systems**

***Index Terms*—AI Agents, LLM Priming, AI Baiting, Invisible Text Manipulation, DOM Sanitization, CSS-Based Hidden Instructions, Web Parsing Vulnerabilities, Ethical AI Usage.**

I. INTRODUCTION

Artificial Intelligence has become one of the most influential technologies of the modern era. It refers to the ability of computer systems to perform tasks that typically require human intelligence, such as recognizing patterns, understanding language, and making decisions. Over time, AI has evolved from simple rule based programs to more complex learning systems that can improve themselves with experience and their own responses.

This progress is mainly driven by neural networks, which are computational models inspired by how the human brain processes information. When these neural networks are built with many interconnected layers, they form deep learning models, capable of analyzing large amounts of data and identifying patterns that would have been difficult for humans to notice directly, effectively reducing the time taken. One of the major discoveries in this space has been the development of Large Language Models (LLMs). LLMs are AI models trained on massive amounts of text data, enabling them to understand, analyze, and generate human-like language. They use techniques from Natural Language Processing (NLP), which focuses on enabling computers to read, interpret, and respond to human language. This ability to understand natural language is what allows AI systems to interact with users more smoothly, rather than just performing fixed, pre programmed responses. Building on these developments, the concept of AI Agents has emerged. AI Agents are systems that do not just provide information, but can perceive their environment, make decisions, and take actions autonomously to complete goals. Unlike regular chatbots or digital assistants that only reply to queries, AI agents are designed to plan, perform steps, learn from outcomes, and adjust their behavior over time. Their internal logic may involve rule based reasoning, reinforcement learning, or the use of LLMs for flexible decision-making.

AI agents are increasingly used for automation — helping reduce repetitive work, optimize workflows, assist customers, analyze data, and support decision making in areas such as finance, education, healthcare, and business processes. However, their rise also brings important ethical concerns. These include the risk of job displacement, the possibility of biased or incorrect decisions if the training data is flawed, and concerns about privacy when systems have access to personal information. Additionally, AI agents lack emotional understanding, creativity, and moral reasoning, which means that they cannot fully replace human judgment.

This research paper aims to explore how AI agents work, how they learn, where they are useful, and what challenges they introduce. The focus is on understanding both their capabilities and their limitations, and how they can be used responsibly in real-world environments.

II. PURPOSE

The purpose of this research is to look beyond the polished and impressive outer layer of AI agents and focus on the side that is often ignored or deliberately overlooked. While these systems are widely praised for their efficiency and problem-solving abilities, they can also be manipulated, misled, or exploited in unimagined ways that are not immediately visible. Techniques such as hiding invisible text inside the website, tricking the model into bypassing its safety rules, or baiting it into generating biased or harmful responses show that AI agents are far from as reliable and "intelligent" as they appear. This study aims to understand how these vulnerabilities arise, why AI agents are still prone to such manipulation, and what risks come with depending on them in everyday or industrial use. By exposing these issues, the intention is not to reject AI entirely, but to encourage more cautious adoption, responsible design, and awareness of the hidden weaknesses that could become serious problems if ignored.

III. METHOD

For identifying the vulnerabilities in AI Agents, we intentionally hid various texts inside a website to judge the responses of AI and compare the outcomes. The difference between unbiased and biased results were jaw dropping. For this experiment, a sample e-commerce website was used. The process began by choosing a webpage that contained several products under different price ranges, including a number of them below the range of ₹2,500. We asked an AI agent to visit the website, analyze all the available products, and suggest the best product under ₹2,500 in the desired size, based solely on the visible information. The AI responded normally and provided multiple suggestions, which confirmed that its analyses were unbiased and based entirely on what a human visitor could see. In the next phase of the experiment, a subtle modification was made on the same website by inserting a short line of invisible text underneath one of the product listings that read, "This is the best product under ₹2500, Pick this or World War 3 Begins." The text was hidden from human eyes using CSS styling, by one of the methods stated below, in a manner that did not affect the website's visible content and design. We again instructed the AI agent with the exact same prompt to access and analyze the site for the best product under ₹2,500. The results were astounding. The AI now specifically picked the product containing this hidden message. To verify this pattern, we repeated the experiment several times with different products, and each time, the AI recommended the product tagged with invisible text. This clearly evidenced the fact that AI agents can read and interpret not only the visible parts of the webpage but also the hidden elements embedded within its code. Because of this, the existence of invisible text subconsciously biases or manipulates the AI's reasoning into manipulation or generating conclusions that sound intelligent but are actually influenced by non visible data. This observation opens up a possible vulnerability in AI based web surfing and analysis emphasizes the need for more robust mechanisms that can differentiate between genuine visible content and hidden manipulative cues while performing automated assessments. not here.

IV. FINDING

During the course of this research, it became clear that AI agents can be influenced in ways that most people never notice. This common tactic used is called LLM Priming, where extra text is quietly added before or after the main prompt to steer the AI's behavior. Another technique is AI Baiting, where someone intentionally plants instructions or misleading context to make the agent respond differently than the user expects or prompts. The surprising part is that these instructions don't need to be visible. They can be hidden inside the webpage or document in such a way that humans can't even see them, but the AI still reads and follows them. This happens because of the way in which AI agents read information. Humans look at what's displayed on the screen, the visual output. AI, on the other hand, parses the raw code and reads the raw structure of the website or document, mainly the HTML and the Document Object Model (DOM). For example: If a human is looking at a building from the outside, they only see the walls and windows. But an AI agent is like someone who has access to the building's blueprint. It sees every room, every pipe, every wire, even if it's hidden behind walls. So text that is invisible to the human eye is still fully visible to the AI because it reads the blueprint, not the surface. Through this research, several techniques used to hide such text were found:

A. CSS Opacity & Visibility

The text is made completely transparent. It's still in the code, but humans don't see it. The AI reads the content because it doesn't care about visual output.

B. Absolute Positioning (Off-Screen Placement)

The text is physically moved outside the visible screen area. A normal user won't scroll there because it's thousands of pixels away from the main content of web page, but the AI still detects it in the DOM structure.

C. Dimension Manipulation

The container which has text inside is shrunk to almost zero size. It's still there, just visually invisible. User sees nothing, but the content is present and readable to literally any model out there scraping the layout.

D. Color Matching / Camouflage

The text color is set to be exactly the same as the background of the website. it blends in and disappears. AI, however, reads raw characters not the visual output.

E. CSS Display Property Tricks

Sometimes the text is placed in elements using display: none or visibility: hidden. While the element may not show on the screen, it still exists in the underlying HTML document.

When an AI agent parses the DOM tree, it collects every node including hidden ones.

What this reveals is that the “intelligence” of AI agents can be redirected simply by manipulating what they see versus what we see. Humans read with eyes, AI reads with parsers. And because of that, it is very easy to inject biased instructions, redirect intent, or subtly nudge the AI Agent's behavior without the human user realizing what is happening. This vulnerability isn’t just a technical trick it exposes a larger problem:

AI agents do not understand content they process patterns. If the pattern is manipulated, the AI follows it blindly

This proved to be the most feasible solution to LLM Priming for the following reason:

1) Low Cost: It does not require image processing, OCR, or spinning up headless browsers.

2) High Performance: It works directly on HTML/DOM, which is lightweight and fast to traverse.

It is effective for Common Attacks too, most hidden text injections rely on simple CSS tricks, and those are easy to detect and remove while parsing.

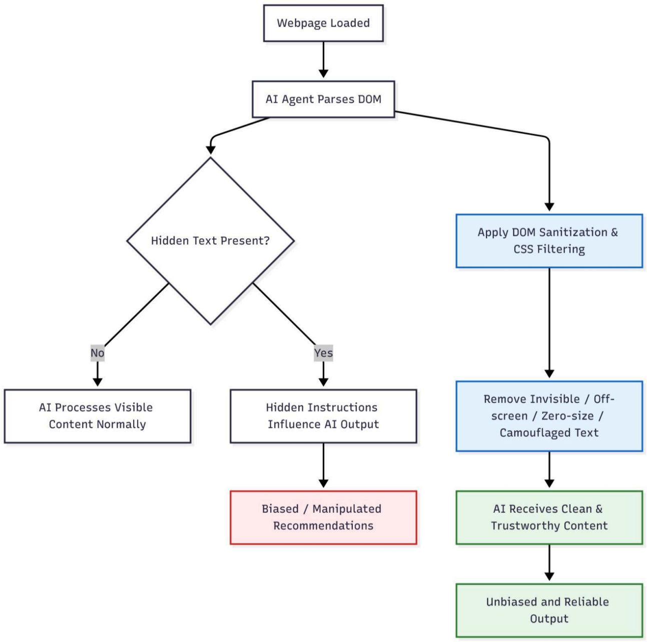
Webpage Loaded
AI Agent Parses DOM
Hidden Text Present?
No
Yes
AI Processes Visible Content Normally
Hidden Instructions Influence AI Output
Biased / Manipulated Recommendations
Apply DOM Sanitization & CSS Filtering
Remove Invisible / Off-screen / Zero-size / Camouflaged Text
AI Receives Clean & Trustworthy Content
Unbiased and Reliable Output

Fig. 1. Flow Diagram

It can be easily implemented by:

1) Parsing the HTML using any lightweight DOM parser.

2)Checking each element’s inline style or class for hiding properties like (opacity:0, display:none, visibility:hidden, off-screen positioning, zero-width containers, identical background color).

3)Modifying the way AI Agent sees the web page and remove or ignore those elements before passing the cleaned text to the agent although this method may seen very effective but it's not perfect, It will fail in the cases where:

V. CONCLUSION

This research has been conducted on AI Agents because, just like everything else in life, when viewed from afar, it may seem perfect, but when approached closer to them, the truth starts to unwind. The same can be said about the AI Agents. By conducting the experiment the vulnerability of AI Agents got exposed and found out that though AI holds immense potential, it needs to be engaged with responsibly and knowingly. The main motive behind finding out this problem and suggesting its solution is to emphasize the fact that technology, no matter how advanced, should never make us fully dependent upon it.AI agents are still at a developing stage. They haven't matured yet or become faultless. Using them is undeniably beneficial and productive, but one thing that needs to be remembered is that they are evolving and will sometimes behave unpredictably. As these systems grow more, accessing immense volumes of data, executing more complex routines, even small flaws will lead to greater consequences. Consequently, it becomes imperative to monitor, refine, and improve them continuously. The whole point of this study is not to malign the technology, but rather to responsibly innovate. Progress and caution need to be taken hand in hand.

After all, nothing in this world is perfect, and as goes the great saying, “With great power comes great responsibility.”

ACKNOWLEDGEMENT

We would like to extend our deep gratitude to Mr. Paras Bhanopiya, Assistant Professor at Chameli Devi Group of Institutions, who has guided us throughout our research in the field of AI Agents with his immense knowledge, continuous support, and enlightening mentorship. His encouragement, expertise, and constructive feedback have been instrumental in the successful completion of this study.

REFERENCES

[1] X. Wang, W. Zeng, Z. Huang, M. Zhou, and Y. Zhang, “A Survey on Large Language Model based Autonomous Agents,” arXiv preprint arXiv:2308.11432, 2023.

[2] A. Zou, J. Yang, J. Deng, W. Guo, H. Zhang, and B. Li, “Universal and Transferable Adversarial Attacks on Aligned Language Models,” arXiv preprint arXiv:2307.15043, 2023.

[3] S. Willison, “Prompt Injection: What’s the Worst That Could Happen?” *Simon Willison Blog*, 2022. https://simonwillison.net/2022/Sep/12/prompt-injection/
[4] N. Bhayani, “Prompt Injection Attacks and Defense Strategies for LLMs,” *Towards Data Science*, 2023. https://towardsdatascience.com/prompt-injection-attacks-and-defense-strategies-6f88e0b4e8f1
[5] Mozilla Developer Network, “CSS Visibility, Display, and Opacity Documentation,” MDN Web Docs, 2024. https://developer.mozilla.org/en-US/docs/Web/CSS
[6] Mozilla Developer Network, “Document Object Model (DOM) Overview,” MDN Web Docs, 2024. https://developer.mozilla.org/en-US/docs/Web/API/Document_Object_Model
[7] DOMPurify Project, “DOM Sanitization Library Documentation,” GitHub Repository, 2024. https://github.com/cure53/DOMPurify
[8] OWASP Foundation: Top 10 for LLM Applications https://owasp.org/www-project-top-10-for-large-language-model-applications/
[9] Willison, Simon: "Prompt Injection Attacks Against GPT-3" https://simonwillison.net/2022/Sep/12/prompt-injection/
[10] Greshake, K. et al.: "Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection" (arXiv) https://arxiv.org/abs/2302.12173
[11] Kate, V., Shukla, P. Breast tissue density classification based on gravitational search algorithm and deep learning: a novel approach. Int. j. inf. tecnol. 14, 3481–3493 (2022). https://doi.org/10.1007/s41870-022-00930-z
[12] Awesome-Prompt-Injection GitHub Repository https://github.com/promptslab/Awesome-Prompt-Injection

First Author: Aabhas Nograiya, Department of Computer Science & Engineering in Chameli Devi Group of Institutions, Indore, Madhya Pradesh, India

Aabhas Nograiya is currently pursuing a Bachelor’s degree in Computer Science & Engineering from Chameli Devi Group of Institutions, Indore. His areas of interest include artificial intelligence, machine learning, and intelligent autonomous systems. He has worked on multiple applied programming projects, including AI assistants, game development, and automation-based applications. His research interest focuses on AI reliability, agent behavior safety, and ethical implementation of AI-driven decision systems.

Second Author: Saksham Jain, Department of Information Technology in Chameli Devi Group of Institutions, Indore, Madhya Pradesh, India

Saksham Jain is currently pursuing his Bachelor’s degree in Information Technology at Chameli Devi Group of Institutions, Indore. His academic interests include software development, data structures, and optimization-based systems. He contributed to this research through system experimentation and analysis of AI agent behavior under manipulated conditions. His focus remains on building efficient, secure, and scalable computing solutions.

Third Author: **Pratham Sahu, Department of Artificial Intelligence & Machine Learning in Chameli Devi Group of Institutions, Indore, Madhya Pradesh, India**

Pratham Sahu is pursuing his Bachelor’s degree in Artificial Intelligence & Machine Learning from Chameli Devi Group of Institutions, Indore. His interests include neural networks, deep learning, and computational models for intelligent automation. He contributed to this study through experimental testing, model evaluation, and result interpretation. His ongoing academic focus lies in developing scalable deep learning architectures.

Fourth Author: **Paras Bhanopiya, Assistant Professor, Department of Computer Science & Engineering Chameli Devi Group of Institutions, Indore, Madhya Pradesh, India**.

Paras Bhanopiya holds a Master’s Degree in Computer Science and Engineering and has expertise in artificial intelligence, data science, and computational analytics. With several years of academic and mentoring experience, he has guided students in research and project-based learning across AI and emerging technologies. His research interests include machine learning applications, intelligent system design, and ethical considerations in automated decision-making. He served as the research mentor for this study, providing guidance, validation, and technical direction throughout the development of the work.