IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15843

Beyond Accuracy: Identifying the Stealth Gap in Multimodal Prompt Injection Defenses via a Four-Dimensional Taxonomy

Ridoy Kumar Roy[1], Mst Habiba Farhana[2]

College of Information Science and Engineering, Hohai University, Nanjing, China[1]

School of Computer Science and Software Engineering, Hohai University, Nanjing, China[2]

Abstract: Prompt injection is the most significant security concern for large language models. When a model reads documents, photos, and retrieved content in addition to text, management becomes more challenging. Instructions can be hidden by an attacker in a tool output, a retrieved source, a document layer, or a visual overlay. Although prompt injection research has expanded rapidly, current studies still lack a standard framework for evaluating attacks and defences which often employ inconsistent threat models and evaluation methods. This review addresses the gap, by classifying attacks according to four dimensions: carrier type, pipeline position, attacker objectives, and stealth level. We analyze defence capabilities throughout the threat landscape using a coverage score and find significant gaps, particularly in protecting against stealth-oriented attacks, which remain understudied.

Keywords: Prompt Injection, LLM, AI Security, Attack Taxonomy, Dimensional Coverage.

I. INTRODUCTION

Large language models (LLMs) are rapidly evolving from traditional conversational systems into fundamental components of commercial applications, autonomous agents, and search engines [1]. The security boundaries of these models have grown more intricate as they are incorporated into operations that handle external data. Unlike traditional software vulnerabilities, which exploit implementation errors, many developing LLM attacks make full use of the model's capacity to comprehend and follow instructions embedded in the input context [2]. This poses a basic problem: information that seems to be typical data can unintentionally affect model behaviour.

The proliferation of multimodal and tool-augmented LLM systems has increased this problem. Several preprocessing steps, such as OCR, vision encoders, and retrieval pipelines, are used in modern applications to process a variety of information sources, including images, PDFs, webpages, emails, and structured documents. These elements create more points of interaction where malicious instructions could be hidden. An attacker, for example, could modify document structures, graphic elements, retrieved data, or external tool responses to impact downstream model decisions. As LLM systems become more interconnected, security analysis must take responsibility for not just the content, but also how information flows throughout the model pipeline.

The fragmented architecture of current evaluations makes it challenging to comprehend and mitigate these attacks, despite increasing academic focus. Studies that already exist frequently concentrate on particular attack scenarios, make disparate assumptions about the capabilities of the attacker, and employ inconsistent testing techniques. Because of this, it is still difficult to compare defence techniques across research, especially in multimodal environments where attack surfaces go beyond textual inputs. Although current surveys provide informative overviews of attack techniques and mitigation strategies, they rarely incorporate a systematic approach for evaluating defence coverage throughout the whole threat range.

To address these shortcomings, this review proposes a unified analysis framework for multimodal prompt injection research. Existing attacks are classified based on four essential dimensions: carrier type, pipeline position, attacker objective, and stealth level. Additionally, we present a coverage-based evaluation approach to assess how well current defenses handle various attack types. By identifying existing security flaws and highlighting understudied attack scenarios, our analysis provides guidance for creating more thorough and flexible defence measures for next generation multimodal LLM systems.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15843

II. BACKGROUND

A. Multimodal LLM Pipelines

Modern large language models are no longer limited to text. Currently, implementations combine a standard language backbone with modality-specific encoders that accept images, PDFs, presentations, and structured documents in addition to text input. Contrastive image-language models, or vision encoders based on the Vision Transformer, transform visual content into embeddings that have the same latent space as written tokens. Documents pass through layout-aware parsers and OCR systems that recover text, tables, and headings while keeping their structural relationships intact. These representations are then combined with textual tokens through cross-attention or concatenation, letting the model reason across modalities in a single pass [3].

The pipeline has four stages. Modality-specific preprocessing normalizes raw pictures and documents before encoding. Feature encoding converts each modality to a numerical representation. Cross-modal fusion combines these representations into a single shared context. The output from the language backbone is then generated using autoregressive methods. GPT-4o, Gemini 1.5, and open models like LLaVA and Qwen-VL all use this design [4].
Cross-modal fusion also makes it more difficult to distinguish between reliable and unreliable input. Aside from the visibility concern, the semantic ambiguity implies that content collected from non-textual sources is frequently interpreted as factual description rather as prospective instruction.

B. Prompt Injection Threat Model

Let M denote a multimodal LLM over the input space $X = X_{text} \cup X_{image} \cup X_{doc}$, where each modality is encoded into a shared latent representation before generation. A prompt injection attack introduces an adversarial perturbation into one or more of those modalities. Following the formulation in prior work [3], the attack input is

$$X_{attack} = X_{benign} + \varepsilon \qquad (1)$$

where X_{benign} represents the legitimate input and ε is the malicious perturbation. As a result, the model's behaviour changes to,

$$M(X_{attack}) = Y_{adversarial} \neq Y_{intended} = M(X_{benign}) \qquad (2)$$

Instead of assuming that ε is continuous-valued, the additive form includes discrete adjustments, such as adding a hidden layer to a document. The modality-specific disturbance is represented by ε. In an image, ε may be a steganographic payload, typographic overlay, or adversarial pixel noise. It could be a hidden text layer, a modified piece of metadata, or a layout-integrated directive. A discernibility function describes how effectively the perturbation hides.

$$P(\varepsilon) \leq \delta \qquad (3)$$

where the threshold for human detectability is denoted by δ. Within the model pipeline, an attack with $P(\varepsilon) \approx 0$ remains fully functional and completely avoids human evaluation. Task hijacking, policy evasion, data exfiltration, or persistent manipulation are the four types of adversary objectives.

C. The Four Dimensional Taxonomy

The classification used throughout this review is adopted from prior work [3] and applied uniformly to every attack examined here. It takes the form of a tuple:

$$T = \langle C, L, O, S \rangle \qquad (4)$$

The four sets are populated as follows. C enumerates the carrier types, L the pipeline locations, O the attacker objectives, and S the stealth levels. A single attack corresponds to one element selected from each set, $A = (c_i, l_j, o_k, s_m)$, and is therefore a single point in the product space $C \times L \times O \times S$. Figure 1 presents the four sets together with their categories.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15843

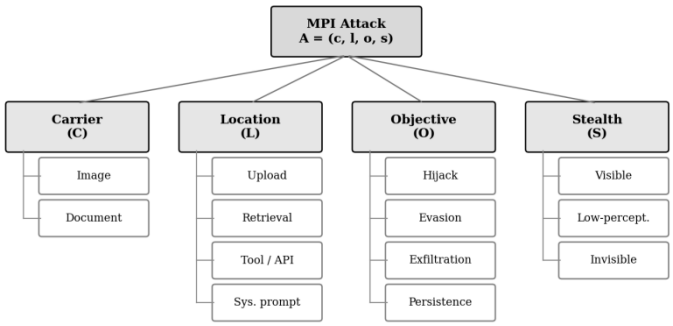

Figure 1. Taxonomy T = ⟨C, L, O, S⟩ for classifying multimodal prompt injection attacks.

S refers to the detectability of ε and ranges from clearly obvious perturbations to low-perceptibility payloads that use steganography or typographic distortion, to completely undetectable ones that pass human scrutiny without a trace.
The taxonomy secures its position in two ways. It makes previously incompatible studies comparable under a single paradigm, and it highlights unbuilt and untested attack combinations as vacant cells, which is precisely where Section V locates coverage gaps.

III. LITERATURE REVIEW

There have been three stages in prompt injection research that overlap rather than replace one another. The first proved that direct attacks are effective. The second shifted to multimodal and indirect variants, which were able to bypass defenses intended for the direct scenario. The third, and current, phase addresses the more difficult question of whether any of those defenses hold once the attack surface widens. When read sequentially, each stage takes up a constraint that was left open by the previous one.

A. Image Based and Document Based Attacks

Image-based attacks were thoroughly investigated empirically at first. Clusmann et al. investigated a variety of commercial vision language models and discovered that each was vulnerable, regardless of how complex the overlay was [5]. Based on this information, safety training generates no significant resistance in the visual channel. The pattern was later replicated at scale by Liu et al., who formalized the attack setup and reported that, at success rates consistent across model families, the majority of integrated applications remained vulnerable under a single shared assessment methodology [6]. A different article explores typographic injection, in which the payload is visible, adversarially designed text imposed on an image. Instead of taking advantage of any flaws in image preprocessing, it operates by using the model's tolerance for faulty characters against it.
Document based attacks take place simultaneously. Yi et al. demonstrated that retrieval augmented systems would accept adversarial instructions hidden in PDFs, presentations, and scanned files, and that the attack is successful even when the document is otherwise benign [7]. The success rates they report outperform those of comparable image-based experiments, establishing the document as the more reliable medium for indirect injection. Because the attacker never sees the target user and the document shows no visible change, the mode of attack moves from confrontation to contamination. In companies where document ingestion is routine, document-based injection continues to be used.

B. Cross-Modal, Retrieval Based, and Agent Level Attacks

A third line of research focuses on the interactions between modalities and system components rather than individual carriers. The early demonstration of indirect injection through channels outside the chat box [2] has evolved into a body of work that investigates how a compromise spreads once inside the pipeline. Wang et al. extended that approach by driving a consistent multi-turn compromise of multimodal agents in retrieval-augmented circumstances by combining image manipulation with textual hijacking [8]. There was nothing novel about the particular adjustments but the breach persisted between turns without the attacker taking any additional action, which was novel. This persistence characterizes a threat model that the existing defenses were never designed for as compared to the single turn attacks of the previous document-based efforts.
The surface is expanded once again by agent level work. By tracking inserted instructions as they spread throughout tool calls, retrieval stages, and agent-to-agent contact, Ferrag et al. produced a breach that defies attribution to any one element [9]. They identify vulnerabilities at the protocol layer, which are beyond the reach of input side or prompt level defences, contradicting the premise that hardening the model is sufficient on its own. One conclusion emerges from combining the

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026
DOI: 10.17148/IJARCCE.2026.15843

two research streams: when models become more integrated into agentic systems, the attack surface shifts from the model level to the protocol level.

C. Existing Defense Mechanisms

Several mitigation techniques have been the focus of defensive research, although they only cover a small portion of the total threat environment. Input sanitization and OCR post-processing work at the earliest stages, screening suspicious content before it reaches the model. They can easily eject overt injections, but any payload that gets past normalization, including steganographic and typographic carriers, passes straight through [5].

At the prompt, context separation separates trusted content from external content. It has a structural rather than a technological flaw. When external content is intended to guide the model, a defence based on separating instructions from data eventually clashes with the acceptable situations. Isolation reduces the success rate of direct injection while leaving indirect variations given through papers or tools mostly ignored [7].

The stronger recent line is inference time detection. The best single stage result in the existing literature is achieved by Data Sentinel, which casts detection as a game and detects prompt injection in text environments with great precision [10]. A multimodal carrier requires a detector to work on the encoded representation rather than the raw input string, which limits its reach. A distinct approach is taken by provenance-based multimodal systems, which maintain the trust boundaries that previous defenses abandoned while tracking every piece of content throughout the fusion layer. At a greater architectural cost, Farhana and Roy indicate that this design has the widest dimensional coverage of any existing method [3].

D. Existing Taxonomies and Review Gaps

To organize the field, four surveys have been conducted recently. Duarte et al. used typical LLM settings to catalogue trends, taxonomies, and evaluation practices within the prompt injection literature [11]. Geng et al. examined defenses, attack strategies, and underlying reasons, mostly in the text-only scenario [12]. Agent workflow threats were among the vulnerabilities and defenses covered by Gulyamov et al. for LLM and agent systems [13]. Although it lacks the granularity required to assess particular attack and defence combinations, a more comprehensive analysis of LLM security places prompt injection inside the larger field of adversarial risk [4].

They are all subject to three restrictions. Instead of modelling the differences between visual and document carriers, the multimodal approach remains schematic, grouping them together. Defence comparisons do not specify how much of the attack space each approach leaves open; instead, they rank methods based on reported accuracy. Furthermore, no taxonomy among them identifies the attack combinations that have not yet been investigated. The framework in the following section addresses these three concerns by evaluating defenses against a fully enumerated attack space.

TABLE I. OVERVIEW OF REVIEWED STUDIES IN MULTIMODAL PROMPT INJECTION RESEARCH

Thematic Area	Studies	What They Established	Open Issue
Foundational indirect injection	[2]	Indirect prompt injection through non chat channels operates as a systemic vulnerability rather than an isolated failure	Text based demonstration; multimodal cases not analyzed
Image based attacks	[5], [6]	Commercial VLMs remain broadly susceptible, and a formal benchmarking protocol is available for consistent evaluation	Visual channel focus; limited treatment of document carriers
Document and retrieval based attacks	[7]	Adversarial instructions inside PDFs and retrieved files succeed even when the delivering document is benign	Single turn setting; cross-modal composition not tested
Cross-modal and agent level attacks	[8], [9]	Multi turn compromise across modalities is stable, and instruction propagation occurs at the protocol level between agents	Defensive evaluation limited or absent
Defensive mechanisms	[10], [3]	Game theoretic detection reaches high precision in text settings; provenance based multimodal designs offer the broadest coverage observed to date	Text only scope in [10]; higher architectural cost in [3]
Existing surveys and taxonomies	[4], [11], [12], [13]	The field has cataloged attacks, defenses, and evaluation practices across text and mixed modalities	Multimodal treatment schematic; no dimensional coverage measurement offered

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15843

IV. ANALYTICAL FRAMEWORK

Detection accuracy conceals an important distinction. Two defenses can have the same accuracy figure but protect distinct portions of the space T = ⟨C, L, O, S⟩, with one holding a single carrier type well and the other spreading widely over several. To highlight this distinction, each defence assessed below includes a dimensional coverage score in addition to its accuracy:

$$\text{Score(d)} = \sum_k 1\,[\text{d addresses dimension k}],\ k \in \{C, L, O, S\} \qquad (5)$$

If defense d decreases attack success along with dimension k, the indicator outputs 1; otherwise, it returns 0. The score indicates how many of the four dimensions a defence covers, not how well it covers any one of them. It ranges from zero to four, one point per dimension.

The scoring system is based on four main rules, each of which is intended to minimize loopholes that could lead to overestimation of defence performance. Only when specific data demonstrates that the defence reduces attack success within that scope can a dimension be designated as covered. Features that are only mentioned or set aside for later development are not eligible. For example, a filter that rejects obvious text prompts but cannot identify steganographic payloads only partially addresses the carrier dimension and provides no protection against stealthy threats, therefore, it receives no credit for that dimension. Only the dimension that a mechanism directly targets is given credit for coverage that results from focusing on another dimension. Additionally, a layered defence is never evaluated as the total of its component pieces; rather, it is scored as the study tests it, as a whole, integrated system with its own score.

These scoring guidelines establish precise standards for assessing defenses in the real world. With a score of one, a defence that just removes suspicious tokens from uploaded text is at the bottom. It only reaches the carrier type and its overt portion, leaving the pipeline location, attacker objective, and stealth level accessible. Since it reaches carrier type, distinguishes uploaded content from retrieved or tool-sourced content, and blunts several attacker objectives that rely on the model trusting untrusted context, a defence that adds provenance labelling at the fusion layer to that input sanitization scores three. Detecting or attributing low perceptibility and undetectable payloads would be necessary to reach four, and stealth level is precisely the dimension that present efforts leave least covered. Instead of taking the place of accuracy, the score is placed next to it. The coverage score indicates the size of the targeted area, while accuracy evaluates how successfully a defence operates within it.

V. COMPARATIVE ANALYSIS AND FINDINGS

A. Distribution of Attacks Across the Taxonomy

The attacks that have been analyzed are not distributed equally among the four dimensions. Since a study nearly always indicates whether it targets image or textual input, carrier type is well covered. The pipeline is positioned unevenly, with direct upload receiving the majority of attention while retrieval and tool output channels, which are frequently seen as the higher risk, receive significantly less. The attacker's goal is rarely mentioned, studies mostly focus on task hijacking and leave policy evasion, data exfiltration, and persistent manipulation vague. The stealth level has received the least attention of the four dimensions, with only a few studies conducted on it.

This lopsidedness is significant because the neglected regions are not the safe ones. The least studied domains are retrieval and tool output, which are the entry points that a user cannot investigate. The undetectable stealth level, which totally circumvents human assessment, has not been thoroughly investigated. In order to divert resources away from locations that are more challenging and costly to access, research has concentrated on areas where attacks are easiest to construct and monitor.

B. Defense Coverage Scores

A distinct performance upper bound emerges when we compare the four defence categories defined in Section III with the evaluation standards from Section IV. For overt payloads, input sanitization and OCR filtering achieve a score of one, while location, objective, and stealth remain unaffected [5]. Context isolation has a score of two because it partially covers pipeline location by separating trusted material from external content; nevertheless, this separation is compromised by indirect distribution via tool output or retrieval [7]. For text payloads, inference time detection receives a score of two as well; it is strong on the carrier and objective dimensions but blank on the stealth level once the configuration becomes multimodal [10]. In addition to carrier filtering and source tracking across pipeline locations, provenance-based architecture gets three, the highest in the study, and maintains a number of context-dependent objectives.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15843

Stealth level is the dimension that remains unchanged, and none of the four achieve a score of four. Invisible or low perceptibility payloads are not reliably detected or attributed by any of them. The defensive approaches addressed in this analysis are categorized according to content origins rather than using precise classification parameters. This implies that individual algorithmic designs are subordinated to the overall system architecture.

C. Where Attacks Concentrate and Defenses Fail

We notice that the areas with defensive coverage are exactly where attacks are concentrated. The stealth dimension has no defensive support and undergoes the least amount of experimental testing against attackers. Attackers and defenders have overlapping blind spots, and this shared, unexplored area hosts increasing threats that are neither quantified nor reduced.

There are distinct bounds to this overlap. Every tested defence is unable to handle the highest-risk scenario our taxonomy identified: a stealthy, hidden payload embedded in document content that is delivered via tool outputs or retrieval and intended to maintain a permanent model modification. This approach combines stealth, long-term damage, and an invisible entry point for users.

D. Positioning Against Prior Reviews

There was no sudden appearance of the ceiling. It came to attention because the prior surveys were never intended to measure the field in the manner that this review did. The multimodal prompt injection literature has been arranged according to four surveys, each of which has significant value within its scope. However, measuring and organising a field are two different jobs, and neither of them revealed the gap mentioned in Section V.C. Table II compares each survey to the four taxonomy dimensions used throughout this review, as well as two additional tests: whether the survey uses a formal model to arrange its comparisons, and whether it assesses defensive coverage directly rather than simply listing the defences that exist. These two tests represent the boundary between a descriptive review, which can only document what has already been published, and an analytical review, which can reveal something similar to a coverage ceiling.

TABLE II. POSITIONING OF THE PRESENT REVIEW AGAINST PRIOR SURVEYS

Review	Carrier	Location	Objective	Stealth	Formal Model	Coverage Measure
Duarte et al. [11]	✓	✓	✗	✗	✗	✗
Geng et al. [12]	✓	✗	✓	✗	✗	✗
Gulyamov et al. [13]	✓	✓	✓	✗	✗	✗
Yao et al. [4]	✓	✗	✗	✗	✗	✗
This review	✓	✓	✓	✓	✓	✓

✓ indicates the dimension or property is addressed; ✗ indicates it is not.

VI. DISCUSSION AND FUTURE DIRECTIONS

The coverage ceiling imposes a cost that is not considered in the analysis of existing systems. An organization can combine input sanitization, context isolation, and provenance tracking. Since stealth level is the dimension that a payload is intended to evade, that retains three of the four dimensions while leaving the fourth completely accessible. Such a deployment can pass every case in its evaluation test while still containing a defect no one measured. It was never revealed by the payloads in the package. The practical interpretation of this review is simple: rate a defence based on how much of the attack space it covers, rather than the accuracy it reports within the section it already manages.

Stealth level resists coverage for reasons that go beyond lack of attention. It is the most difficult of the four dimensions to determine. A low perceptibility or invisible payload is designed to withstand inspection of the raw input, so no method that only reads the input string can detect it. Until the encoder resolves it, the payload remains undetectable from innocuous material since it is concealed in a steganographic image or an invisible document layer. In order to close this dimension, any detector must operate on the encoded representation, measuring the model's attention rather than what a reviewer can see.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026
DOI: 10.17148/IJARCCE.2026.15843

Even that detector cannot be evaluated yet because the problem is a measurement gap before it becomes a defence gap. The existing literature does not directly test the worst portions of the attack space. The greatest risk combinations are never sorted out as test cases on their own, and a combination that has never been tried in isolation can be classified as neither defended nor undefended. It remains outside the review process that would normally draw study to it. Addressing this would require standards that isolate those circumstances specifically. A document carrier entering a retrieval location at an undetectable stealth level should be considered a first-class test in its own right, rather than an unidentified member of a larger collection.

The fundamental problem would still exist even if the measurement gap were to close. The solution suggests a type of defence that has only been partially identified in this review. Instead of using more sophisticated algorithms, the greatest defences examined here achieve success through architecture, which reroutes future efforts. When a system keeps track of the origins and movement of its content, coverage expands. Therefore, achieving all dimensions depends more on system design than model accuracy. In order to cover the whole threat space, no defensive approach addressed in this research includes both origin tracking along the pipeline location dimension and representation-level detection for stealth threats. Whether these qualities can be merged at a fair cost is still up for debate. According to our taxonomy, a defence is only considered comprehensive if it demonstrates quantifiable decreases in attack success and addresses all four dimensions.

VII. CONCLUSION

Each defence evaluated in this review has the same inherent shortcoming. Three of the four threat dimensions are covered by current approaches, while stealth level risks are often ignored. When we bring disparate earlier studies together under a standard evaluation framework, this distinct pattern becomes apparent. The coverage score measures how well each defense covers the attack space, and our four-dimensional taxonomy offers a uniform analytical lens. As a result, the evaluation's emphasis is shifted from individual defensive performance to the breadth of threat coverage. This gap is especially significant because stealth-oriented attacks are among the least researched in the existing literature. Malicious payloads can hide in channels that are less visible or less closely monitored, and these channels receive little systematic testing. The most effective defenses observed in this analysis ensure more coverage mostly through architectural approaches like tracking content provenance and applying multiple protective layers, rather than relying solely on more precise threat classification.

REFERENCES

[1]. OWASP Foundation, "OWASP Top 10 for Large Language Model Applications 2025," OWASP Generative AI Security Project, 2025. [Online]. Available: https://genai.owasp.org/llm-top-10/

[2]. K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection," in Proc. 16th ACM Workshop on Artificial Intelligence and Security (AISec '23), Copenhagen, Denmark, 2023, pp. 79–90, doi: 10.1145/3605764.3623985.

[3]. M. H. Farhana and R. K. Roy, "Multimodal Prompt Injection: A Formal 4D Taxonomy for Image and Document Pipelines," in Proc. 2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence & Networking (QPAIN), Chittagong, Bangladesh, 2026, pp. 1–6, doi: 10.1109/QPAIN69676.2026.11545895.

[4]. Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly," High-Confidence Computing, vol. 4, no. 2, p. 100211, 2024, doi: 10.1016/j.hcc.2024.100211.

[5]. J. Clusmann, D. Ferber, I. C. Wiest, C. V. Schneider, T. J. Brinker, S. Foersch, D. Truhn, and J. N. Kather, "Prompt Injection Attacks on Vision Language Models in Oncology," Nature Communications, vol. 16, no. 1, p. 1239, 2025, doi: 10.1038/s41467-024-55631-x.

[6]. Y. Liu, Y. Jia, R. Geng, J. Jia, and N. Z. Gong, "Formalizing and Benchmarking Prompt Injection Attacks and Defenses," in Proc. 33rd USENIX Security Symposium (USENIX Security '24), Philadelphia, PA, USA, 2024, pp. 1831–1847.

[7]. J. Yi, Y. Xie, B. Zhu, E. Kiciman, G. Sun, X. Xie, and F. Wu, "Benchmarking and Defending against Indirect Prompt Injection Attacks on Large Language Models," in Proc. 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25), Toronto, ON, Canada, 2025, pp. 1809–1820, doi: 10.1145/3690624.3709179.

[8]. L. Wang, Z. Ying, T. Zhang, S. Liang, S. Hu, M. Zhang, A. Liu, and X. Liu, "Manipulating Multimodal Agents via Cross-Modal Prompt Injection," in Proc. 33rd ACM International Conference on Multimedia (MM '25), Dublin, Ireland, 2025, pp. 10955–10964, doi: 10.1145/3746027.3755211.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15843

[9]. M. A. Ferrag, N. Tihanyi, D. Hamouda, L. Maglaras, A. Lakas, and M. Debbah, "From Prompt Injections to Protocol Exploits: Threats in LLM-Powered AI Agents Workflows," ICT Express, vol. 12, no. 2, pp. 353–383, 2026, doi: 10.1016/j.icte.2025.12.001.

[10]. Y. Liu, Y. Jia, J. Jia, D. Song, and N. Z. Gong, "DataSentinel: A Game-Theoretic Detection of Prompt Injection Attacks," in Proc. 46th IEEE Symposium on Security and Privacy (SP '25), San Francisco, CA, USA, 2025, pp. 2190–2208, doi: 10.1109/SP61157.2025.00250.

[11]. J. Damacena Duarte, G. D. Candido, J. R. A. De Britto Filho, J. Souza Neto, E. J. Costa, J. P. J. Da Costa, and L. Peotta De Melo, "A Systematic Review of Prompt Injection Attacks on Large Language Models: Trends, Taxonomy, Evaluation, Defenses, and Opportunities," IEEE Access, vol. 14, pp. 12875–12899, 2026, doi: 10.1109/ACCESS.2026.3656849.

[12]. T. Geng, Z. Xu, Y. Qu, and W. E. Wong, "Prompt Injection Attacks on Large Language Models: A Survey of Attack Methods, Root Causes, and Defense Strategies," Computers, Materials & Continua, vol. 87, no. 1, pp. 1–10, 2026, doi: 10.32604/cmc.2025.074081.

[13]. S. Gulyamov, S. Gulyamov, A. Rodionov, R. Khursanov, K. Mekhmonov, D. Babaev, and A. Rakhimjonov, "Prompt Injection Attacks in Large Language Models and AI Agent Systems: A Comprehensive Review of Vulnerabilities, Attack Vectors, and Defense Mechanisms," Information, vol. 17, no. 1, p. 54, 2026, doi: 10.3390/info17010054.

