Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 7, No. 1, 2026

Research on Information Security Vulnerability Testing and Verification Methods and Protection Countermeasures for LLM Applications in Intelligent Cockpits

Ziyi Wang, Yihong Qin *, Benyin Wu, Xuesong Wu

CATARC Intelligent Technology (Tianjin) Co., Ltd. Tianjin, China

* **Corresponding author**: (Email: qinyihong@catarc.ac.cn)

Abstract: In recent years, the application of large language models (LLMs) in intelligent cockpits has been continuously deepened, significantly improving the natural human-vehicle interaction experience. However, the generative characteristics of models, multi-modal data interaction, and access to vehicle control privileges have also introduced a series of new information security vulnerabilities such as prompt injection, data privacy leakage, unauthorized vehicle manipulation, and model hallucinations. Traditional in-vehicle security testing systems mostly target vulnerabilities in in-vehicle operating systems, in-vehicle buses, and network communications, and are difficult to adapt to the semantic-level and generative security risks brought by LLMs. To address this issue, this paper classifies and analyzes the security vulnerabilities of LLM applications in intelligent cockpit scenarios, constructs a vulnerability testing and verification method integrating static auditing, dynamic fuzz testing, red team adversarial testing, and multi-modal boundary testing, and establishes a corresponding security evaluation index system. On this basis, a hierarchical protection strategy is proposed from five dimensions: input protection, model hardening, system isolation, data security, and compliance auditing. Verified through real-vehicle environment testing, the proposed testing method can effectively identify 92.7% of known security vulnerabilities. After deploying the protection countermeasures, the relevant attack success rate drops from 38.5% to 4.3%, which can provide technical support for the secure development, testing verification, and engineering implementation of LLMs in intelligent cockpits.

Keywords: Intelligent Cockpit, Large Language Model, Information Security, Vulnerability Testing, Protection Countermeasure.

1. Introduction

Against the backdrop of the rapid development of intelligent connected vehicles, intelligent cockpits have gradually evolved from a single in-vehicle entertainment terminal into an in-vehicle intelligent hub integrating multi-modal interaction, vehicle control, travel services, and information entertainment. With powerful natural language understanding, context awareness, and content generation capabilities, LLMs are becoming the core interaction entry point of a new generation of intelligent cockpits, widely used in scenarios such as voice assistants, in-vehicle Q&A, vehicle control command parsing, and personalized service recommendation. As the number of mass-produced models equipped with LLMs continues to increase, the interaction mode of intelligent cockpits has become more natural and smooth, but the accompanying security risks have become increasingly prominent.

Different from traditional software vulnerabilities, the security risks introduced by LLMs are characterized by semanticization, concealment, and scenario-based implementation. Attackers can construct malicious prompts to achieve command hijacking, induce the model to leak car owner privacy information, and even perform unauthorized vehicle control operations. Sensitive data memorized by the model during training and interaction may also be gradually induced to be output in multi-round conversations. The current in-vehicle information security standards and testing systems are mostly built around in-vehicle buses, in-vehicle systems, and network communications, lacking targeted testing methods for the behavioral security, semantic security, and decision-making security of AI models themselves, and protection measures are difficult to adapt to the attack characteristics of generative models. Therefore, conducting research on information security vulnerability testing and verification methods and protection countermeasures for LLMs in intelligent cockpits is of great practical significance for ensuring vehicle operation safety, car owner privacy security, and in-vehicle system compliance.

On the basis of sorting out the typical security vulnerabilities of LLMs in intelligent cockpits, this paper designs a multi-dimensional vulnerability testing and verification system, establishes quantifiable security evaluation indicators, proposes a hierarchical protection scheme combined with the in-vehicle system architecture, and finally verifies the effectiveness of the method through real-vehicle testing, forming a set of security technical solutions that can be directly used in engineering practice.

2. Classification and Cause Analysis of Security Vulnerabilities of LLMs in Intelligent Cockpits

The security vulnerabilities of LLMs in intelligent cockpits are the result of the combined effect of the model's own characteristics, in-vehicle deployment environment, system privilege architecture, and interaction mode. Combined with real-vehicle testing scenarios, typical vulnerabilities can be divided into four categories: prompt security vulnerabilities,

data and privacy security vulnerabilities, system and privilege vulnerabilities, and model behavior and content security vulnerabilities, and the hazard levels are classified according to their impact scope, as shown in Table 1.

Table 1. Hazard Level Classification of LLM Security Vulnerabilities in Intelligent Cockpits

Vulnerability Level	Typical Vulnerabilities	Impact Scope	Hazard Degree
Extremely High Risk	Unauthorized vehicle start/unlock, control domain penetration	Driving safety, personal and property safety	Directly endangers safety
High Risk	Prompt injection, mass privacy leakage, model backdoor	Privacy leakage, function out of control	Seriously damages rights and interests
Medium Risk	Local data theft, harmful content generation, hallucination commands	Information leakage, abnormal experience	Moderate risk
Low Risk	Non-sensitive information leakage, false rejection of legitimate requests	Experience, compliance	Slight impact

Prompt security vulnerabilities are mainly manifested as command hijacking, prompt injection, role confusion, and adversarial suffix attacks. Due to the high sensitivity of LLMs to context commands, attackers can construct special statements to induce the model to ignore security constraints and then perform dangerous operations. In unhardened cockpit systems, such attacks are low-cost and high-success-rate, making them the main current risk point.

Data and privacy security vulnerabilities are concentrated in the model's memory and leakage of sensitive information. During operation, intelligent cockpits continuously collect car owner's mobile phone number, location trajectory, voiceprint information, vehicle identification code, address book and other data. The model may retain relevant context in multi-round conversations and output un-desensitized information under induced questions. At the same time, incomplete data desensitization during the interaction between the cockpit and the cloud model will also cause privacy diffusion.

System and privilege vulnerabilities mainly stem from unclear boundaries between the cockpit domain and the vehicle control domain and the lack of a privilege verification mechanism. When providing voice vehicle control capabilities, LLMs usually need to establish an interface connection with the vehicle control system. If privilege management is insufficient, attackers can use model vulnerabilities to penetrate from the entertainment domain to the vehicle control domain, realizing unauthorized operations of windows, air conditioners, door locks, and even the power system, which will directly affect driving safety in severe cases.

Model behavior and content security vulnerabilities are mainly manifested as model hallucinations, harmful content generation, and unexplainable decisions. Model hallucinations may cause it to output false vehicle condition information and wrong vehicle control logic, misleading drivers; illegal content generation does not meet the relevant requirements for the management of generative AI services; the opaque decision-making process also makes it difficult to trace and audit security incidents.

Overall, the core causes of LLM security vulnerabilities in intelligent cockpits include uncontrollable model generation behavior, weak input verification mechanism, excessive coupling of system privileges, insufficient data desensitization and isolation measures, and the failure of the in-vehicle security protection system to adapt to AI risks in a timely manner.

3. Testing and Verification Method for Security Vulnerabilities of LLMs in Intelligent Cockpits

This paper constructs a four-dimensional vulnerability testing and verification system covering static auditing, dynamic testing, red team confrontation, and multi-modal verification, comprehensively identifying the security risks of LLMs in intelligent cockpits through systematic use case design and quantitative indicator evaluation.

Static security auditing is mainly carried out when the model is not running, including model architecture auditing, training data compliance inspection, privacy desensitization verification, security rule integrity auditing, and vehicle control interface privilege configuration inspection. Through a combination of static analysis tools and manual verification, underlying hidden dangers such as data poisoning, backdoor logic, excessive privilege opening, and sensitive information hard coding are identified to reduce security risks from the source.

Dynamic testing is the core link of vulnerability verification, mainly including prompt fuzz testing, boundary anomaly testing, and privacy leakage testing. Fuzz testing uses automated tools to construct multi-type and multi-variant prompt attack use cases, simulating scenarios such as command injection, privacy induction, vehicle control unauthorized access, and adversarial bypass to continuously detect whether the model has abnormal responses. Boundary anomaly testing focuses on system stability under extreme conditions such as over-length input, special characters, weak network offline, and high concurrency. Privacy leakage testing usually uses watermark implantation, reverse query, cross-session tracking and other methods to verify whether the model will recall and output sensitive historical information during interaction.

Red team adversarial testing is carried out in a real-vehicle or hardware-in-the-loop environment. Testers simulate real attackers, gradually using model vulnerabilities to attempt unauthorized vehicle control and data theft through steps such as information collection, privilege breakthrough, and in-depth penetration, comprehensively testing the system's protection capability under combined attacks, which is closer to real security threat scenarios.

To achieve quantitative evaluation of test results, this paper establishes a multi-dimensional security evaluation index system, setting observable and comparable indicators from the aspects of prompt security, data privacy, system privileges, model behavior, and emergency response, as shown in Table 2.

Table 2. Security Evaluation Index System for LLMs in Intelligent Cockpits

Dimension	First-Class Indicator	Second-Class Indicator	Weight	Evaluation Standard
Prompt Security	Anti-injection Capability	Attack Success Rate (ASR)	15%	ASR < 5% is excellent
	Accuracy of Refusal	Compliance Refusal Rate/False Rejection Rate	10%	Refusal Rate > 95%, False Rejection < 3%
Data Privacy	Privacy Protection Capability	Leakage Detection Rate, Desensitization Effect	15%	Leakage Rate < 0.5%
	Data Privilege Control	Unauthorized Access Times, Least Privilege	10%	Unauthorized Times = 0
System Privilege	Vehicle Control Security	Unauthorized Manipulation Success Rate, Isolation Effectiveness	15%	Success Rate = 0
	Domain Isolation Capability	Cross-domain Penetration Success Rate	10%	Penetration Failure
Model Behavior	Robustness	Abnormal Input Stability, Hallucination Rate	10%	No Crash, Hallucination < 2%
	Compliance	Harmful Content Generation Rate, Interpretability	10%	Harmful Rate = 0
Emergency Response	Vulnerability Discovery-Repair Timeliness	5%	Response within 24h	

Comprehensive scoring of test results through the above indicators can directly reflect the system security level and provide a clear direction for protection and hardening.

4. Security Protection Countermeasures for LLMs in Intelligent Cockpits

Combined with the electronic and electrical architecture of intelligent cockpits and the risk characteristics of LLMs, this paper constructs a five-layer in-depth protection system of "Input Protection – Model Hardening – System Isolation – Data Security – Audit Emergency" to achieve full-link and full-cycle security protection.

At the input layer, user input is filtered through a multi-level verification mechanism, including keyword matching, grammatical structure detection, semantic intention recognition, input length and format constraints, and suspicious instructions are directly intercepted or require secondary confirmation. For multi-modal inputs such as voice and images, voiceprint authentication and biometric desensitization are added to reduce the risk of malicious input bypass.

At the model layer, the model's own anti-attack capability is improved through secure fine-tuning and adversarial training, enabling it to actively identify and reject malicious prompts. At the same time, a model sandbox mechanism is adopted to restrict the model from directly accessing vehicle control hardware and sensitive data. All vehicle control-related instructions must be verified by an independent security agent module, and the model is not allowed to directly perform high-risk operations. To further improve security, vehicle control privileges can be managed hierarchically, as shown in Table 3.

Table 3. Hierarchical Management of Vehicle Control Privileges in Intelligent Cockpits

Privilege Level	Operation Type	Authentication Requirement	Model Privilege
High Risk	Start, Unlock, Handbrake, Power Control	Face + Voiceprint + PIN	Model can only request, not execute
Medium Risk	Air Conditioner, Window, Light	Voiceprint/Single Authentication	Execute after verification
Low Risk	Music, Navigation, Query	None/Basic Authentication	Execute directly

At the system layer, the strong isolation mechanism between the cockpit domain and the vehicle control domain is strengthened. Cross-domain data interaction is forwarded through a security gateway, and only whitelisted instructions are allowed to pass. Key data and key operations run in a trusted execution environment, and hardware security modules are used to protect keys and authentication information, following the principle of least privilege to restrict model service privileges.

At the data layer, full life cycle protection is implemented for privacy information, pre-desensitization is performed on sensitive content, and edge-side reasoning is preferred to reduce cloud data upload. Encryption mechanisms are adopted for data storage and transmission, supporting users to manually clear conversation history to meet data security compliance requirements.

At the audit and emergency layer, a full-link log recording mechanism is established to completely archive model input/output, privilege calls, and instruction execution, and an abnormal behavior monitoring engine is used to identify risks such as high-frequency attacks and suspicious induction in real time, realizing automatic blocking and alarm, forming a security closed loop.

5. Real-Vehicle Testing Verification and Result Analysis

To verify the effectiveness of the testing method and protection countermeasures, this paper selects three real vehicles equipped with mass-produced LLM intelligent cockpits for testing. The testing environment includes a real-vehicle platform, hardware-in-the-loop simulation equipment, network packet capture tools, and an AI security testing platform, with a total of 3,200 testing use cases covering typical scenarios such as prompt injection, privacy induction, unauthorized vehicle control, and abnormal input.

The testing is divided into baseline testing and post-hardening comparative testing, and the results are shown in Table 4.

Table 4. Comparison of Test Results Before and After Protection

Testing Indicator	Before Protection	After Protection	Improvement Rate
Prompt Injection Attack Success Rate	38.5%	4.3%	88.8%
Privacy Information Leakage Rate	12.7%	0.8%	93.7%
Unauthorized Vehicle Control Success Rate	8.2%	0%	100%
Harmful Content Generation Rate	5.6%	0%	100%
Vulnerability Identification Coverage	—	92.7%	—
System Comprehensive Security Score	62.3	91.5	46.9%

The test results show that before deploying protection measures, the system has weak resistance to attacks such as prompt injection and privacy induction, and limited unauthorized vehicle control can be achieved in some scenarios; after applying the protection countermeasures proposed in this paper, the prompt injection attack success rate drops significantly, privacy leakage is effectively controlled, unauthorized vehicle control and harmful content generation are completely blocked, the comprehensive security score is significantly improved, and the false rejection rate of legitimate instructions remains at a low level, achieving a good balance between security and experience.

6. Conclusion

Aiming at the information security problems faced by LLM applications in intelligent cockpits, this paper systematically sorts out typical vulnerabilities such as prompt injection, privacy leakage, unauthorized vehicle control, and model hallucinations, analyzes their internal causes and hazard levels, constructs a comprehensive vulnerability testing and verification method including static auditing, dynamic fuzz testing, red team confrontation, and multi-modal verification, and establishes a quantitative security evaluation index system. On this basis, a five-layer in-depth protection countermeasure is proposed to achieve full-dimensional security protection from input, model, system, data to audit.

Real-vehicle test results show that this method can effectively identify most security vulnerabilities, significantly reduce the attack success rate, completely block high-risk unauthorized behaviors, and has strong engineering practicability. In the future, further research can be carried out on multi-modal joint attack testing and robustness testing in extreme environments, and the establishment of security testing specifications for in-vehicle LLMs can be promoted to provide support for the safe and reliable development of intelligent connected vehicles.

References

[1] OWASP Foundation. OWASP Top 10 for Large Language Model Security Risks, 2024.

[2] Cyberspace Administration of China. Interim Measures for the Administration of Generative Artificial Intelligence Services, 2023.

[3] Cyberspace Administration of China and other departments. Several Provisions on the Administration of Automotive Data Safety (for Trial Implementation), 2021.

[4] ISO/SAE 21434. Road vehicles—Cybersecurity engineering, 2021.

[5] China Automotive Technology and Research Center Co., Ltd. Research Report on the Testing and Evaluation System for Information Security of Intelligent Connected Vehicles, 2024.