M.TECH ACADEMIC RESEARCH PROJECT

ADVANCED LLM SECURITY ASSESSMENT REPORT

Analysis ID:	log_analysis_ab342194
Date Generated:	2025-12-11T14:44:50.058842
Framework Version:	LLM Security Framework v4.0
Research Focus:	OWASP LLM Top 10 & CVSS 4.0 Integration
MITRE ATT&CK:	Enterprise Mappings Included

OVERALL RISK ASSESSMENT

Severity: HIGH
CVSS 4.0 Score: 7.9
LLM Risk Score: 2.0
Priority: HIGH

MITRE ATT&CK; MAPPING
Techniques: T1059.007, T1564.001
Tactics: Execution, Defense Evasion
CONFIDENTIAL - ACADEMIC RESEARCH
This report contains sensitive security analysis for academic research purposes only.

EXECUTIVE SUMMARY

Metric	Value	Assessment
Prediction	LLM02_Insecure_Output	
Confidence	20.0%	
Risk Level	INFO	
CVSS 4.0 Score	7.9	
LLM Risk Score	2.0	
Overall Severity	HIGH	
Remediation Priority	HIGH	

MITRE ATT&CK; Framework Mapping:
Techniques: T1059.007, T1564.001
Tactics: Execution, Defense Evasion
Description: Insecure output handling enables code execution and defense evasion

Analysis Overview:
This comprehensive security assessment analyzed the input prompt using advanced ensemble detection methodology integrated with CVSS 4.0 scoring framework and MITRE ATT&CK; mappings. The analysis provides multi-dimensional risk assessment combining traditional vulnerability scoring with LLM-specific risk factors.

Key Insights:
- **Attack Type:** LLM02_Insecure_Output
- **Detection Confidence:** 20.0%
- **Automation Potential:** Single reusable prompt, trivially scriptable
- **Safety Impact:** No safety impact beyond baseline

Research Significance:
This assessment demonstrates the integration of traditional security scoring (CVSS 4.0) with LLM-specific risk factors and MITRE ATT&CK; framework, providing a comprehensive framework for evaluating LLM security threats in academic and enterprise environments.

CVSS 4.0 SCORING ANALYSIS

Metric	Value	Description
AV	N	Attack Vector - How the vulnerability is exploited
AC	L	Attack Complexity - Conditions beyond attacker control
PR	N	Privileges Required - Level of privileges needed
UI	R	User Interaction - Requirement for user participation
VC	L	Vulnerable System Confidentiality - Impact on confidentiality
VI	L	Vulnerable System Integrity - Impact on integrity
VA	N	Vulnerable System Availability - Impact on availability
SC	L	Subsequent System Confidentiality - Impact on other systems
SI	L	Subsequent System Integrity - Impact on other systems
SA	N	Subsequent System Availability - Impact on other systems

Score Type	Value	Severity	Vector String
Base Score	7.9	HIGH	AV:N/AC:L/PR:N/UI:R/VC:L/VI:L/VA:N/SC:L/SI:L/SA:N

LLM SUPPLEMENTAL RISK ASSESSMENT

Metric	Level	Weight	Description
Safety Impact	N	0.0	No safety impact beyond baseline
Automation Potential	H	1.5	Single reusable prompt, trivially scriptable
Value Density	H	1.5	Core proprietary asset, sensitive data

LLM Risk Score Calculation (M.Tech Formula):
Raw Product: 0.0
Normalization Constant: 7.5
Calculation: min((0.0 × 1.5 × 1.5) × 7.5, 10.0)
Final LLM Risk Score: 2.0
Severity: LOW

Research Methodology:
The LLM Supplemental Risk Assessment extends CVSS 4.0 with LLM-specific factors:

- **Safety Impact (SI):** Measures potential for harmful content generation
- **Automation Potential (AP):** Assesses attack scalability and scriptability
- **Value Density (VD):** Evaluates target model's business criticality

This multi-dimensional approach provides comprehensive risk assessment for LLM systems.

OWASP LLM TOP 10 ANALYSIS

ID	Category	Description	Status
LLM01	Prompt Injection	Manipulating LLM through crafted inputs	
LLM02	Insecure Output	LLM generates harmful content	
LLM03	Data Poisoning	Training data manipulation	
LLM04	Model Denial of Service	Resource exhaustion attacks	
LLM05	Supply Chain	Vulnerable components/dependencies	
LLM06	Information Disclosure	Sensitive data exposure	
LLM07	Plugin Abuse	Unauthorized plugin usage	
LLM08	Excessive Agency	Overprivileged model access	
LLM09	Overreliance	Uncritical trust in LLM outputs	
LLM10	Model Theft	Unauthorized model access/exfiltration	

Detailed Analysis of LLM02_Insecure_Output:
Enhanced ensemble voting: LLM02_Insecure_Output

Threat Indicators:

Context Analysis:
Text Length: 348 characters
Word Count: 36
Entropy: 4.0587
Special Characters: No
Insecure Patterns: No

MITRE ATT&CK; FRAMEWORK MAPPING

Technique ID	Name	Tactic	Description
T1059.007	Command and Scripting Interpreter: JavaScript	Execution	Adversaries may abuse JavaScript for execution.
T1564.001	Hide Artifacts: Hidden Files and Directories	Defense Evasion	Adversaries may use LLM-generated code to hide malicious artifacts.

MITRE ATT&CK; Tactical Analysis:
The detected LLM attack maps to the following MITRE ATT&CK; tactics:
- Execution
- Defense Evasion

Security Implications:
Insecure output handling enables code execution and defense evasion

Recommended MITRE Mitigations:
- MM1050
- MM1049

THREAT INTELLIGENCE ANALYSIS

Attack Category	Probability	Risk Level
LLM02_Insecure_Output	100.0%	HIGH

Advanced Threat Patterns Detected:
The ensemble model analyzed multiple threat dimensions:

- **Semantic Patterns:** Contextual understanding of attack intent
- **Syntactic Patterns:** Structural analysis of prompt construction
- **Behavioral Patterns:** Attack sequence and escalation detection
- **Contextual Patterns:** Multi-turn conversation analysis

Ensemble Advantage:
Combining multiple detection approaches reduces false positives and improves accuracy in identifying sophisticated LLM attacks.

ATTACK PATTERN ANALYSIS

Common Patterns for LLM02_Insecure_Output:
- XSS Payload: alert()
- Code Injection: System commands
- Unsanitized HTML: Direct markup rendering
- JavaScript Execution: eval() patterns
- CSS Injection: Style-based attacks
- Iframe Injection: Embedded malicious content
- Data URL Injection: data:text/html payloads

Advanced Mitigation Strategies:

Input Validation:
- Semantic analysis for intent detection
- Pattern matching for known attack signatures
- Context-aware filtering

Output Sanitization:
- Content safety classification
- Code execution prevention
- PII detection and redaction
- HTML/JavaScript sanitization

Model Hardening:
- Safety fine-tuning
- Prompt engineering
- Response filtering
- Rate limiting and usage controls

SECURITY RECOMMENDATIONS

■■ URGENT: Enhance input validation rules

■■ URGENT: Implement output content filtering

Schedule immediate security patch deployment

Conduct penetration testing for similar vulnerabilities

Update incident response procedures

MITRE ATT&CK; Based Recommendations:
- Implement MITRE mitigation MM1050
- Implement MITRE mitigation MM1049

Academic Research Recommendations:

Short-term (1-3 months):
- Implement ensemble detection in production
- Develop custom detectors for organization-specific threats
- Create automated response workflows

Medium-term (3-12 months):
- Integrate with security orchestration platforms
- Develop predictive threat intelligence
- Implement adaptive defense mechanisms

Long-term (1+ years):
- Contribute to OWASP LLM Security Standard
- Publish research findings in academic journals
- Develop open-source security tools

TECHNICAL IMPLEMENTATION DETAILS

Ensemble Model Architecture:

Base Model: RoBERTa-base (125M parameters)
Training Data: 10,000+ labeled LLM security examples
Classes: 11 (10 OWASP LLM categories + Benign)
Accuracy: 94.2% on test dataset

Detection Methodology:
- Multi-layer transformer architecture
- Attention mechanism for pattern recognition
- Contextual semantic analysis
- Threat indicator extraction

Performance Metrics:
- Inference Time: ~100ms per request
- Memory Usage: ~500MB
- Support: Batch processing capable

Current Analysis Details:

Detection Time: 8.503s
Model Used: Ensemble Detection
Ensemble Version: Complete Fused Model v1.0
Framework: PyTorch + Transformers

Research Validation:
- Cross-validation accuracy: 92.8%
- False positive rate: 3.1%
- Precision/Recall: 94.1%/93.8%
- F1-Score: 93.9%