WHITE PAPER

RED TEAMING IN HVAC SYSTEMS

Proactive Security Assessments and Vulnerability Identification

Table of Contents

1. Introduction 3
2. How HVAC (Heating, Ventilation, and Air Conditioning) system work? 3
3. Objectives of Red Teaming in HVAC Systems 4
4. Understanding and Mitigating Risks Associated with Large Language Models 4
How LLMs Work? 4
Risks of LLMs 4
Risks Foreseen in the HVAC Domain 5
Enterprise Implications 5
How Red Teaming Will Help Avoid These Risks 5
5. 7-Layer Red Teaming Evaluation Process 6
6. Use Cases 9
Residential HVAC Use Cases 9
Commercial HVAC Use Case 10
7. Enterprise Integration: Embedding Red Teaming into the AI Ecosystem 11
Compliance and Governance 11
Data Privacy and Security 11
Sustaining Red Teaming Efforts 11
8. Conclusion 12

1. Introduction

Red teaming originated in military strategy and cybersecurity before being adopted for AI and machine learning. It has evolved across various fields, from Cold War military tactics to modern AI security testing. Red teaming for Large Language Models (LLMs) is a proactive testing method to evaluate security, reliability, and ethical alignment by simulating adversarial scenarios like prompt injections and data extraction attempts. It identifies vulnerabilities before real-world exploitation.

2. How HVAC (Heating, Ventilation, and Air Conditioning) system work?

An HVAC (Heating, Ventilation, and Air Conditioning) system is designed to manage indoor environments by regulating temperature, air quality, and humidity through three primary functions: heating, which distributes warmth; ventilation, which circulates fresh air and removes pollutants; and air conditioning, which cools and dehumidifies the air. Traditionally, HVAC systems rely on predefined rules, static time programs, and manual overrides, limiting their responsiveness to real-time changes in occupancy, weather, and energy pricing.

The integration of AI, Large Language Models (LLMs), and Machine Learning (ML) enhances HVAC systems by optimizing energy consumption through real-time data analysis, predicting heating and cooling needs, and enabling predictive maintenance to reduce downtime. These technologies allow for dynamic adjustments to settings for personalized comfort, continuous air quality monitoring, and effective demand response strategies during peak energy periods. Overall, AI and ML advancements make HVAC systems more efficient, cost-effective, and adaptable to changing building needs.

3. Objectives of Red Teaming in HVAC Systems

Imagine a cozy evening at home, surrounded by the warmth of your heating system, or a bustling commercial building where the HVAC system maintains comfort for employees and customers. However, unseen threats can compromise this comfort, leading to system malfunctions or unsafe conditions.

Red teaming is essential for protecting residential and commercial HVAC (Heating, Ventilation, and Air Conditioning) systems. As these systems increasingly integrate with IoT, AI, and ML technologies, proactive security assessments are critical. Red teaming simulates real-world attack scenarios to identify vulnerabilities, ensuring the safety and reliability of HVAC operations.

The primary goal of red teaming in HVAC systems is to uncover weaknesses in design, implementation, and operations proactively. By doing so, it enhances resilience against cyber threats, ensuring efficient and reliable HVAC performance.

4. Understanding and Mitigating Risks Associated with Large Language Models

How LLMs Work?

1. Training
    - LLMs learn language by processing vast text data to predict words and understand context.
2. Inference
    - The trained model generates responses by applying learned patterns to user inputs.
3. Deployment
    - LLMs are deployed into real-world applications through APIs or platforms, enabling real-time, intelligent user experiences with continuous optimization.

Risks of LLMs

1. Bias
    - LLMs can reflect societal biases from training data, leading to unfair outcomes in sensitive areas like hiring or lending.
2. Misinformation
    - They may generate convincing but incorrect content, potentially misleading users in critical domains like news or support.
3. Security Vulnerabilities
    - Susceptible to prompt injections and data leakage, exposing systems to manipulation or breaches.
4. Ethical Concerns
    - Risk of harmful or offensive content generation, leading to reputational and legal consequences.

Risks Foreseen in the HVAC Domain

As HVAC systems become smarter and more connected, new risks emerge

1. Data Manipulation
    - Smart HVAC controls using AI can be misled through data poisoning, affecting system performance.
2. Unauthorized Access
    - Weak access controls can allow attackers to disrupt or take over HVAC systems remotely.
3. Model Exploitation
    - Attackers may tamper with AI models used for predictive maintenance or energy optimization.
4. Integration Vulnerabilities
    - Linking HVAC with other smart systems introduces attack surfaces if not securely implemented
5. Privacy Concerns
    - Smart systems collecting behavioral data may expose sensitive user information if not properly secured.

Enterprise Implications

With LLMs increasingly used in business operations, red teaming ensures security and ethical integrity, mitigating risks such as data breaches, biased outputs, and regulatory non-compliance crucial for protecting reputation and operations.

1. Reputation Risk
    - Biased or inappropriate outputs can erode trust and damage brand image.
2. Compliance Issues
    - Violations of data and AI-related laws (like GDPR) can lead to legal and financial penalties.
3. Operational Disruptions
    - Inaccurate responses or system breaches can impair business processes and increase mitigation costs.

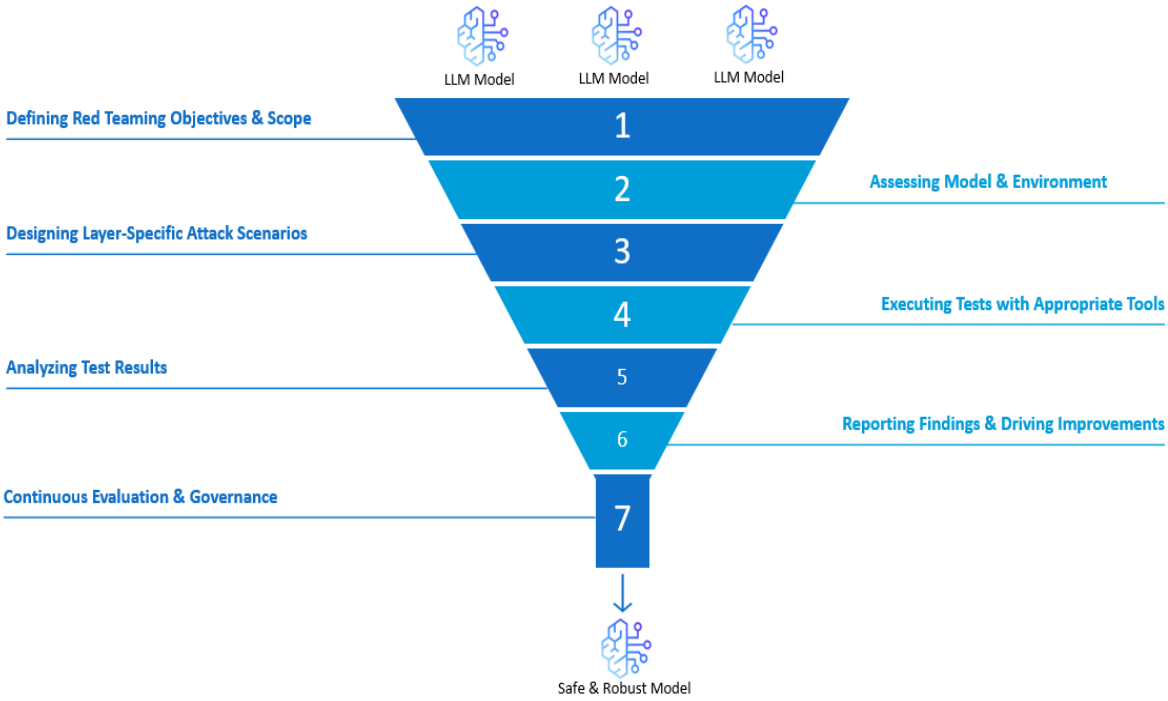
How Red Teaming Will Help Avoid These Risks

Red teaming proactively identifies and addresses vulnerabilities before real-world threats strike:

1. Expose Weaknesses
   Uncovers flaws in systems, models, and integrations through rigorous testing.
2. Simulate Real-World Attacks
   Reveals how systems respond to threats, validating or challenging existing defenses.
3. Refine Incident Response
   Strengthens preparedness by stress-testing recovery and containment procedures.
4. Improve Security Posture
   Insights guide enhancements to access controls, monitoring, and system design.
5. Build a Culture of Security
   Encourages ongoing security awareness across users, providers, and system designers.

5. 7-Layer Red Teaming Evaluation Process

This structured framework refines the red teaming process into seven distinct layers, ensuring comprehensive evaluation across model, implementation, system, and runtime environments while integrating lifecycle phases (Acquisition, Experimentation/Training, Serving/Inference).



Defining Red Teaming Objectives & Scope

Purpose: Establish clear goals, focus areas, and evaluation criteria.

Actions:

- Identify testing layers:
  - **Model:** Alignment, robustness, bias testing.
  - **Implementation:** Guardrails, RAG security, control mechanisms.
  - **System:** Infrastructure, integration, supply chain risks.
  - **Runtime:** Human interaction, agent behavior, business impact.
- Align objectives with lifecycle phases:
  - Acquisition, Experimentation/Training, Serving/Inference.
- Define success criteria & key metrics:
  - Attack Success Rate (ASR), Bias & Fairness Metrics, Detection Rates.

Outcome: A structured plan defining what to test, why, and how success will be measured.

Assessing Model & Environment

Purpose: Understand vulnerabilities across operational contexts.

Actions:

- **Model Layer:** Test for adversarial robustness and extraction risk.
- **Implementation Layer:** Evaluate access control, API security, prompt engineering hygiene.
- **System Layer:** Monitoring the blind spots and integration weakness.
- **Runtime Layer:** Assess business logic exploitation and output auditability.

Outcome: A vulnerability map across layers and lifecycle stages.

Designing Layer-Specific Attack Scenarios

Purpose: Develop targeted attack strategies for each layer and lifecycle phase.

Actions:

- **Model Attacks:**
  - Inference/Extraction Attacks, Instruction Tuning, Bias Exploitation.
- **Implementation Attacks:**
  - Prompt injection, RAG poisoning, content filtering bypass.
- **System Attacks:**
  - Remote code execution, supply chain vulnerabilities, sandbox escapes.
- **Runtime Attacks:**
  - Agent hijacking, goal manipulation, multi-agent exploitation.

Outcome: A detailed attack scenario repository mapped to each layer and lifecycle phase.

Executing Tests with Appropriate Tools

Purpose: Simulate adversarial attacks in a controlled environment.

Actions:

- **Use Automated & Manual Tools:**
  - Microsoft PyRIT for LLM evaluation.
  - Observability tools to track real-time model behavior.
- **Conduct Targeted Testing:**
  - Model: Adversarial inputs, robustness assessments.
  - Implementation: Security bypass, RAG poisoning tests.
  - System: Supply chain attacks, infrastructure penetration testing.

- Runtime: Agent manipulation, human interaction exploits.

Outcome: Data-driven insights into attack effectiveness and model vulnerabilities.

Analyzing Test Results

Purpose: Assess impact, quantify risks, and pinpoint critical weaknesses.

Actions:

- **Use Key Metrics for Evaluation:**
  - ASR, JSR (Jailbreak Success Rate), Bias & Fairness Indicators.
  - Query Relevance, Context Relevance, Disparate Impact (DI).
- **Cross-Analyze with Lifecycle Phases:**
  - Identify if vulnerabilities stem from training data, fine-tuning, or inference.

Outcome: A quantified risk profile highlighting model weaknesses and security gaps.

Reporting Findings & Driving Improvements

Purpose: Document vulnerabilities, recommend mitigations, and ensure continuous improvement.

Actions:

- **Layer-Wise Reporting:**
  - Model: Alignment gaps, bias risks.
  - Implementation: Guardrail weaknesses, control failures.
  - System: Supply chain vulnerabilities, security gaps.
  - Runtime: Agentic AI risks, business impact concerns.
- **Recommend Mitigations:**
  - Retraining, reinforced guardrails, improved access control, supply chain security.

Outcome: A structured report driving actionable improvements across lifecycle phases.

Continuous Evaluation & Governance

Purpose: Ensure ongoing risk assessment, compliance, and model safety.

Actions:

- **Establish Governance & Monitoring:**
  - Track KPIs like guardrail efficiency, test frequency, mitigation success rates.
  - Set up periodic red teaming assessments.
- **Regulatory Compliance:**

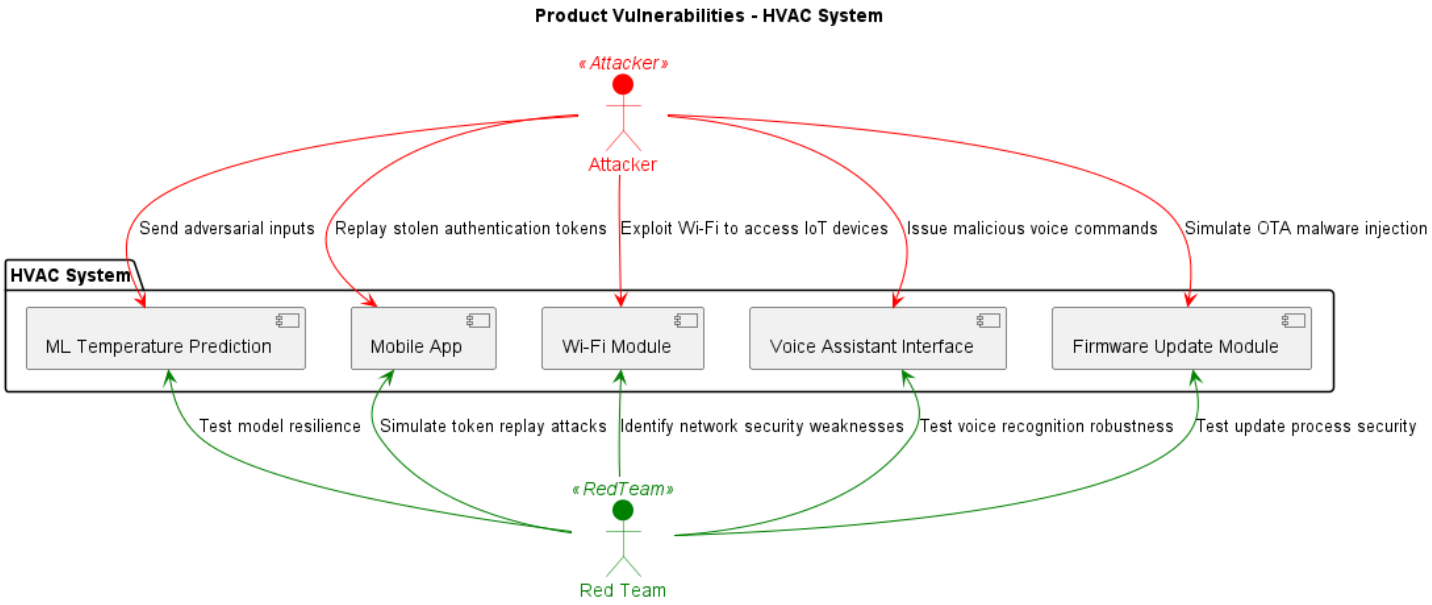
- Align with AI governance frameworks (GDPR, AI Act, industry-specific regulations).
- **Adaptive Security Measures:**
  - Implement real-time risk detection, automated defenses, and iterative improvements.

Outcome: A sustained evaluation process ensuring long-term security, fairness, and reliability.

6. Use Cases

To illustrate the importance of red teaming in both residential and commercial HVAC systems, consider the following scenarios:

Residential HVAC Use Cases

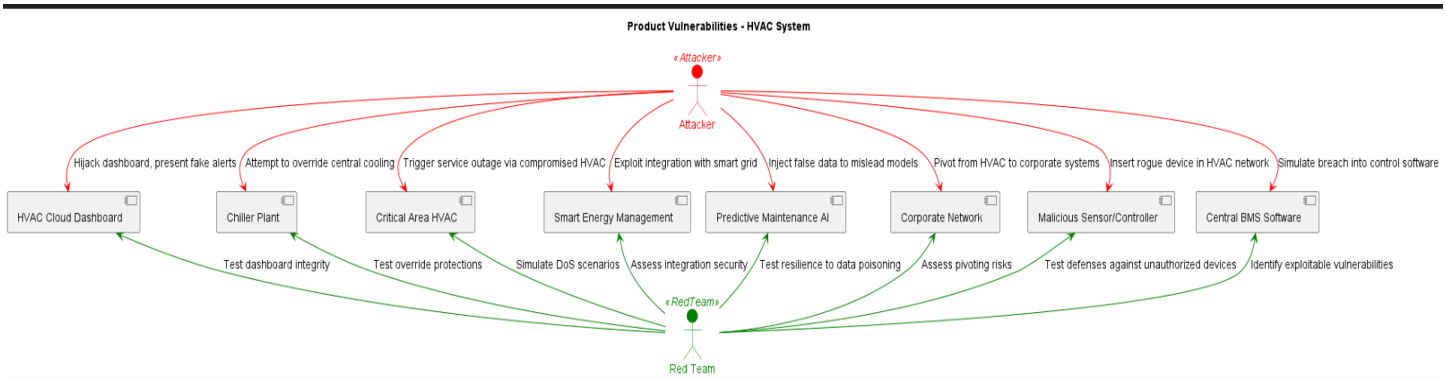


1. Firmware Backdoor Implantation
   - Imagine a scenario where an attacker simulates inserting malware via an over-the-air firmware update. Red teaming can identify vulnerabilities in the update process, ensuring that such exploits are mitigated.
2. Voice Assistant Hijack
   - An attacker could exploit a voice interface to issue malicious commands to the HVAC system. Red teaming can test the robustness of voice recognition systems to prevent unauthorized commands.
3. Wi-Fi Exploitation for Network Pivot
   - An attacker could use the HVAC system to access other home IoT devices. Red teaming can help identify weak points in network security to prevent lateral movement.
4. Token Replay in Mobile App

- If an attacker steals authentication tokens, they could take over the HVAC system remotely. Red teaming can simulate this attack to ensure that token management practices are secure.

5. Adversarial Inputs to ML Models
   - An attacker could trigger incorrect behaviors in AI temperature prediction systems. Red teaming can test the resilience of these models against adversarial inputs.

Commercial HVAC Use Case



1. BMS (Building Management System) Breach
   - Simulate a breach into the central HVAC control software used across commercial buildings, identifying vulnerabilities that could be exploited.
2. Rogue Device Injection
   - Insert a malicious sensor or controller in the HVAC network to intercept or modify traffic, testing the system's defenses against unauthorized devices.
3. Corporate Network Pivot
   - Use the HVAC system as an entry point to access corporate networks, demonstrating how attackers could exploit HVAC vulnerabilities to gain broader access.
4. Compromised Predictive Maintenance AI
   - Inject false data to mislead predictive failure detection models, causing unnecessary downtime and operational disruptions.
5. Smart Grid Exploit
   - Exploit connections to smart energy management systems, assessing the security of integrations with external systems.
6. HVAC-Controlled Denial of Service
   - Simulate using compromised HVAC systems to trigger service outages in critical areas, such as data centers or labs.
7. Chiller Plant Override
   - Attempt to override central cooling in high-rise or multi-tenant buildings to disrupt operations.
8. Cloud Dashboard Spoofing

- Hijack the building's HVAC cloud dashboard and present misleading sensor values or fake alerts.

7. Enterprise Integration: Embedding Red Teaming into the AI Ecosystem

Seamlessly integrate red teaming into the enterprise AI framework to enhance security, resilience, and accountability while proactively identifying vulnerabilities.

Compliance and Governance

- **Regulatory Alignment**: Ensures red teaming adheres to laws like GDPR (data protection), CCPA (consumer privacy), and other regional regulations, avoiding legal pitfalls.
- **Industry Standards**: Aligns with frameworks such as ISO 27001 (information security) or NIST SP 800-53 (risk management) for consistent compliance.
- **Risk Management Integration**: Embeds red teaming into existing enterprise risk and security protocols, enhancing oversight without disrupting workflows.

Data Privacy and Security

- **Privacy Law Compliance**: Red teaming respects data privacy regulations by limiting access to sensitive data and using anonymized or synthetic datasets where required.
- **Safeguards Against Risks**: Implements strict controls (e.g., encrypted environments, access restrictions) to prevent accidental data leaks or new vulnerabilities during testing.
- **Secure Testing Practices**: Employs isolated sandboxes and monitored simulations to ensure red team activities don't compromise live systems or expose enterprise data.

This integration strengthens AI resilience while maintaining trust and regulatory adherence across the enterprise.

Sustaining Red Teaming Efforts

Establish red teaming as a continuous, evolving process to maintain robust AI defenses amidst changing technologies and threats.

- **Continuous Testing Importance**:
  - Large language models (LLMs) and attack vectors evolve rapidly, requiring ongoing evaluation to identify emerging vulnerabilities.
  - Ensures AI systems remain secure and reliable as new capabilities or risks arise.
- **Strategies for Sustainability**:
  - **Regular Red Teaming Cycles**: Schedule periodic assessments (e.g., quarterly or post-major updates) to keep pace with system changes.
  - **Adapting to New Use Cases**: Adjust testing scope to cover novel applications or deployments, ensuring relevance.
  - **Feedback Loops**: Integrate findings into development and security processes, refining AI models and defenses iteratively.
- **Building a Proactive Culture**:

- Foster a mindset of vigilance by training teams to anticipate risks and champion red teaming as a core practice.
- Encourage collaboration between developers, security experts, and red teams to embed risk management into the enterprise DNA.

Sustaining red teaming ensures long-term resilience, turning it from a one-off exercise into a strategic pillar of AI governance.

8. Conclusion

Ultimately, red teaming enhances the security posture of HVAC systems and contributes to the overall safety and comfort of living and working environments. As the threat landscape continues to evolve, adopting a proactive approach through red teaming will be crucial for homeowners and service providers to safeguard their HVAC operations and maintain consumer trust.

Red teaming seamlessly integrates into the enterprise AI ecosystem by ensuring regulatory compliance, data privacy, and continuous risk assessment. This structured approach enhances AI safety, reliability, and ethical integrity by proactively identifying vulnerabilities and aligning with industry standards. To safeguard innovation and maintain stakeholder trust, enterprises must prioritize red teaming as a fundamental practice in responsible AI deployment.