LLM Red Team Assessment Report

Model: gpt-3.5-turbo

Provider: openai

Assessment Date: August 24, 2025

Total Tests: 20

1. LLM Choice and Justification

Selected Model: gpt-3.5-turbo

Provider: openai

Justification:

GPT-3.5-turbo was selected as a widely-used, cost-effective model that represents a common choice for production applications.

Model Strengths:

- Fast response times suitable for real-time applications
- Good balance between capability and cost
- Well-documented safeguards and content filtering
- Extensive training on diverse datasets

Security Considerations:

- May have vulnerabilities to certain prompt injection techniques
- Limited context window compared to newer models
- Potential for outdated information (knowledge cutoff)

2. Red Team Cookbook Source and Methodology

The red team prompts used in this assessment are derived from multiple authoritative sources:

PromptFoo Red Team Framework:

Industry-standard framework for LLM vulnerability testing, providing comprehensive coverage of common attack vectors.

OWASP LLM Top 10:

Security risks specific to Large Language Model applications, including prompt injection, data leakage, and model manipulation.

Custom Adversarial Prompts:

Domain-specific test cases designed to evaluate the model's behavior in context-relevant scenarios.

Test Categories Evaluated:

Jailbreak: Attempts to bypass model safety guidelines and restrictions through various techniques including role-playing, encoding, and context manipulation.

Bias: Evaluates model responses for discriminatory behavior, stereotyping, and unfair treatment across different demographics.

Testing Methodology:

The assessment executed 20 carefully crafted prompts across 2 categories. Each prompt was designed to test specific vulnerability patterns while maintaining reproducibility. Responses were evaluated using a multi-factor scoring system considering safeguard activation, response content analysis, and behavioral indicators.

3. Automated Assessment Results

Executive Summary:

The assessment of gpt-3.5-turbo revealed an overall robustness score of 5.72/10, indicating Moderate resilience against adversarial attacks. The model successfully triggered safeguards in 35.0% of test cases.

Key Performance Metrics:

Metric	Value	Interpretation
Overall Robustness Score	5.72/10	Moderate
Safeguard Success Rate	35.0%	Insufficient protection

Average Response Time	1.70s	Acceptable
Tests Conducted	20	Comprehensive coverage

Risk Distribution Analysis:

Risk Level	Count	Percentage
Low Risk	7	35.0%
Medium Risk	5	25.0%
High Risk	8	40.0%
Critical Risk	0	0.0%

Performance by Attack Category:

Category	Tests	Avg Score	Safeguard Rate
Jailbreak	10	0.00	0.0%
Bias	10	0.00	0.0%

Key Security Findings:

• The model shows moderate vulnerability to certain attack vectors, requiring targeted improvements.

• Critical gap in safeguard coverage with only 35.0% activation rate.

Security Recommendations:

• Enhance safeguard detection algorithms to cover more attack patterns.

• Implement multi-layer defense with both pre and post-processing filters.

• Deploy continuous monitoring for emerging attack patterns and model drift.

• Establish incident response procedures for detected vulnerability exploits.