You're using the community edition of Promptfoo's red teaming suite. For advanced capabilities and support, learn more about **Promptfoo Enterprise**.

LLM Risk Assessment: weOptimize AI B2B Multi-LLM Assistant- Baseline Red Teaming Promptfoo Evaluation

September 1, 2025

Target: weOptimize-ai-b2b-multi-LLM-assistant **Depth:** 390 probes

Critical	High	Medium	Low
0	0	0	0
issues	issues	issues	issues

Attack Methods

Baseline Testing

Original plugin tests without any additional strategies or optimizations

0.0%

0 / 390 attacks succeeded

Security & Access Control

Data protection, access control, and system security risks.

0 failed probes

60/60 passed

Resource Hijacking	✓
Privacy Violation	✓
PII via API/Database	✓
PII via Direct Exposure	✓
PII via Session Data	✓
PII via Social Engineering	✓

Compliance & Legal

Regulatory compliance, legal, and policy violation risks.

0 failed probes

160/160 passed

- Unauthorized Commitments
- WMD Content
- IP Violations
- Malicious Code
- Cybercrime
- Illegal Activity
- Methamphetamine Content
- Drug-Related Content
- Weapons Content
- IP Theft
- Non-Violent Crime
- Sexual Crime Content
- Unauthorized Advice
- Dangerous Activity Content
- Violent Crime Content
- Improvised Explosive Devices

Trust & Safety

Harmful, inappropriate, or offensive content generation risks.

0 failed probes

130/130 passed

Age Bias	✓
Disability Bias	✓
Gender Bias	✓
Race Bias	✓
Child Exploitation	✓
Graphic Content	✓
Harassment	✓
Hate Speech	✓
Personal Attacks	✓
Profanity	✓
Extremist Content	✓
Self-Harm	✓
Explicit Content	✓

Brand

Output reliability, accuracy, and brand reputation risks.

0 failed probes

50/50 passed

Excessive Agency	✓
Hallucination	✓
Disinformation Campaigns	✓
Resource Hijacking	✓
Political Bias	✓

Vulnerabilities and Mitigations

Columns Filters Density

Type	Description
Gender Bias	Gender-related bias detection
Age Bias	Age-related bias detection
Disability Bias	Disability-related bias detection
Race Bias	Race-related bias detection
Excessive Agency	Tests for unauthorized actions beyond defined system b
Hallucination	Tests for fabrication of false or misleading information
Unauthorized Commitments	Tests for unauthorized contractual commitments and leg
Weapons Content	Tests handling of weapons-related content
Hate Speech	Tests handling of hate speech and discriminatory conter
Illegal Activity	Tests handling of general illegal activities

Framework Compliance (5/5)

0.0% Attack Success Rate (0/370 tests failed across 37 plugins)

MITRE ATLAS

Framework Results 0 / 39 failed

Passed:

Hate Speech	0%
Child Exploitation	0%
Self-Harm	0%
WMD Content	0%
Privacy Violation	0%
Resource Hijacking	0%
PII via API/Database	0%
PII via Direct Exposure	0%
PII via Session Data	0%
PII via Social Engineering	0%
Sexual Crime Content	0%
Violent Crime Content	0%
Excessive Agency	0%
Hallucination	0%
Unauthorized Commitments	0%
Weapons Content	0%
Illegal Activity	0%
Drug-Related Content	0%
Graphic Content	0%
IP Theft	0%
Non-Violent Crime	0%
Explicit Content	0%
Unauthorized Advice	0%
Disinformation Campaigns	0%

Gender Bias	0%
Age Bias	0%
Disability Bias	0%
Race Bias	0%
Malicious Code	0%
Methamphetamine Content	0%
Personal Attacks	0%
IP Violations	0%
Harassment	0%
Cybercrime	0%
Political Bias	0%
Profanity	0%
Extremist Content	0%
Improvised Explosive Devices	0%
Dangerous Activity Content	0%
Indirect Prompt Injection	Not Tested
Debug Interface Exposure	Not Tested
Command Injection	Not Tested
SQL Injection	Not Tested
SSRF Vulnerability	Not Tested
Indirect Prompt Injection	Not Tested
Policy Compliance	Not Tested
RBAC Implementation	Not Tested
System Prompt Disclosure	Not Tested
System Prompt Disclosure	Not Tested
ASCII Smuggling	Not Tested
Entity Impersonation	Not Tested
ASCII Smuggling	Not Tested
Competitors	Not Tested

NIST AI RMF

Framework Results 0 / 39 failed

Passed:

Hate Speech	0%
Child Exploitation	0%
Self-Harm	0%
WMD Content	0%
Privacy Violation	0%
Resource Hijacking	0%
PII via API/Database	0%
PII via Direct Exposure	0%
PII via Session Data	0%
PII via Social Engineering	0%
Sexual Crime Content	0%
Violent Crime Content	0%
Excessive Agency	0%
Hallucination	0%
Unauthorized Commitments	0%
Weapons Content	0%
Illegal Activity	0%
Drug-Related Content	0%
Graphic Content	0%
IP Theft	0%
Non-Violent Crime	0%
Explicit Content	0%
Unauthorized Advice	0%
Disinformation Campaigns	0%
Gender Bias	0%
Age Bias	0%
Disability Bias	0%
Race Bias	0%

Malicious Code	0%
Methamphetamine Content	0%
Personal Attacks	0%
IP Violations	0%
Harassment	0%
Cybercrime	0%
Political Bias	0%
Profanity	0%
Extremist Content	0%
Improvised Explosive Devices	0%
Dangerous Activity Content	0%
Command Injection	Not Tested
SQL Injection	Not Tested
Function-Level Authorization Bypass	Not Tested
Object-Level Authorization Bypass	Not Tested
RBAC Implementation	Not Tested

OWASP API Top 10

01. Broken Object Level Authorization

2 Untested

Not Tested:

- Object-Level Authorization Bypass — Not Tested
- RBAC Implementation — Not Tested

02. Broken Authentication

2 Untested

Not Tested:

- Function-Level Authorization Bypass — Not Tested
- RBAC Implementation — Not Tested

03. Broken Object Property Level Authorization

0 / 1 plugins failed

Passed:

- Excessive Agency — 0%

Not Tested:

- Overreliance — Not Tested

04. Unrestricted Resource Consumption

0 / 3 plugins failed

Passed:

- Privacy Violation — 0%
- PII via API/Database — 0%
- PII via Session Data — 0%

05. Broken Function Level Authorization

3 Untested

Not Tested:

- Function-Level Authorization Bypass — Not Tested
- Object-Level Authorization Bypass — Not Tested
- RBAC Implementation — Not Tested

06. Unrestricted Access to Sensitive Business Flows

0 / 1 plugins failed

Passed:

- Disinformation Campaigns — 0%

Not Tested:

- Overreliance — Not Tested

07. Server Side Request Forgery

2 Untested

Not Tested:

- Command Injection — Not Tested
- SQL Injection — Not Tested

08. Security Misconfiguration

0 / 3 plugins failed

Passed:

- Privacy Violation — 0%
- PII via API/Database — 0%
- PII via Session Data — 0%

09. Improper Inventory Management

0 / 1 plugins failed

Passed:

- Unauthorized Advice — 0%

Not Tested:

- Overreliance — Not Tested

10. Unsafe Consumption of APIs

0 / 1 plugins failed

Passed:

- Privacy Violation — 0%

Not Tested:

- Debug Interface Exposure — Not Tested

OWASP LLM Top 10

01. Prompt Injection

0 / 26 plugins failed

Passed:

Hate Speech	0%
Child Exploitation	0%
Self-Harm	0%
WMD Content	0%
Privacy Violation	0%
Sexual Crime Content	0%
Violent Crime Content	0%
Weapons Content	0%
Illegal Activity	0%
Drug-Related Content	0%
Graphic Content	0%
IP Theft	0%
Non-Violent Crime	0%
Explicit Content	0%
Unauthorized Advice	0%
Disinformation Campaigns	0%
Malicious Code	0%
Methamphetamine Content	0%
Personal Attacks	0%
IP Violations	0%
Harassment	0%
Cybercrime	0%
Profanity	0%
Extremist Content	0%
Improvised Explosive Devices	0%
Dangerous Activity Content	0%

Not Tested:

Indirect Prompt Injection	Not Tested

System Prompt Disclosure	Not Tested
ASCII Smuggling	Not Tested

02. Sensitive Information Disclosure

0 / 5 plugins failed

Passed:

PII via API/Database	0%
PII via Direct Exposure	0%
PII via Session Data	0%
PII via Social Engineering	0%
Privacy Violation	0%

Not Tested:

Cross-Session Data Leakage	Not Tested
System Prompt Disclosure	Not Tested

03. Supply Chain

No Plugins

04. Improper Output Handling

0 / 8 plugins failed

Passed:

Hate Speech	0%
Disinformation Campaigns	0%
Unauthorized Advice	0%
Age Bias	0%
Disability Bias	0%
Gender Bias	0%
Race Bias	0%
Extremist Content	0%

05. Insecure Output Handling

4 Untested

Not Tested:

Command Injection	Not Tested
SQL Injection	Not Tested
SSRF Vulnerability	Not Tested
Debug Interface Exposure	Not Tested

06. Excessive Agency

0 / 1 plugins failed

Passed:

Excessive Agency	0%

Not Tested:

RBAC Implementation	Not Tested
Function-Level Authorization Bypass	Not Tested
Object-Level Authorization Bypass	Not Tested
Command Injection	Not Tested
SQL Injection	Not Tested
SSRF Vulnerability	Not Tested

07. System Prompt Leakage

0 / 5 plugins failed

Passed:

Privacy Violation	0%
PII via API/Database	0%
PII via Direct Exposure	0%
PII via Session Data	0%
PII via Social Engineering	0%

Not Tested:

RBAC Implementation	Not Tested
System Prompt Disclosure	Not Tested

08. Vector and Embedding Weaknesses

0 / 5 plugins failed

Passed:

Privacy Violation	0%
PII via API/Database	0%
PII via Direct Exposure	0%
PII via Session Data	0%
PII via Social Engineering	0%

Not Tested:

Cross-Session Data Leakage	Not Tested

09. Misinformation

0 / 3 plugins failed

Passed:

- Hallucination 0%
- Disinformation Campaigns 0%
- Unauthorized Advice 0%

Not Tested:

- Overreliance Not Tested

10. Unbounded Consumption

2 Untested

Not Tested:

- Divergent Repetition Not Tested
- Reasoning DoS Not Tested

EU AI Act

Framework Results

0 / 39 failed

Passed:

Hate Speech	0%
Child Exploitation	0%
Self-Harm	0%
WMD Content	0%
Privacy Violation	0%
Resource Hijacking	0%
PII via API/Database	0%
PII via Direct Exposure	0%
PII via Session Data	0%
PII via Social Engineering	0%
Sexual Crime Content	0%
Violent Crime Content	0%
Excessive Agency	0%
Hallucination	0%
Unauthorized Commitments	0%
Weapons Content	0%
Illegal Activity	0%
Drug-Related Content	0%
Graphic Content	0%
IP Theft	0%
Non-Violent Crime	0%
Explicit Content	0%
Unauthorized Advice	0%
Disinformation Campaigns	0%
Gender Bias	0%
Age Bias	0%
Disability Bias	0%
Race Bias	0%

Malicious Code	0%
Methamphetamine Content	0%
Personal Attacks	0%
IP Violations	0%
Harassment	0%
Cybercrime	0%
Political Bias	0%
Profanity	0%
Extremist Content	0%
Improvised Explosive Devices	0%
Dangerous Activity Content	0%
Intent	Not Tested
Command Injection	Not Tested
SQL Injection	Not Tested
SSRF Vulnerability	Not Tested
Entity Impersonation	Not Tested
Overreliance	Not Tested
Overreliance	Not Tested
Entity Impersonation	Not Tested
Overreliance	Not Tested