AI Risk Repository Chatbot: Tester Performance Evaluation Guide

# Executive Summary

This document provides a comprehensive framework for external testers to evaluate the performance of the AI Risk Repository Chatbot. It includes structured testing protocols, evaluation rubrics, and detailed feedback collection instruments to assess both objective performance metrics and subjective user experience factors.

# 1. Testing Overview

## 1.1 What You Are Testing

The AI Risk Repository Chatbot is an intelligent assistant designed to help researchers, policymakers, and practitioners explore and understand AI-related risks. The system has access to a comprehensive database of 1,612 documented AI risks categorized across 7 domains.

## 1.2 System Capabilities

**What the System CAN Do:**

* ✓ Answer questions about AI risk categories and taxonomies
* ✓ Provide statistics and data about AI risks
* ✓ Cite academic sources with direct links to documents
* ✓ Explain how risks are organized and classified
* ✓ Search across multiple dimensions (domain, entity, timing, intent)
* ✓ Process queries in multiple languages
* ✓ Provide detailed explanations of specific risks

**What the System CANNOT Do:**

* ✗ Provide mitigation strategies or solutions
* ✗ Suggest governance frameworks
* ✗ Offer benchmarks or evaluation criteria
* ✗ Make predictions about future risks
* ✗ Give opinions or subjective assessments
* ✗ Access real-time or updated information
* ✗ Provide information outside the repository scope

# 2. Performance Evaluation Framework

## 2.1 Evaluation Dimensions

|  |  |  |
| --- | --- | --- |
| Dimension | What We Measure | How to Evaluate |
| Accuracy | Correctness of information provided | Verify facts against known data |
| Relevance | How well responses match queries | Rate alignment with question intent |
| Completeness | Thoroughness of responses | Check if all aspects addressed |
| Clarity | Ease of understanding | Assess language and structure |
| Speed | Response generation time | Measure time from query to response |
| Reliability | Consistency across similar queries | Test variations of same question |
| Error Handling | Grace in handling edge cases | Try ambiguous or invalid queries |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## 2.2 Performance Scoring Rubric

Use this 5-point scale to evaluate each dimension:

**5 - Excellent:** Exceeds expectations, professional quality

**4 - Good:** Meets expectations, minor issues only

**3 - Satisfactory:** Acceptable performance, some notable issues

**2 - Below Average:** Significant issues affecting usability

**1 - Poor:** Critical failures or unusable responses

# 3. Structured Testing Protocol

## 3.1 Test Session Structure

Each testing session should follow this structured approach to ensure comprehensive evaluation:

**0-5 minutes:** Setup and familiarization

**5-20 minutes:** Category A: Basic functionality testing

**20-35 minutes:** Category B: Domain-specific testing

**35-50 minutes:** Category C: Edge case testing

**50-65 minutes:** Category D: Performance stress testing

**65-75 minutes:** Category E: Comparative evaluation

**75-90 minutes:** Documentation and feedback form completion

## 3.2 Category A: Basic Functionality Testing

Test the system's ability to handle fundamental queries:

### Test Queries:

1. "What types of AI risks exist?"

2. "How many AI risks are in the repository?"

3. "What are the 7 domains of AI risk?"

4. "Tell me about privacy risks from AI"

5. "How are AI risks categorized?"

6. "What percentage of risks are caused by humans?"

### Evaluation Criteria:

For each query, assess:

* Response time (seconds): \_\_\_\_\_\_\_\_
* Accuracy (1-5): ⭐⭐⭐⭐⭐
* Completeness (1-5): ⭐⭐⭐⭐⭐
* Citations provided: Yes ☐ No ☐
* Citations work: Yes ☐ No ☐ N/A ☐
* Language appropriate: Yes ☐ No ☐
* Would a researcher find this useful: Yes ☐ No ☐

## 3.3 Category B: Domain-Specific Testing

Test the system's depth of knowledge in specific domains:

### Discrimination & Fairness:

• "What are the main discrimination risks in AI?"

• "How many discrimination risks are post-deployment?"

• "Show me unintentional discrimination risks"

### Privacy & Surveillance:

• "What privacy risks come from AI systems?"

• "Are privacy risks mostly intentional or unintentional?"

• "How do privacy risks relate to surveillance?"

### Misinformation:

• "What types of misinformation can AI create?"

• "Are misinformation risks pre or post deployment?"

• "Who causes misinformation risks - humans or AI?"

## 3.4 Category C: Edge Case and Error Handling

Test how the system handles challenging or unusual inputs:

### Ambiguous Queries:

• "AI?"

• "Tell me about it"

• "What about risks?"

• "More"

• "Explain"

### Out-of-Scope Queries:

• "What are the best mitigation strategies?"

• "Which governance framework should I use?"

• "What will AI risks look like in 2030?"

• "How do I prevent these risks?"

• "What benchmarks measure AI safety?"

### Technical Stress Tests:

• Very long query (500+ words)

• Query with special characters: !@#$%^&\*()

• Query with typos: "Waht are teh mian rsiks?"

• Mixed language query

• Rapid sequential queries (5 in 10 seconds)

## 3.5 Category D: Performance Under Load

Evaluate system performance under various conditions:

**Simple vs Complex Queries:** Compare response times for basic vs elaborate questions

**Sequential Query Handling:** Send 10 queries in rapid succession

**Context Retention:** Ask follow-up questions to test conversation memory

**Multi-domain Queries:** Ask questions spanning multiple risk domains

**Citation-heavy Responses:** Request detailed information requiring many sources

## 3.6 Category E: Comparative Evaluation

Compare the chatbot's performance against alternative methods:

### Comparison Points:

|  |  |  |  |
| --- | --- | --- | --- |
| Aspect | This Chatbot | ChatGPT/Claude | Manual Search |
| Speed to answer |  |  |  |
| Accuracy of facts |  |  |  |
| Source citations |  |  |  |
| Domain expertise |  |  |  |
| Ease of use |  |  |  |
| Trust in results |  |  |  |

# 4. Detailed Feedback Collection Forms

## 4.1 Quantitative Performance Metrics

Please provide ratings for each performance aspect:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Metric | 1 | 2 | 3 | 4 | 5 |
| Response Accuracy |  |  |  |  |  |
| Response Completeness |  |  |  |  |  |
| Response Clarity |  |  |  |  |  |
| Response Speed |  |  |  |  |  |
| Citation Quality |  |  |  |  |  |
| Error Handling |  |  |  |  |  |
| Overall Usefulness |  |  |  |  |  |
| System Reliability |  |  |  |  |  |
| User Experience |  |  |  |  |  |
| Trust in System |  |  |  |  |  |

## 4.2 Qualitative Feedback

1. What aspects of the chatbot's performance impressed you most?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2. What aspects need the most improvement?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3. Describe any responses that were particularly good or bad:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4. What features or capabilities do you wish the chatbot had?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5. How does this compare to your expectations?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6. Would you recommend this tool to colleagues? Why or why not?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7. What would make you trust the system's responses more?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

8. Describe any technical issues or bugs encountered:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

9. How could the response format be improved?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

10. Additional comments or suggestions:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# 5. Performance Issue Classification

## 5.1 Issue Severity Levels

**Critical (Red):** System failures, incorrect core information, security issues

**Major (Orange):** Significant inaccuracies, very slow responses, broken features

**Moderate (Yellow):** Minor inaccuracies, suboptimal responses, UI issues

**Minor (Green):** Formatting issues, slight delays, preference matters

## 5.2 Issue Tracking Form

Document each issue encountered:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Issue # | Query/Context | What Happened | Expected Behavior | Severity |
| 1 |  |  |  |  |
| 2 |  |  |  |  |
| 3 |  |  |  |  |
| 4 |  |  |  |  |
| 5 |  |  |  |  |
| 6 |  |  |  |  |
| 7 |  |  |  |  |
| 8 |  |  |  |  |
| 9 |  |  |  |  |
| 10 |  |  |  |  |

# 6. Statistical Analysis Framework

## 6.1 Response Time Analysis

Track response times across different query types:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Query Type | Min (s) | Max (s) | Avg (s) | Sample Size |
| Simple factual |  |  |  |  |
| Complex analytical |  |  |  |  |
| Multi-domain |  |  |  |  |
| Taxonomy queries |  |  |  |  |
| Statistical queries |  |  |  |  |
| Edge cases |  |  |  |  |

## 6.2 Accuracy Assessment

Evaluate accuracy across different information types:

**• Numerical statistics (counts, percentages):** Accurate ☐ Mostly Accurate ☐ Some Errors ☐ Many Errors ☐

**• Domain names and categories:** Accurate ☐ Mostly Accurate ☐ Some Errors ☐ Many Errors ☐

**• Risk classifications (entity, timing, intent):** Accurate ☐ Mostly Accurate ☐ Some Errors ☐ Many Errors ☐

**• Relationships between concepts:** Accurate ☐ Mostly Accurate ☐ Some Errors ☐ Many Errors ☐

**• Technical terminology usage:** Accurate ☐ Mostly Accurate ☐ Some Errors ☐ Many Errors ☐

**• Citation relevance and validity:** Accurate ☐ Mostly Accurate ☐ Some Errors ☐ Many Errors ☐

# 7. User Journey Performance Mapping

## 7.1 Typical Research Journey

Walk through a typical research scenario and evaluate performance at each step:

**Initial Exploration:** User asks broad question about AI risks

Performance Rating: ⭐⭐⭐⭐⭐

Issues Encountered: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Domain Discovery:** User learns about the 7 risk domains

Performance Rating: ⭐⭐⭐⭐⭐

Issues Encountered: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Deep Dive:** User focuses on specific domain of interest

Performance Rating: ⭐⭐⭐⭐⭐

Issues Encountered: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Data Gathering:** User requests statistics and specifics

Performance Rating: ⭐⭐⭐⭐⭐

Issues Encountered: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Source Verification:** User checks citations and sources

Performance Rating: ⭐⭐⭐⭐⭐

Issues Encountered: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Follow-up Questions:** User asks clarifying questions

Performance Rating: ⭐⭐⭐⭐⭐

Issues Encountered: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Cross-domain Analysis:** User explores relationships between domains

Performance Rating: ⭐⭐⭐⭐⭐

Issues Encountered: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## 7.2 Performance Bottlenecks

Identify where users might encounter performance issues:

|  |  |  |  |
| --- | --- | --- | --- |
| Journey Stage | Bottleneck Type | Impact (1-5) | Suggested Fix |

# 8. Comparative Performance Analysis

## 8.1 Feature Comparison Matrix

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | AI Risk Chatbot | General AI (GPT/Claude) | Winner |
| Domain expertise |  |  |  |
| Response speed |  |  |  |
| Source citations |  |  |  |
| Factual accuracy |  |  |  |
| Query understanding |  |  |  |
| Error handling |  |  |  |
| Conversation flow |  |  |  |
| Technical depth |  |  |  |
| User friendliness |  |  |  |
| Trust/Credibility |  |  |  |

## 8.2 Use Case Suitability

Rate suitability for different use cases (1-5 scale):

• Academic research on AI risks: ⭐⭐⭐⭐⭐

• Policy development and analysis: ⭐⭐⭐⭐⭐

• General public education: ⭐⭐⭐⭐⭐

• Risk assessment for AI projects: ⭐⭐⭐⭐⭐

• Teaching and education: ⭐⭐⭐⭐⭐

• Quick fact checking: ⭐⭐⭐⭐⭐

• In-depth analysis: ⭐⭐⭐⭐⭐

• Exploratory research: ⭐⭐⭐⭐⭐

# 9. Final Performance Assessment

## 9.1 Overall Performance Summary

|  |  |
| --- | --- |
| Category | Grade (A-F) |
| Technical Performance |  |
| Response Quality |  |
| User Experience |  |
| Reliability |  |
| Domain Coverage |  |
| Overall System |  |

## 9.2 Recommendation

Based on your evaluation, would you recommend this system for:

Production use by researchers: Yes ☐ No ☐ With improvements ☐

Public deployment: Yes ☐ No ☐ With improvements ☐

Integration with other tools: Yes ☐ No ☐ With improvements ☐

Replacement of manual search: Yes ☐ No ☐ Partially ☐

## 9.3 Priority Improvements

List the top 5 improvements needed (in priority order):

1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Impact: High ☐ Medium ☐ Low ☐

Effort: High ☐ Medium ☐ Low ☐

2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Impact: High ☐ Medium ☐ Low ☐

Effort: High ☐ Medium ☐ Low ☐

3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Impact: High ☐ Medium ☐ Low ☐

Effort: High ☐ Medium ☐ Low ☐

4. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Impact: High ☐ Medium ☐ Low ☐

Effort: High ☐ Medium ☐ Low ☐

5. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Impact: High ☐ Medium ☐ Low ☐

Effort: High ☐ Medium ☐ Low ☐

# 10. Appendices

## Appendix A: Test Query Bank

Additional queries for extended testing:

### Factual Queries

• "How many risks are in each domain?"

• "What percentage of risks are AI-caused vs human-caused?"

• "Which domain has the most pre-deployment risks?"

### Analytical Queries

• "How do discrimination risks differ from privacy risks?"

• "What patterns exist in post-deployment risks?"

• "Which combinations of entity and intent are most common?"

### Exploratory Queries

• "Tell me about emerging AI risks"

• "What should policymakers know about AI risks?"

• "How has the understanding of AI risks evolved?"

## Appendix B: Performance Benchmarks

Target performance metrics for reference:

**Response Time:** <2 seconds for simple, <5 seconds for complex

**Accuracy:** >95% for factual information

**Citation Rate:** >80% of claims should have citations

**Error Rate:** <5% queries should result in errors

**User Satisfaction:** >4/5 average rating

*MIT FutureTech - AI Risk Repository | Tester Evaluation Documentation*