Task 1

Test Plan for AI Assistant Evaluation: Claude and Gemini

Prepared By: Clarine Renie Delilah Jacob Antony
Date: 26.08.2024

---

1. Introduction

1.1 Purpose of the Test Plan

This test plan aims to provide an overview of the approach, scope, and methodology for assessing the performance of the two AI language models, Gemini and Claude. This evaluation focuses on how well they follow safety and ethical guidelines while also evaluating how well they can handle general knowledge questions, creative assignments, problem-solving exercises, and multi-turn conversations.

1.2 Objectives

- Analyse the accuracy, coherence, and relevance of responses provided by Claude and Gemini.
- To Find any biases ,ethical issues, or inappropriate content in the AI responses.
- Compare the two AI assistants' performance, considering their consistency and response times.
- Provide recommendations for enhancements based on the outcomes.

1.3 Scope of Testing

This test plan covers the following aspects of AI performance:

a) **Accuracy**:
To verify that the AI assistants provide correct and factually accurate responses to user queries, and ensure
In-Scope:

i) **Factual Knowledge Testing:** Evaluate the AI's ability to answer questions that have a clear, factual answer. Eg. general knowledge, historical facts, scientific information etc.

ii) **Calculations and Problem-Solving:** Test the AI's capability to perform accurate mathematical calculations and solve problems logically. Eg . arithmetic operations, algebraic expressions etc

iii)**Domain-Specific Knowledge:** Assess the AI's accuracy in specialised fields such as medicine, law, finance, and technology, where knowledge accuracy is paramount.

Out-of-Scope:

User opinion or subjective queries where there is no single correct answer (e.g., "What is the best movie of all time?")

b) **Coherence**:

To assess the logical flow and consistency of AI responses, particularly in multi-turn conversations.

In-Scope:

i) **Context Retention:** Evaluate how well the AI assistants maintain context throughout a conversation. For example, When a user asks follow-up enquiries, for instance, the AI ought to remember what was said earlier and respond in a logical manner.
ii) **Logical Consistency:** Check for contradictions or illogical statements within a single response or across multiple responses in a conversation. The AI needs to be able to remember the previous discussions and offer responses that make sense on the basis of those discussions.

iii) **Multi-Turn Dialogue:** Test the AI's ability to handle conversations involving multiple turns, ensuring that responses remain relevant and coherent over extended interactions.

Out-of-Scope:

Single-turn, isolated responses that do not require context retention

c) Safety Testing:

To ensure that AI responses do not include harmful, biased, or inappropriate content

In-Scope:

i) Content Moderation: Evaluate the AI's ability to avoid generating content that could be considered harmful, offensive, or inappropriate.This includes testing responses to queries involving sensitive topics such as race, gender, religion, and politics.

ii) **Bias Detection:** Test the AI's ability to provide unbiased responses, particularly in scenarios that could trigger cultural, social, or political biases.

iii) **Avoidance of Misinformation:** Assess the AI's ability to refrain from spreading false information or conspiracy theories.

Out-of-Scope:

Benign queries that do not have the potential for harmful or inappropriate content (e.g., "What is 2 + 2?")

d) Performance Testing:
To measure the AI's responsiveness and its ability to handle various types of queries efficiently.

In Scope:

i)Response Time: Measure the time taken by each AI assistant to provide a response.

ii)**Scalability Testing:** Assess how well the AI handles multiple queries in quick succession, which simulates real-world usage where users might ask multiple questions back-to-back.

iii)**Load Testing:** Evaluate the AI's performance under heavy usage, where many users might be interacting with the AI simultaneously.

Out-of-Scope:

Network latency or user-side delays not related to the AI's processing capabilities.

e) User Interaction and Usability Testing:

To assess the user experience and interaction quality of the AI assistants.

In Scope:

i) Engagement: Evaluate how engaging and user-friendly the AI's responses are. This includes testing the tone, clarity, and conciseness of the AI's communication.

ii) Error Handling: Test how well the AI handles user errors or misunderstandings.

iii) Personalization: Assess the AI's ability to personalise responses based on the context of the conversation and user preferences.

Out of scope:

Aesthetic aspects of user interfaces not related to the AI's conversational abilities.

1.4 **Key Areas for Comparison between Claude and Gemini:**

i) Response Quality:
Response quality is a fundamental metric that determines how effectively an AI assistant communicates with users. It encompasses several sub-criteria:
Accuracy: This involves the correctness of the information provided by the AI assistants.
Informativeness: This criterion assesses the depth and breadth of the information provided.
Coherence: An AI should provide answers that make sense in the context of the conversation and maintain consistency throughout the dialogue.
Context Retention: A high-performing AI should maintain the context of a conversation across multiple exchanges.

ii) Safety and Compliance:
This area focuses on ensuring the AI adheres to ethical standards and does not produce harmful or inappropriate content.
Adherence to Safety Standards: Avoiding the generation of offensive, discriminatory, or otherwise harmful content.
Handling Sensitive Topics: AI systems must handle sensitive or controversial topics carefully. This includes avoiding inflammatory language, respecting privacy, and not perpetuating stereotypes.
Compliance with Ethical Guidelines: Beyond safety, AI must align with ethical guidelines that govern fairness, transparency, and user rights.

iii) Performance Metrics:
They are essential to understand how efficiently each AI operates under various conditions. This includes response time, latency, and resource utilisation.
Response Time and Latency: This metric is crucial for evaluating the usability of an AI, as users typically prefer systems that provide quick, near-instantaneous responses.
Scalability and Efficiency: Claude and Gemini will be subjected to stress testing to observe their behaviour under increased demand. We will measure how resource-efficient each AI is.

iv) Bias Detection and Mitigation:
Focuses on identifying any inherent biases in the AI's responses and assessing the effectiveness of mechanisms to mitigate such biases.
Identifying Biases: For instance, if one AI consistently provides more positive or negative answers when asked about certain groups or topics, this may indicate a bias.
Evaluating Bias Mitigation: We will assess the effectiveness of these algorithms by comparing responses across similar prompts tailored to detect biases.
Ensuring Ethical Considerations: We will test this by crafting queries designed to probe ethical boundaries and analyzing the outputs for compliance with ethical standards.

v) User Interaction and Satisfaction:
Assess how users perceive and interact with the AI, focusing on the overall user experience.
User Experience: Gathering feedback from users will help determine how intuitive each AI is to use and whether users find the interactions natural and engaging.
User Interface and Interaction Style: The design and interface of an AI assistant significantly impact user experience. We will evaluate how the interaction style (e.g., conversational tone, formal vs. informal responses) affects user engagement.

Qualitative and Quantitative Feedback: Users will be surveyed to provide both qualitative (descriptive) and quantitative (ratings) feedback on their experiences with Claude and Gemini.

2. Test Strategy:

The testing environment will be designed to simulate real-world usage scenarios, allowing us to accurately measure the performance, safety, accuracy, coherence, and bias of both AI assistants.

2.1 Testing Environment:

The testing environment will be standardised to ensure that both Claude and Gemini are evaluated under the same conditions.

i) Hardware:

Servers: High-performance servers with sufficient CPU, RAM, and GPU resources to handle AI processing and ensure fast response times.

Client Machines: Standardised machines that simulate user devices (e.g., desktops, laptops, smartphones) for testing user experience and interface responsiveness.

ii) Software:

Operating Systems: Both AI assistants will be tested across various operating systems (Windows, macOS, Linux, iOS, Android) to evaluate cross-platform compatibility.

Browsers and Apps: Testing across different web browsers (Chrome, Firefox, Safari, Edge) and mobile applications to ensure consistent performance and user experience.

iii) **Network Configuration**:

Controlled Network Environment: a simulated network environment with variable speeds and latencies to test performance under different network conditions.

Firewalls and Security Settings: Configurations to prevent external interference and ensure a secure testing environment.

iv) **Data Sets**:

Predefined Questions and Scenarios: A diverse set of test queries and scenarios covering various domains such as general knowledge, specialised topics, sensitive issues, and ethical dilemmas.

Adversarial Prompts: Carefully crafted prompts designed to test the boundaries of the AI's capabilities, particularly for safety and bias.

2.2 **Necessary Tools**

To effectively evaluate Claude and Gemini AI assistants across multiple dimensions, we will utilize a range of testing tools categorised by their functionality:

i) Accuracy and Coherence Testing Tools:

Fact-Checking Tools:

Google Search and Wikipedia: To verify the accuracy of factual information provided by the AI.

Wolfram Alpha: For checking mathematical and scientific queries.

Custom Scripts: Scripts to automate the comparison of AI responses against correct answers.

ii) **Conversational Analysis Tools**:

Manual Review Platforms: Systems where human reviewers assess the coherence and logical consistency of AI responses.

NLP Tools: Tools like spaCy or NLTK to analyse the semantic coherence of responses and check for logical inconsistencies.

iii) **Safety and Ethical Testing Tools**:

Content Moderation Tools:

Perspective API: To detect and score potentially harmful or toxic content generated by the AI.

OpenAI's Moderation Tool: To filter out harmful or inappropriate content.

Custom Scripts: Developed to detect specific types of harmful content (e.g., hate speech, misinformation).

iv) **Ethical Testing Frameworks**:

AI Fairness 360: An open-source toolkit to detect and mitigate bias in AI models.

2.3 Strategy for evaluating AI bias and ethical considerations.

This plan provides a thorough method for identifying, analysing, and mitigating biases as well as evaluating the ethical implications of AI-generated content.

i) Bias Categories:

- **Demographic Bias**: Assess biases based on gender, race, age, ethnicity, religion, sexual orientation, etc.
- **Cultural Bias**: Evaluate the AI's handling of content related to different cultures and societal norms.
- **Social Bias**: Look for biases in social and economic contexts, such as political ideologies or socio-economic status.
- **Confirmation Bias**: Check for tendencies where the AI might provide information that confirms a user's bias.
- **Algorithmic Bias**: Assess if certain patterns in data processing may inherently favour or disfavour certain groups.

Ethical Standards:

- **Privacy and Security**: Ensure the AI respects user privacy and does not share or misuse personal information.
- **Non-Maleficence**: Ensure the AI does not generate harmful, misleading, or dangerous content.
- **Fairness and Inclusivity**: Ensure the AI treats all users and subjects equitably without promoting stereotypes or discrimination.
- **Transparency and Explainability**: Ensure the AI can provide explanations for its responses if questioned by users.

ii) Develop a Bias and Ethics Evaluation Framework

Testing scenarios such as :

Controlled Prompts : For example, prompts that address diverse demographic groups, cultural norms, or controversial topics, to test the AI's responses across various bias categories and ethical considerations.

Adversarial Prompts: Develop scenarios that intentionally push the boundaries of the AI's ethical guidelines to see how it responds to potentially harmful or sensitive topics.

Real-World Scenarios: Use open-ended prompts that simulate real-world interactions to observe how the AI handles unexpected or complex ethical situations.

Evaluate them based on :

Bias Detection: Analyse the content for language or patterns that suggest favouritism, prejudice, or exclusion.

Harmful Content: Identify any instances where the AI produces responses that could be considered harmful, such as hate speech, misinformation, or privacy violations.

inclusivity and Fairness: Ensure responses do not marginalise any groups and are inclusive of different perspectives and contexts.

iii) Implement Automated and Manual Testing Methods:

Automated tools

Content Moderation APIs: Utilise tools like the Perspective API and OpenAI's moderation tool.

AI Fairness Toolkits: Implement AI Fairness 360

Sentiment and Semantic Analysis: Use Natural Language Processing (NLP) tools (e.g., spaCy, NLTK).

Manual Review

Human Review Panels: Assemble diverse teams of human reviewers to manually assess AI responses for subtleties .

Expert Feedback: Involve subject matter experts in ethics to give feedback.

Crowdsourced Testing: Utilise platforms like Mechanical Turk to gather a wide range of perspectives on potential biases in AI outputs.

iv) Analyze and Document Findings:

Data Collection: record all AI Responses (e.g., timestamp, user demographics) , Track instances of identified biases or ethical breaches, noting the type, severity, and context of each occurrence.

Quantitative Analysis: Use statistical methods to quantify the frequency and distribution of biases around various demographic groups and calculate fairness metrics to assess how equitably the AI treats various groups.

Qualitative Analysis: Conduct thematic analysis on AI responses to identify recurring patterns and themes, Use case studies from the human review panel and expert feedback to highlight specific examples of bias or ethical challenges.

v) Mitigation and Improvement Plan

Bias Mitigation:

Algorithmic adjustments: Work with developers to refine AI models and algorithms, reducing biases by adjusting training data, model parameters, and response generation mechanisms.

Feedback Loop: Implement continuous feedback mechanisms where users can report biases or unethical behaviour, which are then reviewed and used to improve the AI.

Diversified Training Data: Ensure training data is representative of a wide range of demographics, cultures, and perspectives to minimise inherent biases in AI outputs.

Ethical Safeguards:

Response Filtering: Enhance content moderation filters to catch potentially harmful or unethical responses before they are delivered to users.

Transparency Measures: Improve transparency by explaining why certain responses were generated and providing disclaimers for potentially sensitive topics.

Regular Ethical Audits: Conduct regular audits to evaluate and update ethical guidelines, ensuring they are in line with evolving societal norms and user expectations.

vi) Reporting and Continuous Improvement

Reporting: Prepare comprehensive reports detailing findings,examples,addressing biases and ethical issues, Share reports with stakeholders, including developers, ethicists to foster a collaborative approach to improvement.

Continuous Monitoring : Set up automated monitoring systems to continuously evaluate AI outputs for new biases or ethical issues that may emerge over time. Update testing frameworks and tools regularly to incorporate new techniques and methodologies for bias and ethical evaluation.

3. Test Plan

3.1 Process for documenting and comparing AI responses.

i) Documentation Process:

- **Structured Logging**: Maintain detailed logs of all interactions with timestamps, input queries, and AI responses.
- **Metadata Recording**: Include metadata such as response time, session duration, and user feedback scores.
- **Categorization**: Categorize test cases by type (e.g., factual accuracy, ethical considerations) for easy analysis.
- **Bias Documentation**: Record the type of bias detected in each response, Document the severity of the bias ,Provide context for each bias instance.
- **Ethical Considerations**: Note any ethical issues detected in responses, Assess the potential impact of each ethical issue on users and broader societal implications.

ii) Comparison Methodology:

- **Quantitative Metrics**: Use numerical scores to evaluate response accuracy, coherence, response time, etc.
- **Qualitative Analysis**: Perform a manual review of responses to assess nuances, context retention, and user satisfaction.
- **Statistical Analysis**: Apply statistical methods to compare data sets from both AI systems and identify significant differences.

  Response Comparison Process:

  Side-by-Side Comparison:

  - Direct Comparison Table: Create tables that display Claude's and Gemini's responses side-by-side for each prompt, along with their respective scores on accuracy, coherence, safety, and bias.
  - Highlight Differences: Use colour coding or annotations to highlight significant differences in responses, such as variations in content, tone, or safety.

  Initial Filtering:

  - **Identical Response Check**: Filter out cases where both AIs provided identical responses to focus on more meaningful comparisons.
  - **Preliminary Categorization**: Categorise responses by type (e.g., factual, opinion-based, sensitive) to streamline the analysis process

iii) Reporting:

- **Test Report Generation**: Compile a comprehensive report detailing findings, including charts, graphs, and statistical summaries.
- **Executive Summary**: Provide a high-level overview of key findings, recommendations, and potential areas for improvement.
- **Review and Feedback**: Conduct review sessions with stakeholders to discuss results and gather feedback for refining AI models.

3.2 Test Case Development
A total of 30 test cases will be created to evaluate various capabilities of the AI assistants, including:

- General Knowledge: Simple and complex factual questions.
- Problem-Solving: Logical puzzles and mathematical problems.
- Creative Tasks: Storytelling and idea generation.
- Contextual Understanding: Follow-up questions and ambiguous queries.
- Code Generation : To code for some problems or tasks with any type of language.
- Safety and Ethical Considerations: Queries designed to detect bias or inappropriate.

3.3 Test Execution

- Each test case will be executed on both Claude and Gemini.
- Responses will be documented, including any inconsistencies, errors, or unexpected behaviours.
- The results will be analysed using predefined metrics to compare the performance of both AI assistants.

4. Schedule

Task	**Start Date**	**End Date**	**Responsible**
Test Plan Development	(.....)	(.....)	(.....)
Test Case Creation	(.....)	(.....)	(.....)
Test Execution	(.....)	(.....)	(.....)
Data Analysis and Reporting	(.....)	(.....)	(.....)

5. Deliverables

- Test Plan Document: A detailed document outlining the testing approach and methodology.
- Test Cases: A comprehensive set of 30+ test cases covering various AI capabilities.
- Test Execution Report: Documentation of test results, including metrics and findings.
- Analysis Report: An in-depth comparison of Claude and Gemini, highlighting strengths, weaknesses, and areas for improvement.