Test Report: Multilingual Debt Collection Chatbot

1. Introduction

This test report documents the evaluation of a multilingual chatbot developed for debt collection via web and WhatsApp platforms in India. The chatbot supports six languages, English, Hindi, Marathi, Tamil, Telugu, and Hinglish and handles key user intents such as loan repayment confirmation, account closure, refusal to pay, do-not-disturb (DND) requests, and general queries.

The objective of testing was to verify that the chatbot accurately detects user intent across different languages and correctly follows the corresponding workflow. A total of 6 major workflows were tested, each containing 60 test cases (10 per language), resulting in **360 total test cases**.

Initial testing revealed multiple issues, particularly with language detection such as misclassification between similar scripts (e.g., Hindi and Marathi) and difficulty handling code-mixed inputs like Hinglish. To address this, the language detection logic was significantly improved and made more robust. Following these updates, the detected language was consistently accurate across all test cases.

Additionally, several test cases failed because the chatbot either misclassified the user's intent or failed to trigger the correct workflow path altogether. To improve intent recognition, we introduced enhanced matching logic using regular expressions, Levenshtein distance for fuzzy matching, and a more comprehensive set of prompt patterns for each workflow. These changes substantially increased the accuracy of the system, especially in handling informal, multilingual, and region-specific inputs.

This report outlines the test structure, workflows, methodology, results, and key observations from the manual testing process. The current implementation is functional and accurate, but further optimisation is planned to enhance speed, modularity, and scalability in future iterations.

2. Description of Workflows

The chatbot operates through six primary workflows, each designed to handle a specific user intent related to loan repayment and debt collection. These workflows are triggered by analysing the user’s input message and determining both the language and underlying intent. Below is a description of each workflow:

Workflow 1: Promise to Pay (PTP) Intent

This workflow is activated when a user expresses willingness to repay the loan, either immediately or at a future date. The chatbot responds positively, may prompt for a specific payment date, and updates the backend (Supabase) to reflect the user's intent to pay. Some sample prompts are - I will pay 1000 tomorrow, Will make the payment on 26th July, दोन दिवसांत पैसे देईन.

Workflow 2: Refusal to Pay

Triggered when the user denies the obligation to repay or uses language that clearly communicates refusal. The chatbot responds in a calm and non-confrontational manner while logging the refusal intent for follow-up or escalation. Some sample prompts are - अभी पैसे नहीं हैं, மீண்டும் மீண்டும் மெசேஜ் அனுப்பாதீர்கள்.

Workflow 3: Do Not Disturb (DND)

If the user asks to stop receiving messages using phrases like “stop messaging me” or “do not contact again” the DND workflow is triggered. The bot acknowledges the request, confirms compliance, and flags the user in the system to avoid future contact. Some sample prompts are - ab koi msg mat bhejna, లిస్ట్ నుండి నా నెంబర్ తీసేయండి.

Workflow 4: Escalation

This workflow handles situations where the user becomes aggressive, frustrated, or threatens legal action. The chatbot responds respectfully and de-escalates the situation, while sending a Slack notification to alert a human agent for manual review. Some sample prompts are - I will take legal action, मला कॉल करू नका, नाहीतर मी तक्रार करीन.

Workflow 5: General Testing / Neutral Conversation

This pathway covers generic or ambiguous inputs such as “hello,” “who is this?” or polite messages that do not express any clear intent. The bot responds in a neutral tone without triggering further backend updates. Some sample prompts are - 3 मंथ पहले पे किया था शायद, எனக்கு புரியலை இந்த மெசேஜ்.

Workflow 6: Account Closed / Loan Already Paid

Activated when a user states that the loan is already paid or the account has been closed. The chatbot acknowledges the message, thanks the user, and sends a notification to the collection team. The Supabase status is also updated to reflect loan clearance. Some sample prompts are - Main pehle hi payment kar chuka hoon, నా వద్ద ఎలాంటి బకాయిలు లేవు.

3. Testing Methodology

The chatbot was manually tested across six primary workflows: Promise to Pay (PTP), Refusal to Pay, Do Not Disturb (DND), Escalation, General/Neutral Conversations, and Account Closed or Loan Already Paid. A total of 360 test cases were executed across the six workflows, with input messages written in a mix of six languages, English, Hindi, Marathi, Tamil, Telugu, and Hinglish. Each workflow was tested using diverse multilingual prompts, including regionally varied and informal inputs.

Test cases were designed to simulate realistic user inputs, including formal, informal, and regionally varied phrasings. The inputs were manually entered to test them.

For each test case, the following aspects were evaluated:

- **Language detection**: Whether the system correctly identified the input language.
- **Intent recognition**: Whether the chatbot correctly classified the message into the appropriate workflow.
- **Response correctness**: Whether the bot responded in the correct tone and language based on the workflow.
- **Backend actions**: In workflows involving Supabase or Slack, whether the relevant backend update or notification was triggered.

Initially, many test cases failed not only due to language detection issues such as confusion between similar scripts (e.g., Hindi and Marathi) and informal Hinglish inputs, but more critically because the chatbot often **misclassified user intent**, leading to the wrong workflow being triggered. For example, clear messages asking to stop communication were not routed to the DND workflow, and prompts indicating loan repayment were sometimes not recognised under the account closure path.

To address these challenges, we strengthened the input processing pipeline by adding robust normalisation, fine-tuned regex-based pattern matching, and fuzzy matching using Levenshtein distance. We also expanded the phrase lists and detection logic for each workflow, ensuring better alignment between user input and system behaviour.

Final testing was conducted manually by executing each case and comparing the chatbot's response to a predefined list of expected outcomes. While the replies may not always match word-for-word, they were evaluated based on correctness of intent detection, language used, and appropriate routing to the intended workflow.

4. Summary of Results

A total of **360 test cases** were executed to evaluate the chatbot's performance across six workflows. Each workflow had **60 test cases**, covering a diverse range of user inputs across English, Hindi, Marathi, Tamil, Telugu, and Hinglish. The goal was to verify whether the chatbot correctly detects the user's intent and routes the message to the appropriate workflow path.

After final improvements to the intent detection logic, including better language normalisation, regex matching, and fuzzy logic, the chatbot was retested. The overall performance is now significantly improved, with the system accurately routing inputs in the vast majority of cases.

Workflow	Total Cases	Passed	Failed	Pass Rate
Promise to Pay (PTP)	60	52	8 (They do pass, but not with the message we want.)	0.8667
Refusal to Pay	60	60	0	1
Do Not Disturb (DND)	60	60	0	1

Escalation	60	60	0	1
General Testing	60	60	0	1
Account Closed	60	58	2	0.9667
Total	**360**	**350**	**10**	**0.9722**

Based on the final testing, the chatbot shows high reliability in detecting and handling debt-related intents across languages. While most workflows achieved near-perfect accuracy, a few minor misclassifications persist in edge-case phrasing, which will be addressed in future iterations.

5. Observations

The testing process highlighted both the strengths and limitations of the current chatbot implementation. The following observations were made:

- **Improved detection accuracy**: The enhanced intent recognition logic, including regex and fuzzy matching, significantly improved accuracy, especially for ambiguous or informal prompts.
- **Language handling robustness**: Language detection errors observed in early stages—especially between Hindi and Marathi were fully resolved post-normalization, making multilingual interaction much more reliable.
- **Workflow-specific challenges**: The **DND** and **Account Closed** workflows were the most error-prone during initial testing, largely due to the diversity of ways users express those intents. These have since been strengthened with additional phrases and matching techniques.
- **Low error severity**: In most failed cases, the bot still responded in a polite, controlled manner. Even when it misclassified intent, it avoided escalating the conversation or producing offensive outputs.
- **Scalability potential**: The current rule-based and regex-driven detection logic works well for controlled cases but may require more dynamic, learning-based models for scaling to larger user bases with more linguistic variance.

6. Limitations and Future Work

Despite strong results in manual testing, the current version of the chatbot has several limitations that constrain its real-world scalability and performance:

- **Lack of real user feedback**: All testing was conducted internally with curated prompts. There is currently no data on how real users interact with the system in production, which limits our understanding of performance under unpredictable conditions.
- **No automated scoring or confidence mechanism**: The system does not yet include confidence scores or fallback mechanisms for uncertain cases. As a result, ambiguous inputs might still lead to misclassification.

- **Rule-based logic limitations**: Although the detection logic is robust, it is still largely based on regular expressions, keyword matching, and fuzzy similarity. This limits the chatbot's flexibility in understanding complex or unexpected inputs.
- **No payment or reminder integration**: While the bot can handle debt-related communication, it currently lacks automated payment reminder scheduling and integration with payment gateways.
- **Incomplete test result tracking**: The pass/fail data was recorded manually, and real-time analytics or performance dashboards are not yet implemented.

Future Work

To overcome these limitations and further strengthen the chatbot, the following improvements are planned:

- Integrating user feedback tracking to learn from real-world interactions and improve intent handling accuracy.
- Adding a scoring or decision module to help determine when the chatbot should escalate to a human.
- Exploring lightweight ML-based classification models to enhance intent detection beyond pattern matching.
- Implementing real-time dashboards and analytics to monitor chat success rates, intent coverage, and workflow effectiveness.
- Optimising current workflows for speed and modularity, with a focus on scalable deployment across organisations.