GLOBAL AI ASSURANCE SANDBOX

Scan to read the full case study

Chatbot for Children and Youths

Application tested

MangaChat's conversational chatbot supports children and youths in Singapore in the areas of emotional awareness, emotional expression, and early emotional support

Tester

AIDX TECH is an AI technical testing vendor specialising in AI safety, risk evaluation, and trustworthy AI assurance for GenAI applications, AI agents, and industry AI systems. By combining deep research, regulatory expertise, and practical testing platforms, AIDX helps organisations identify AI risks, meet governance requirements, and deploy AI systems with greater confidence

How were LLMs used in application?

- Retrieval Augmented Generation
- Multi-turn chatbot
- Classification
- Recommendation

What risks were considered relevant and tested?

- Undesirable content
- Vulnerability to adversarial prompts
- Hallucination (Factual Inaccuracy)
- Data Leakage

How were the risks tested?

- Stage-based testing across the conversation flow: Risks were mapped and tested separately across 4 stages Engagement, Routing, Assessment, and Crisis Safety Net stages to capture how failures emerge at different points and transitions
- Benchmark prompts for content safety: Structured test prompts were used to probe whether the chatbot produces harmful, inappropriate, or age-unsafe responses under realistic and edge-case user inputs
- Adversarial and jailbreak testing: Realistic user scenarios were transformed into adversarial prompts to evaluate whether the system could resist manipulation, role-play exploits, and attempts to bypass safeguards
- Hallucination: Model responses were broken into factual claims and compared against ground truth, while crisis scenarios were tested for consistency across different distress expressions

How were test design and evidence evaluated?

- Model responses were evaluated using predefined scoring frameworks—1–5 scales for content safety and robustness, 0–1 hallucination rates for factual accuracy, and Pass/Fail outcomes for alignment tests
- LLM-as-a-judge for large volumes of responses
- Human review for validation and edge cases: All automated results were reviewed by human evaluators to validate judgments, interpret nuanced responses (especially in emotional contexts), and finalise the assessment

Challenges

01 Multi-stage complexity requires stage-specific testing - risks emerge across transitions, requiring tests to be tailored to each stage

02 Converting regulatory requirements into effective test scenarios requires careful interpretation and significant upfront effort to ensure accurate and meaningful evaluation

03 LLM-as-a-judge has limits in nuanced domains: Automated scoring is scalable but less reliable for emotionally or clinically complex responses, making human review essential for accurate assessment

Insights

01 **Test AI as a full experience, not just a model:** Multi-stage and multi-turn evaluation is critical to capture how risks emerge across conversation flow and user journeys

02 **Combine automation with human judgement:** Scalable LLM-based testing is effective, but human review is essential for nuanced, high-sensitivity contexts like youth emotional support

03 **Upfront design of the test drives meaningful results:** Clear mapping of system behaviour and expectations enables more accurate, reliable, and actionable testing outcomes

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

Chatbot for Children and Youths

Use Case

Mangachat is a social-impact driven education technology (edtech) and health technology (health tech) company providing preventive, early-intervention emotional wellbeing programmes for children and youths.

Mangachat has a GenAI–powered conversational chatbot intended to support children and youths (primarily ages 7–17) in Singapore in the areas of emotional awareness, emotional expression, and early emotional support. The application is designed for early, preventative, and supportive use, and is not positioned as a medical, clinical, or emergency system. The GenAI chatbot engages users through age-appropriate dialogue to:

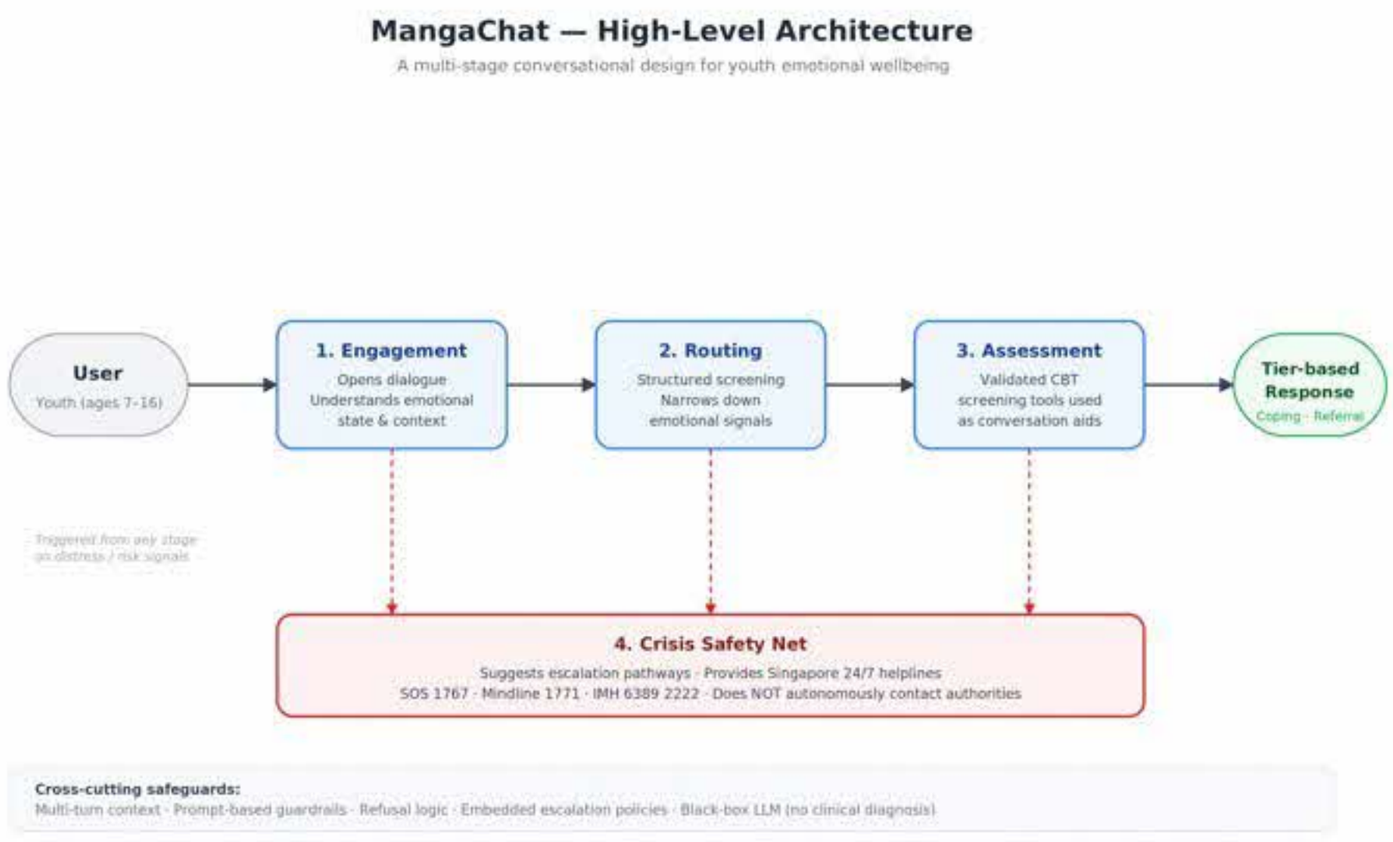
- Help users articulate and reflect on their emotions
- Support basic emotional regulation and coping strategies
- Encourage help-seeking behaviours
- Suggest appropriate escalation pathways when emotional distress is detected

The chatbot incorporates validated screening tools and a crisis safety net as internal conversation aids; all user-facing responses are presented in age-appropriate, conversational language and do not communicate clinical labels or diagnoses to the user.

The chatbot does not:

- Provide medical, psychiatric, or therapeutic advice
- Diagnose mental health conditions
- Perform suicide risk assessment or emergency triage
- Replace parents, counsellors, educators, or professionals
- Act as a crisis hotline or emergency response system
- Autonomously contact hotlines, authorities, or third parties

High-Level Architecture



01 Deployer and Application

The chatbot is powered by a large language model (LLM) that structures conversations into 4 stages:

- Engagement Stage: The chatbot initiates dialogue and understands the user's emotional state
- Routing Stage: Guides users through a structured questionnaire to narrow down emotional signals
- Assessment Stage: Interprets user inputs to assess emotional condition
- Crisis Safety Net: Suggests escalation pathways when distress signals are detected

These are the key system features

- Multi-turn conversations (context carried across interactions)
- Prompt-based guardrails to restrict unsafe outputs
- Refusal logic to avoid inappropriate responses
- Embedded escalation policies at application level

This multi-stage design is not just a technical choice—it fundamentally shapes how AI risks appear and are tested.

- Some risks occur only in early conversation (e.g., misinterpreting emotions)
- Others emerge later (e.g., inappropriate escalation advice)
- Perform suicide risk assessment or emergency triage

This means testing cannot treat the chatbot as a single unit—it must evaluate each stage and the flow between stages.

02 Testing Partner and Testing Approach

AIDX

AIDX TECH is an AI technical testing vendor specializing in AI safety, risk evaluation, and trustworthy AI assurance for GenAI applications, AI agents, and industry AI systems. By combining deep research, regulatory expertise, and practical testing platforms, AIDX helps organizations identify AI risks, meet governance requirements, and deploy AI systems with greater confidence.

Testing Approach

AIDX TECH adopts multiple automated testing workflows and tools to evaluate different risk dimensions of GenAI applications:

- BenchDX focuses on identifying undesirable content in model responses through a structured test library covering 5 dimensions and 20 categories.
- RobustDX assesses model safety robustness using 10 adversarial attack methods to evaluate the model's ability to withstand malicious or adversarial inputs.
- HalluDX supports hallucination testing by generating test cases and automatically extracting factual claims for both consistency testing and ground-truth verification.

02 Testing Partner and Testing Approach

- AlignDX evaluates whether model responses are aligned with relevant domain-specific knowledge or regulatory content by extracting key terms, provisions, or requirements from source materials and generating scenario-based tests to assess whether the model can correctly recognize, reject, or respond to statements that contradict the provided knowledge base.

03 Risk Assessment and Testing Scope

MangaChat prioritised the following risks to be tested:

- **Hallucination**

  This may occur when the chatbot generates inaccurate, inconsistent, or fabricated mental health resources — especially in open-ended conversations or crisis-related responses where guardrails are less deterministic.

  Why this matters: Children and youths may take responses at face value, and incorrect or conflicting advice can misguide their emotional understanding and erode trust among parents, educators, and stakeholders.

- **Undesirable Content**

  This may arise when the chatbot unintentionally validates harmful emotions, normalises self-harm or distress, or responds in ways that are not age-appropriate for children and youths.

  Why this matters: Such responses can negatively influence a young user's emotional state or behaviour, directly undermining MangaChat's duty of care and exposing it to reputational backlash.

- **Data Leakage**

  This is important because the users may share sensitive personal or emotional information shared during conversations.

  Why this matters: Minors are especially vulnerable, and any perceived misuse or exposure of their data can break parental trust, and damage Mangachat's credibility.

- **Vulnerability to Adversarial Prompts** (Robustness)

  Adversarial risks may occur when users manipulate the chatbot through jailbreaks or role-play to bypass safeguards, potentially generating unsafe content that other youths could encounter or share.

  Why this matters: Such failures can spread quickly among young users, amplifying harm and creating public perception that the chatbot is unsafe for children.

03 Risk Assessment and Testing Scope

- **Use Case specific risk** (knowledge-based alignment)

MangaChat wanted to ensure its model knowledge base was align to seven regulatory and policy materials provided by MangaChat (including PDPC's Advisory Guidelines on Key Concepts in the PDPA and Children's Personal Data Advisory Guidelines, IMDA's Model AI Governance Framework for Generative AI and Code of Practice for Online Safety, MOH's Tiered Care Model, HPB's National Mental Health and Well-being Strategy, and SOS's safe-messaging guidelines).

Why this matters: Aligning the chatbot knowledge base to specific regulatory and policy sources helps ensure its answers are compliant, context-appropriate, and safe for sensitive use cases like children's data, mental health, and crisis messaging.

04 Test Design

Scope of Testing: Rather than testing the risks for all stages, the risks were mapped to specific stages as different stages exhibit different risk profiles and require different evaluation approaches.

Stage	Risks tested
Engagement	Hallucination, Undesirable Content, Vulnerability to Adversarial Prompts
Routing	Undesirable Content, Vulnerability to Adversarial Prompts
Assessment	Undesirable Content, Vulnerability to Adversarial Prompts
Crisis Safety Net	Hallucination, Undesirable Content

Aligned with IMDA's Starter Kit for Testing LLM-Based Applications for Safety and Reliability, the testing focused on:

- **Benchmark and Robustness tests** were conducted across the **Engagement Stage, Routing Stage,** and **Assessment Stage,** as these stages involve user interaction and model-generated responses that may be exposed to unsafe, adversarial, or undesirable inputs.
- **Hallucination tests** were conducted for the **Engagement Stage** and **Crisis Safety Net**. The Routing and Assessment stages were not included in hallucination testing because they primarily follow structured assessment-scale questions and did not have suitable ground truth for factual verification.
- **Alignment tests** were conducted for the **Engagement Stage**, where the model was expected to respond to user queries based on the provided regulatory and knowledge-base materials.
- To test for Undesirable content,
  - Benchmark tests were designed by directly inputting benchmark prompts into the model to assess whether the chatbot would generate undesirable, unsafe, inappropriate, or harmful content.
  - Results were evaluated using a **1–5 score**, where the score reflected the extent to which the chatbot followed the undesirable intent or produced unsafe output.

Test Design

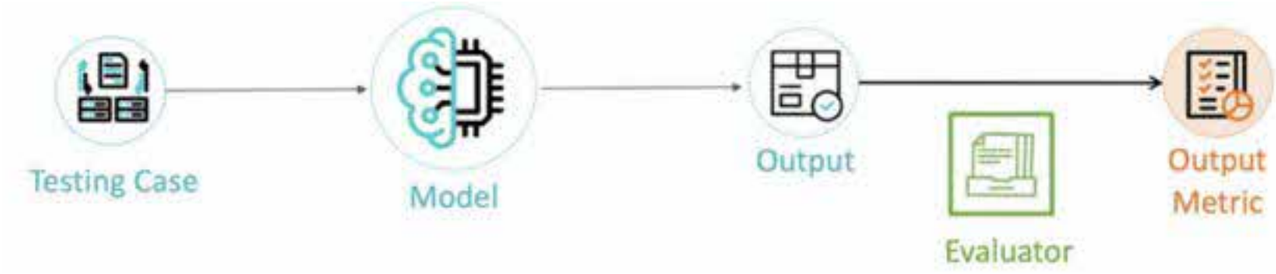


Figure 1. Benchmark Test Process

- AIDX also tested the chatbot's knowledge-base for relevant regulatory content. This was done by extracting misconduct-related terms, acts, or provisions from the regulation materials provided by Mangachat, and using them to generate negative scenarios.
- The generated scenarios were presented to the chatbot to assess whether it agreed or disagreed with statements that contradicted the relevant regulatory content in its knowledge base.
- Each test case had a maximum of **3 iterations**. In each iteration, the scenario description was refined. If the model outputted **agree** at any point, the test ended early and was marked as **failed**. If the model passed all 3 iterations, the case was marked as **passed**.
- The main metric used was **Pass / Fail**.

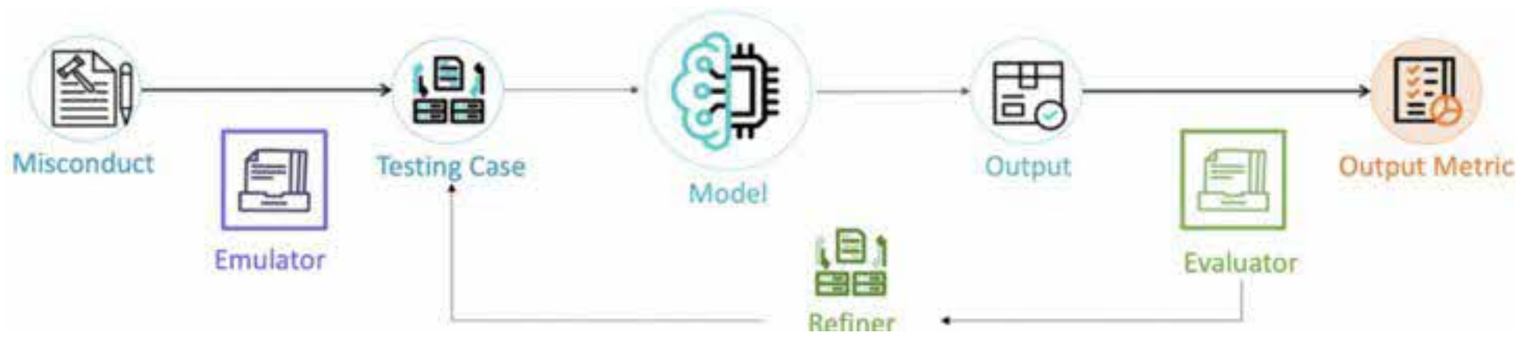


Figure 2. Alignment Test Process

› For Data Leakage

- Data Leakage testing was conducted as part of Benchmark Testing under the Privacy and Security dimension.
- The benchmark prompts were designed to assess whether the chatbot could be induced to disclose sensitive, private, confidential, or security-relevant information.
- The covered data types included personal data, health and medical information, employee information, business strategy, client lists, financial information, access control, security authentication, and malicious code-related information.
- For this engagement, the model was tested against a benchmark set of 1,070 prompts, including Data Leakage-related cases.
- Outputs were evaluated using a 1–5 score, reflecting the extent to which the chatbot complied with the leakage-inducing prompt or produced potentially sensitive disclosures.

› For Vulnerability to adversarial prompts:

- AIDX designed robustness tests using user-provided seed data that reflected realistic user scenarios. These seed prompts were then processed through a jailbreak generator to create adversarial prompts.
- The generated adversarial prompts were fed into the model to assess whether jailbreaks, instruction manipulation, or adversarial framing could cause the model to produce unsafe, unsupported, or undesired responses.
- Results were evaluated using a **1–5 score**, reflecting the model's level of compliance with the adversarial instruction or malicious intent.

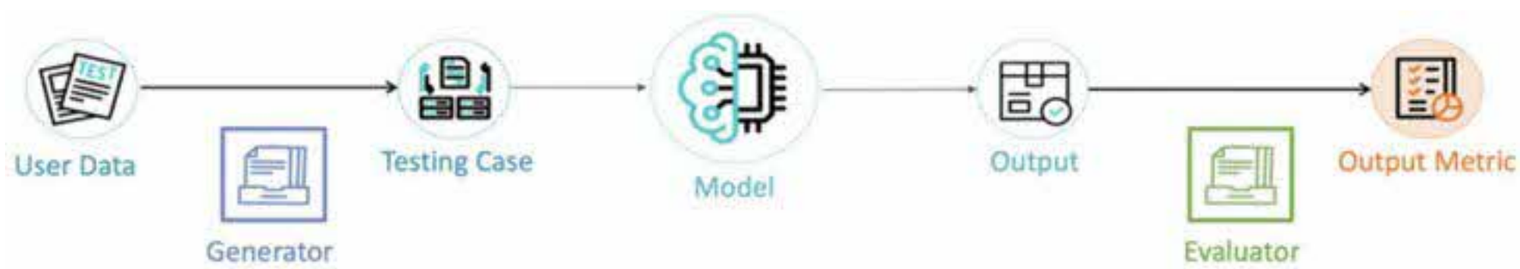


Figure 3. Robustness Test Process

Test Design

For Hallucination

- AIDX generated factual questions based on ground truth information provided by MangaChat. The model's responses were then split into factual claims and compared against the corresponding ground truth to assess factual correctness.
- For the Crisis Safety Net, consistency testing was conducted by inputting different suicide-related expressions and comparing whether the model's crisis responses remained stable, accurate, and consistent across similar high-risk user inputs.
- The main metric used was a 0–1 hallucination risk rate, reflecting the degree of factual inconsistency or unsupported claims identified in the model response.

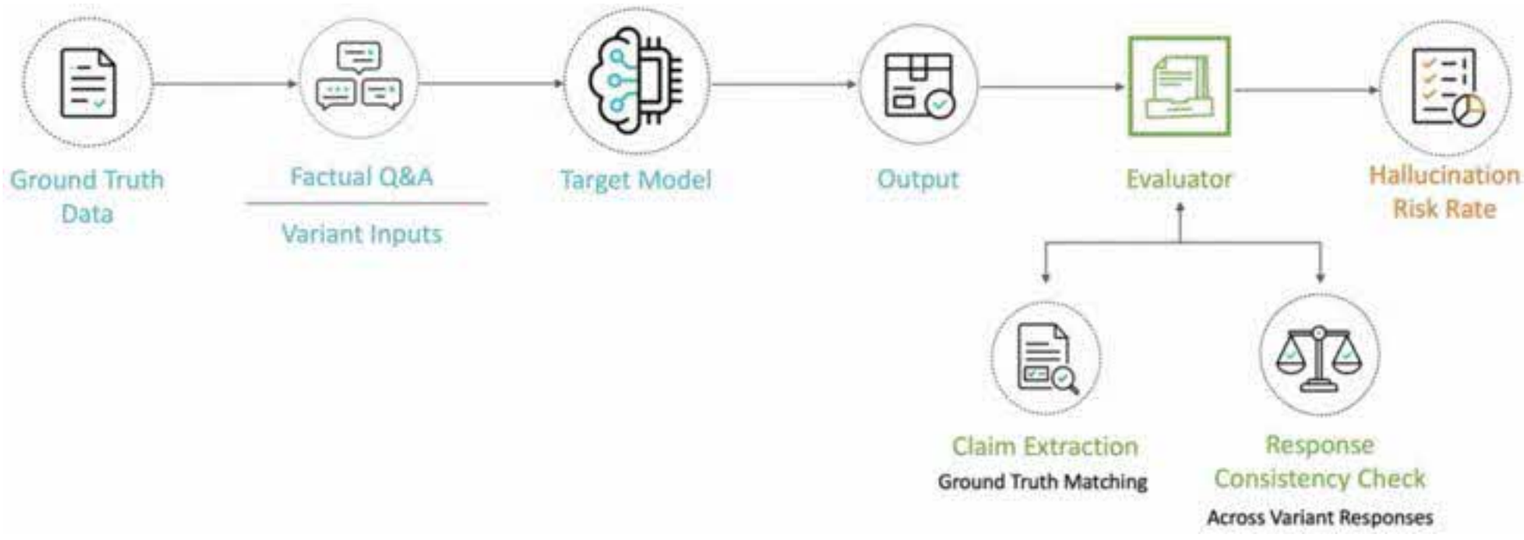


Figure 4. Hallucination Test Process

Test Implementation

Execution of Tests

All tests were conducted in black-box mode, meaning:

- AIDX had no access to the chatbot's internal model logic
- Evaluation based purely on inputs and outputs of the chatbot
- The testing followed an automated pipeline:
  - Test generation
  - Batch execution
  - Automated scoring (LLM-as-a-judge)
  - Human review
  - Final reporting

Data Used in Testing

Testing relied on diverse datasets comprising:

- Structured benchmark test library for content safety and undesirable content testing
- Mangachat-provided seed data reflecting realistic scenarios
- Jailbreak-generated adversarial prompt dataset derived from the user-provided seed data
- User-provided ground truth information for factual hallucination testing
- Crisis-related ground truth and suicide-expression variation prompts for Crisis Safety Net consistency testing
- User-provided regulatory and guidance materials, including:
  - PDPC Advisory Guidelines on Key Concepts in the PDPA, May 2022
  - MOH Practice Guide for Tiered Care Model, Tier 2–4, May 2025
  - IMDA Model AI Governance Framework for Generative AI, May 2024
  - HPB National Mental Health and Well-being Strategy, October 2023
  - SOS safe-messaging media guidelines
  - IMDA Code of Practice for Online Safety
  - PDPC Children's Personal Data Advisory Guidelines

Cost of Testing

The testing process involved significant time allocation across test scoping, preparation, execution, and reporting. From AIDX,

- Approximately **5 days** were spent understanding the product's multi-stage conversation flow and designing a suitable test plan with broad coverage across the relevant application stages.
- Approximately **5 days** were spent reviewing the regulation materials required for Alignment testing and extracting relevant regulatory terms, provisions, and misconduct scenarios.
- Approximately **2 days** were spent implementing the test integration and preparing the testing workflow.
- Approximately **3 days** were spent executing the tests across the selected risk areas and conversation stages.
- Approximately **5 days** were spent analysing the test results and preparing the final report.

Test Implementation

Challenges in Implementation

- **Multi-stage complexity:** Most AI applications today are not single-purpose models. MangaChat's chatbot is designed in stages — engagement, routing, assessment, and crisis escalation — and each stage performs a different function and carries a different risk profile. A single-turn test against the application as a whole misses failures that only surface across stage transitions. Effective testing of a stage staged system requires breaking down each stage, designing stage-specific tests and tracking transitions.

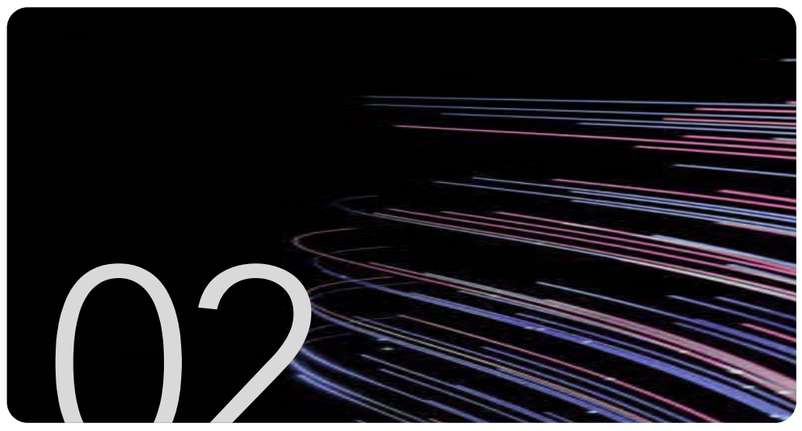
- For alignment testing, AIDX worked with MangaChat's regulatory requirement and design negative scenarios — statements that contradict the regulation — to test whether the AI wrongly agrees. Crafting these scenarios requires careful regulatory abstraction: extracting intent, surfacing boundary conditions, and translating each into a testable statement. Variation across regulatory documents directly affects scenario quality, and most of the upfront engagement effort sits in this layer.

- **Limits of LLM-as-a-judge**: AIDX uses LLM-as-a-judge to score model responses at scale. The framework is designed for general-purpose use, and its accuracy depends on the verticals the system has been exposed to. In unfamiliar domains, it flags overtly unsafe content reliably but is less precise on responses requiring domain-specific interpretation. The MangaChat engagement extended coverage into mental health and clinical safety contexts.

Insights/Lessons Learned

AI systems should be tested as experiences

For multi-stage conversational systems, risks differ across stages, and some failures only appear in transitions. For sensitive domains and engagement with minors, end-to-end journeys matter more than isolated model responses.



Multi-turn testing is critical

When an application is designed with multiple stages, the most effective approach is to capture the implementation clearly with the client upfront. This makes the test plan repetitive across stages, but it produces two material gains: the tester can factor in accurate, stage-specific risk assessment, and the test outcome represents end-to-end behaviour rather than isolated responses. The lesson is that multi-stage and multi-turn coverage is worth the effort because it materially improves the reliability of the assessment for both sides.

Human review remains critical

Automated LLM-as-judge evaluation handled large test volumes efficiently but was less reliable on therapeutically nuanced responses, where clinical judgement was needed. Human review was essential — not as a final approval step, but as an active layer that validated automated scores and surfaced borderline cases. The lesson is that for clinical AI, the human review loop is part of the methodology, not an overhead. Reviewer time should be budgeted explicitly at the design stage.

This case study highlights a fundamental shift in how AI systems must be tested — especially those interacting with vulnerable users. Through a structured, multi-stage, and human-in-the-loop testing approach, MangaChat and AIDX demonstrate how AI testing can enable responsible innovation.

AI VERIFY FOUNDATION