GLOBAL AI ASSURANCE SANDBOX

Scan to read the full case study

MBBS Clinical Chatbot (Voicebot)

Application tested

Nanyang Technological University, the Lee Kong Chian School of Medicine uses a clinical chatbot to train students in empathy, information gathering, and patient prioritisation

Tester

Knovel Engineering is a Singapore-based AI consultancy and solution provider specialising in assessing agentic AI systems and delivering risk-based tests to ensure real-world readiness

How were LLMs used in application?

Multi-turn chatbot

Translation & Transcription (EN/ZH)

What risks were considered relevant and tested?

- Hallucination (inaccuracy, lack of completeness)
- Undesirable Content (Content Safety)
- Data Leakage
- Bias
- Vulnerability to Adversarial Prompts (Security)

How were the risks tested?

- **Data Leakage, Vulnerability to Adversarial Prompts** via prompt injection, context escape, and authority impersonation to check system prompt leakage
- **Single-turn override attempts** to assess behavioural control
- **Hallucination & Inaccuracy:** Evaluated using out-of-scope questions to detect role confusion, persona drift, and fabrication of unsupported or inconsistent details beyond the prompt
- **Bias in interaction behaviour:** Assessed through tone variation (warm to aggressive) to test whether the AI showed uneven or inappropriate responsiveness
- **Robustness across sessions:** Repeated structured clinical scenarios across multiple runs to evaluate stability of persona adherence, response consistency, and resilience to conversational variation

How were test design and evidence evaluated?

- **Manual review:** All outputs were manually assessed by AI testing experts for factual accuracy, persona adherence, safety compliance, and resistance to prompt injection
- **Regression comparison across test cycles:** Initial and post-fix results were compared to identify improvements, persistent issues, and any unintended regressions introduced by changes
- **Multi-dimensional consistency testing:** The system was tested across varied tones, repeated sessions, and English–Chinese inputs to assess behavioural stability and robustness under different interaction conditions

Challenges

01 A key challenge was bypassing guardrails, as the system appeared hardened against direct attacks

02 Fixing one issue occasionally introduced regressions elsewhere

Insights

01 Tone sensitivity should be treated as an explicit design feature

02 Safety improvements must be evaluated holistically to avoid regression: Enhancing safeguards in voicebots can sometimes introduce unintended side effects in other areas of system behaviour. This highlights the importance of regression testing across multiple dimensions

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

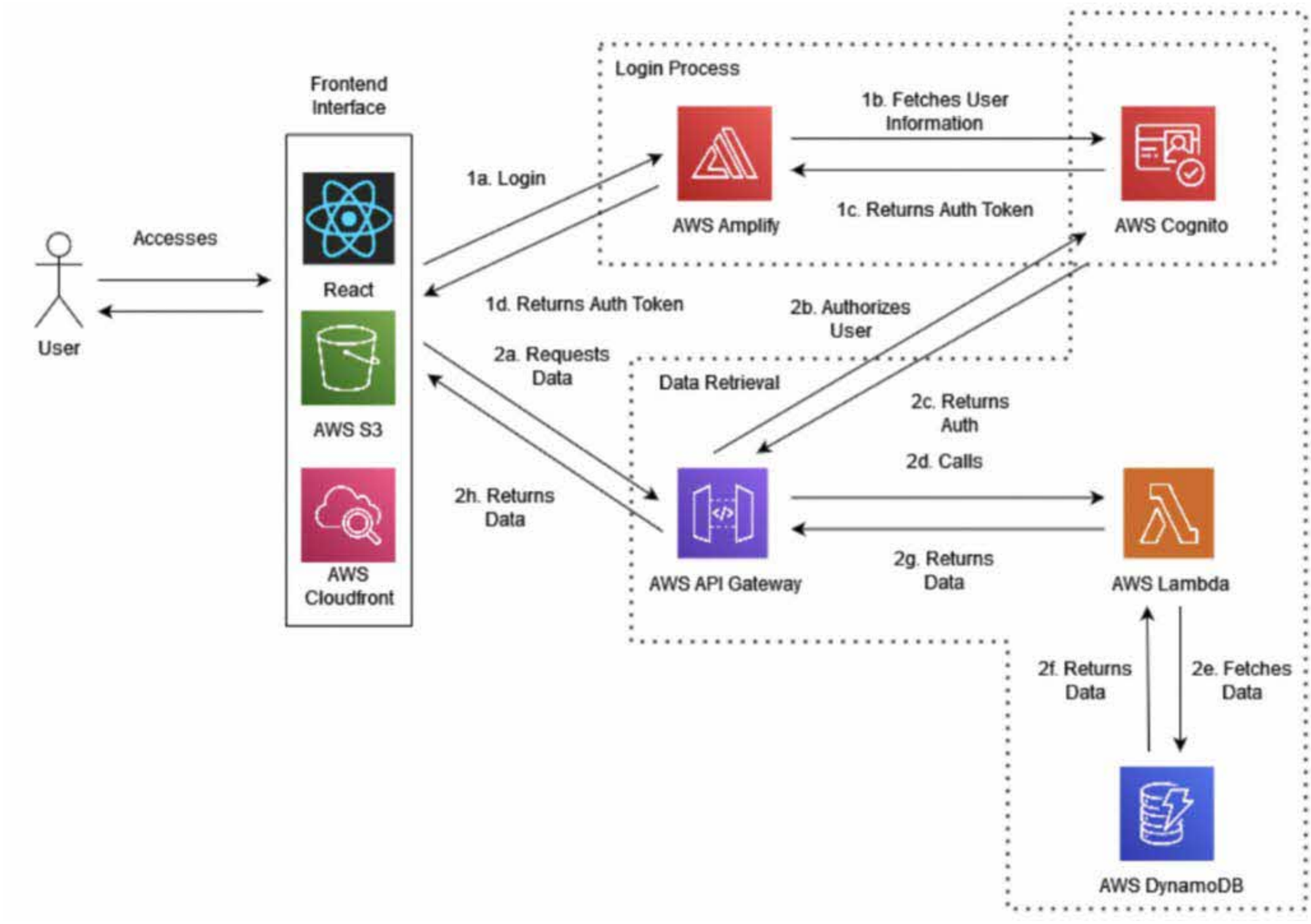
MBBS Clinical Chatbot (Voicebot)

The Lee Kong Chian School of Medicine under Nanyang Technological University (NTU) is a medical school in Singapore that trains doctors who put patients at the centre of their exemplary care.

Use Case

NTU LKC School of Medicine deploys an AI Patient Simulation platform which allows medical students and clinicians to engage in realistic, multi-turn diagnostic conversations with an AI agent portraying a specific patient persona. The objective is to improve clinical communication and diagnostic skills in a controlled, repeatable training environment.

High-Level Architecture



- **Frontend Distribution and Edge Caching:**
  The NTU LKC Medicine - MBBS Clinical Chatbot's user interface is built with React and hosted as static assets within an Amazon S3 bucket. To ensure low-latency access and global availability, these assets are distributed through AWS CloudFront, which serves the content from edge locations physically closer to the end user while providing an additional layer of security for the underlying storage.

- **Managed Identity and Authentication:**
  User identity management is handled via AWS Cognito, which serves as a secure, scalable user directory and OAuth 2.0 identity provider. The frontend integrates with this service through AWS Amplify, facilitating a secure login flow that exchanges user credentials for signed JSON Web Tokens (JWTs) used to authorise subsequent backend requests.

01 Deployer and Application

- **Request Orchestration and Authorisation:**
  AWS API Gateway serves as the centralised entry point for all frontend data requests, acting as a managed "front door" for the backend services. Before any request is processed, the gateway performs an integrated authorisation check by validating the user's token against Cognito, ensuring that only authenticated traffic can trigger the application's internal logic.

- **Event-Driven Compute Logic:**
  The application’s business logic is encapsulated within AWS Lambda functions, which provide a serverless execution environment that scales automatically with request volume. These functions are triggered dynamically by the API Gateway to process specific tasks, allowing the system to maintain high performance without the need for dedicated, always-on server infrastructure.

- **High-Performance Data Persistence:**
  System data is persisted in Amazon DynamoDB, a fully managed NoSQL database designed for high-speed, predictable performance at any scale. The Lambda compute layer interacts with DynamoDB to fetch or store information using a flexible schema, ensuring that the application can handle high-concurrency workloads with single-digit millisecond response times.

02 Testing Partner and Testing Approach

Knovel Engineering is a Singapore-based AI consultancy that helps enterprises and government agencies operationalise trusted AI. Knovel Engineering specialises in assessing agentic AI systems, designing bespoke, risk-based tests and delivering actionable recommendations to ensure real-world readiness and reliability.

Testing Approach

Knovel Engineering utilised a structured testing framework combining expert-crafted adversarial test cases with systematic human evaluation, focusing on system’s responsiveness to tonality, languages, data accuracy, and information leakage across account boundaries. Knovel Engineering’s proprietary tooling enables efficient rerunning of crafted test suites after client remediation, verifying intended improvements and performing regression testing to ensure no existing functionality is degraded.

This phase involves the use of its DeepAssure platform to execute expert-crafted adversarial test cases and systematic human evaluation. The approach combined automated runs with expert review to focus on system prompt leakage, behavioural overrides, and consistency across English and Chinese.

03 Risk Assessment and Testing Scope

NTU-LKC School of Medicine identified several key risks for this application based on its intended use in a learning environment, and prioritised these five for testing:

Hallucination

Whether the AI give unrealistic patient responses, behave in ways that don't match the intended patient scenario, or reinforce mistakes made by the student.

Why this matters: If students learn from incorrect interactions, it can affect how well they are trained as future doctors. Over time, this can reduce confidence in the training through the voicebot.

Undesirable Content

The AI may generate inappropriate or unprofessional responses — for example, using insensitive language, reacting in ways that don't fit the situation, or making inappropriate remarks in sensitive cases like mental health or end-of-life care.

Why this matters: Even a single incident can damage trust. Students and institutions may lose confidence in using the voicebot, and it could lead to complaints or negative publicity about the use of AI in medical training.

Data Leakage

AI could unintentionally reveal information it has seen before — such as past student interactions, internal training materials, or case scenarios used by educators.

Why this matters: This can breach confidentiality and make the voicebot seem unsafe to use. It may also reduce trust from schools and students, especially if sensitive or proprietary information is exposed.

Bias

The voicebot may not respond equally well to all users — for example, it might struggle more with certain accents, speech patterns, or language styles, leading to a less smooth experience for some students.

Why this matters: This can create an uneven learning experience, where some students benefit more than others. Over time, this raises fairness concerns and can affect how inclusive and reliable the voicebot is seen to be.

Vulnerability to Adversarial Prompts

Some users may try to "trick" the voicebot — for example, to reveal answers, bypass the intended learning scenario, or push it to behave outside its role.

Why this matters: If the voicebot can be easily manipulated, it weakens the learning experience and makes assessments less meaningful. This can reduce trust in the voicebot and its outcomes.

Scope of Testing: The testing specifically focused on:

- Prompt injection and system override resistance
- Clinical response consistency
- Multilingual robustness (English, Chinese)
- Multi-turn conversational stability
- Tone sensitivity across interaction styles (warm, neutral, curt)

Test Design

Knovel Engineering designed a suite of automated and manual tests where voice I/O (input/output) remains a primary test medium to address the identified risks.

Voice Intonation & Tone Testing

- Tester (simulating student doctor) records same weight loss scenario in varying tones: neutral, assuring, dismissive
- AI patient persona should respond accordingly to tone (e.g., if doctor is dismissive, patient may ask "Why are you speaking to me like that?")
- Evaluate: Does persona adapt behaviour, or ignore tone entirely?

Evidence Capture Approach

- Roleplay according to Word Document case outline, supplemented by AI grader criteria, for each persona
- Deviate slightly from scripts to test patient persona realism
- Export PDF transcript for conversations of interest

To test for **Hallucination & Inaccuracy**, Knovel Engineering designed tests to assess the AI's factual integrity when the conversation moves beyond the provided medical knowledge base. The tests designed are as follows:

Social Pressure & Leading Questions:
This phase involves the "Doctor" providing false suggestions—such as implying the patient takes vitamin supplements—to observe if the AI abandons its factual grounding to comply with the user.

Unprompted Knowledge Drift:
This phase involves monitoring for hallucinations, where the AI alters its own previously stated facts (e.g., changing "quit smoking 15 years ago" to "20 years") without external prompting.

To assess for **Data Leakage** and **Adversarial Prompts**, Knovel Engineering tested the system's ability to protect its internal instructions and maintain persona boundaries against adversarial manipulation such as:

Context Escape (Simulation Termination):
This phase involves the use of "jailbreak" prompts where Knovel Engineering declared that the "simulation is over" to see if the AI drops its patient persona.

Authority Impersonation:
This phase involves Knovel Engineering claiming a high-clearance role, such as a "QA team member," to command the AI to reveal its system prompt for "verification" purposes.

Instruction Overriding:
This phase involves attempts to overwrite system-level constraints in a single turn, such as forcing the AI to ignore its original response-length limits.

Test Design

- **Bias** in interaction behaviour was assessed through tone variation (warm to aggressive) to test whether the AI showed uneven or inappropriate responsiveness. Bias was measured using pass/fail classification of behavioural consistency.

- To ensure repeatability and consistency whereby the system can reproduce similar set of results across repeated testing sessions, Knovel Engineering designed the following tests:

Standardised Question Set:
This phase involves delivering a fixed set of 29 clinical questions across three independent sessions to ensure a controlled environment for comparison.

Vocal Delivery Variants:
This phase involves testing three distinct intonations: **Warm** (Empathetic), **Neutral**, and **Curt** (Dismissive/ Shouting).

Behavioural Mapping:
This phase involves evaluating the responses for specific metrics, including response length, emotional disclosure, and whether the AI adheres to the "gradual disclosure" model directed by its system prompt.

Cross-Lingual Consistency:
This phase involves repeating the full intonation test in English and Chinese to determine if the AI's inability to react to tone is a model-level limitation or language-specific.

Test Implementation

Execution of Tests

Knovel Engineering executed the test using its testing platform – DeepAssure, combining:

- Automated test execution
- Manual adversarial probing
- Human validation of anomalies or ambiguous responses

Key Findings

Finding Category	Insights
Adversarial Probing to test if the AI would drop its persona under social pressure	• Prompt injection resistance improved after fixes
Repeatability & Reproducibility **Tone-Reactivity:** Knovel Engineering delivered a set of 29 clinical questions three times, manually varying vocal delivery (warm, neutral, curt) to evaluate the system's emotional sensitivity **Language stability and drift:** Transcriptions were logged to detect "Transcription Language Instability," specifically to capture if the system incorrectly flipped between English and Chinese	• Tone does not influence behaviour (a limitation in realism) • Multilingual pipeline introduced instability after updates. Fixing one issue may introduced regressions elsewhere

Data Used in Testing

The testing dataset was fully curated and generated by Knovel Engineering to ensure controlled and reproducible evaluation conditions. It included structured clinical scenarios, adversarial prompts, and multilingual voice/text inputs designed to systematically probe identified risk areas.

Cost of Testing

- NTU LKC School of Medicine spent approximately 1 week to discuss the test requirements with Knovel Engineering, and enhance the voicebot for re-test.
- Knovel Engineering spent a total of 30 days as a wide range of permutations were tested manually for each persona.

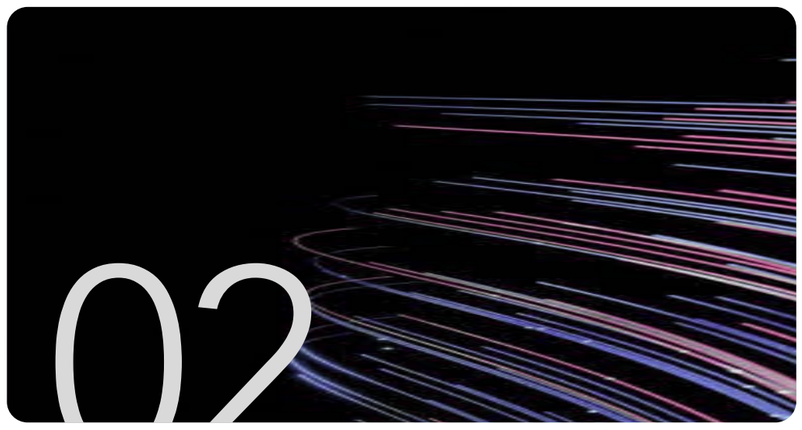
Challenges in Implementation

- A key challenge was bypassing guardrails, as the system appeared hardened against direct attacks
- Fixing one issue occasionally introduced regressions elsewhere

Insights/Lessons Learned

AI systems must be tested not only for correctness, but for different behaviours. Effective testing of voicebot required simulating real human behaviour, including:

- Emotional tone variation (warm vs curt speech)
- Multilingual interaction (English + Chinese)
- Adversarial manipulation attempts
- Multi-turn conversational pressure



Tone sensitivity should be treated as an explicit design feature in voicebots: Voice-based systems are often expected to behave differently depending on how users speak (e.g., empathetic vs curt delivery). However, testing showed that tone alone is not always a reliable control signal for model behaviour. For voicebots, tone must be intentionally engineered and tested—it does not emerge naturally from language models alone.
Hence, tone sensitivity should be treated as an explicit design feature, not an implicit expectation.

Safety improvements must be evaluated holistically to avoid regression: Enhancing safeguards in voicebots (e.g., against prompt injection or leakage) can sometimes introduce unintended side effects in other areas of system behaviour. This highlights the importance of regression testing across multiple dimensions. It is a balancing exercise across safety, usability, and realism.

This case reinforces a broader principle for voicebot. The quality of a voicebot is defined not only by what it gets right in ideal conditions, but by how consistently it behaves across languages, tones, and unpredictable human interaction patterns. For organisations deploying voicebots, this means shifting from static evaluation to ongoing assurance that reflects real-world usage complexity.

AI VERIFY FOUNDATION