

CASE STUDY

When Omnichannel Banking Turns One Complaint Into Many

Sector: Banking / Customer Grievance Redressal / Agentic AI /
Regulatory Technology

73%

Faster Resolution

80%+

SLA Breach Reduction

CASE BACKGROUND

The more times a customer follows up on a banking complaint, the more work they can unintentionally create for the bank.

A failed payment is reported through the mobile app. With no resolution yet, the customer sends an email. Still concerned, they follow up on WhatsApp.

From the customer's perspective, nothing has changed. There is one failed transaction and one unresolved complaint.

Inside the bank, the picture can be very different.

Each interaction may enter the grievance system as a separate ticket. Different agents can begin investigating the same transaction. Different responses can be prepared. And each ticket can start carrying its own resolution timeline.

The customer has contacted the bank three times because the problem remains unresolved. The bank may now be processing the same problem three times because the customer contacted it three times.

The bank has more work without having received a new problem.

Across the banking system, the scale of grievance management makes that duplication significant. More than 1 crore complaints were received by India's 95 commercial banks in FY24, according to the figures referenced in the project research. Of these, 9.34 lakh complaints were escalated to the RBI Ombudsman, a 33% year-on-year increase. The regulatory consequences extend beyond resolution timelines: under RB-IOS, banks whose maintainable complaints exceed their peer-group average can also be charged for the RBI's cost of handling them.

The operational burden grows further when many customers are reporting the same underlying problem.

A UPI disruption across a city, for instance, can trigger hundreds or thousands of individual complaints within a short period. Each represents a customer who needs an answer, while collectively they may be signalling one systemic incident.

This was the operating problem behind Samadhan, an accelerator we developed for retail banking grievance redressal.

The idea was to give the bank a way to understand a complaint before processing it: whether it is new, whether the same customer has already reported it elsewhere, whether it relates to another issue, whether a surge of complaints points to a larger incident, what regulatory requirements apply, and how much of the subsequent investigation and response can safely be handled by AI.

THE PROBLEM

That required us to connect two problems that are often treated separately: unifying a fragmented complaint journey and automating enough of the resolution process to make it faster without compromising regulatory control.

One Customer, One Issue, One Regulatory Clock

The first engineering decision was to separate the channel through which a complaint arrives from the complaint itself.

Samadhan uses an adapter-based ingestion architecture. Complaints can enter through **WhatsApp, email, Telegram, social channels, the RBI CMS or branch interactions** and are normalised into a common complaint format before entering the processing pipeline.

This also makes the channel layer extensible. Additional sources such as IVR, SMS, internet banking, ATM interactions or another social platform can be introduced through a new adapter against the existing contract, without redesigning the identity resolution, deduplication, triage or dispatch logic underneath it.



THE APPROACH

Once a complaint enters the system, Samadhan resolves the customer identity and determines how that interaction relates to existing open cases.

An explicit ticket reference can connect it directly to an existing complaint. The system can also recognise an existing channel thread or compare the intent of the incoming message with open tickets for the same customer.

High-confidence intent matches can be merged. Intermediate matches can become related or child tickets. A genuinely different grievance creates a new ticket.

These decisions remain traceable. Every merge creates an audit event identifying the rule that triggered it, while specified low-confidence merges involving an assigned ticket require supervisory approval.

The operating unit therefore becomes the underlying customer issue, regardless of how many times or where the customer has communicated it.

In the project baseline, a complaint raised through three channels could create three tickets and three SLA clocks. Samadhan consolidates those interactions into one canonical ticket carrying one RB-IO clock.

For the bank, this reduces repeated investigation and gives agents one history of the problem. For the customer, every follow-up adds context to the existing grievance instead of beginning another disconnected resolution journey.

When Hundreds of Tickets Are Actually One Incident

Customer-level deduplication solves only one form of repetition.

Consider a UPI disruption affecting a particular geography. Hundreds or thousands of customers may report failed transactions around the same time.

These complaints belong to different customers and need their own case records. Yet investigating every one independently would overlook the pattern emerging across them.

Samadhan uses live aggregation across complaint category, geography and time to identify incident clusters.

Once that relationship is established, the common resolution narrative can be handled at the parent incident level while individual tickets remain intact. Each affected customer still receives a personalised response containing their own name, transaction reference, channel and language.

The bank can therefore see an operational incident developing through its grievance traffic while continuing to manage the regulatory and customer-specific requirements attached to every complaint.

This also changes the economics of investigation. One underlying failure does not have to trigger hundreds of teams or agents independently discovering the same root cause before the organisation recognises what is happening.

The Model Proposes. The Rules Decide.

Understanding the complaint is one part of grievance resolution. Deciding what AI should be permitted to do with it carries a different level of responsibility.

A balance enquiry, a UPI dispute and a suspected fraud case cannot pass through the same autonomous workflow.

Samadhan therefore assigns complaints to resolution classes that define the level of automation permitted.

Informational cases such as balance enquiries, IFSC lookups or branch hours can be processed directly when the necessary safety conditions are satisfied.

Operational cases such as UPI disputes, card blocks or NEFT traces can be investigated and an action requested, with execution remaining separately authorised.

Sensitive matters including fraud, KYC, deceased-account cases and court orders stay with human specialists. AI can investigate the case and prepare the draft, but it cannot send the final response autonomously.

The same separation exists inside triage.

The LLM reads the unstructured complaint and proposes signals such as intent, sentiment, language and severity. A deterministic rule engine then establishes the final resolution class, severity and regulatory treatment using objective inputs such as the disputed amount, fraud indicators and regulator flags.

That distinction is important in a regulated workflow. The model contributes its ability to interpret unstructured language, while the decision governing what happens next remains bound to explicit rules.

The agent itself also operates within defined limits. It can retrieve customer information, check transaction details and retrieve relevant policy material, but its tools are read-only and its run is capped at four tool calls. Facts used in the investigation come from the underlying systems and retrieved sources, rather than the model's memory.

And one boundary remains fixed throughout the system: the AI never moves money.

A Banking Reply Needs Evidence Behind It

Once the investigation is complete, Samadhan still needs to **establish what supports the answer it is preparing to give the customer.**

RBI master circulars and internal SOPs are embedded into a retrieval layer. When a complaint is processed, the system retrieves the passages relevant to that particular issue along with paragraph references and relevance scores.

The drafted response can therefore cite the regulatory material used to formulate it.

This grounding continues through the agentic investigation. Tool calls and their results are persisted in an agent trace, giving the bank a record of what information the agent retrieved and how the case progressed.

Even a grounded draft does not automatically reach the customer. Samadhan uses a four-condition safety gate before unattended dispatch. The case must belong to a category permitted for autonomous handling, meet the confidence threshold, contain no hard-block condition and satisfy the required assurance state.

If any condition fails, the complaint moves into the human queue. The reviewer receives the prepared response together with the reason automatic dispatch was stopped.

The bank therefore gets automation where the system has enough evidence and authority to proceed, while sensitive or uncertain cases arrive with a human who can see both the draft and the reasoning trail behind it.

Keeping Language Attached to the Complaint

The scale of Indian banking introduces another practical challenge: customers do not necessarily complain in the language in which a bank's internal grievance operation works.

The project context considers a customer population spanning 22 scheduled languages.

Samadhan treats language as a property of the individual complaint.

If a customer writes in Tamil for one interaction and English for another, the system can determine the appropriate response language from each message. The stored customer preference serves as a fallback.

The human-review experience was designed around the same requirement.

For cases that need an agent, the pre-translation English version is retained alongside the customer-language response. A reviewer who does not read the customer's language can still verify the substance before approving it. Citations and the critic verdict remain connected to the English source so that the review can later be reconstructed.

This allows multilingual communication to remain part of the same governed complaint process instead of becoming an isolated translation step at the end.

The Regulatory Record Builds Alongside the Complaint

Complaint resolution in banking can continue beyond the customer response.

Cases entering the Ombudsman process may require an Annexure-A containing complaint details, chronology, investigation findings and other information required for regulatory submission.

Preparing that document manually means bringing together information generated across the life of the complaint. **The project baseline estimates that this compliance drafting consumes approximately 20 person-days every month.**

Samadhan generates the Annexure-A directly from the resolved ticket state in one action, producing a hashed PDF for regulatory submission.

That capability depends on maintaining a trustworthy history throughout the case.

Every significant action enters an append-only, SHA-256-chained audit trail. The database itself rejects updates or deletions to that history. Privileged workflows support four-eyes control, with maker and checker required to be different people. SLA calculations account for business hours and pauses, while the audit trail records the progression of the complaint through resolution.

The compliance document can consequently be produced from information that has already been captured and governed during the grievance lifecycle.

Why the Architecture Could Not Wait for the AI

The customer may experience Samadhan as one complaint journey. Underneath it, different parts of that journey have very different processing characteristics.

A customer-facing channel expects an acknowledgement quickly. Identity resolution and ticket processing are relatively deterministic. An AI agent may need more time to retrieve evidence, call tools and formulate its findings.

Connecting all of these services synchronously would create an architectural weakness: a slow LLM call could hold up the entire chain and eventually block the customer-facing webhook.

We therefore built Samadhan around an event-driven architecture using Apache Kafka.

Each stage publishes an event and consumes the events it needs instead of depending on a chain of direct service calls. A slow AI operation can continue processing without holding the customer interaction open. Kafka can redeliver events when a consumer fails, preserve the processing history for replay and audit, and allow new capabilities to subscribe to existing events without requiring upstream services to be redesigned.

The control plane and ingestion services are built in Go, while the AI plane uses Python for functions including deduplication, triage, embeddings, RAG, agentic investigation and drafting.

PostgreSQL serves as the system of record, Qdrant stores regulatory content and intent vectors, Redis supports caching, and MinIO handles objects such as attachments and generated Annexure-A PDFs. Incoming attachments are also scanned for malicious content before they proceed through the grievance workflow.

The architecture is containerised and designed for deployment within the bank's network boundary. The model layer sits behind a provider abstraction, allowing the development model to be replaced by an on-premise Indic model in production without changing application call sites. In production, the same images are intended to run on Kubernetes inside the bank's environment, keeping the pipeline, datastores and model endpoint within that controlled boundary.

For a banking AI system working with grievance and transaction context, this architecture gives the agent access to the information required for investigation while preserving control over where that information travels.

THE OUTCOME

What Changed in Practice

The impact becomes visible when the entire complaint journey is measured.

In the project evaluation, average complaint turnaround moved from approximately 45 minutes to 12 minutes, a reduction of roughly 73%.

A customer issue raised through three channels could **move from three tickets to one**, carrying one regulatory clock instead of three.

Annexure-A preparation, estimated at approximately 20 person-days per month, could be generated from the case state in one action.

And through consolidated SLA tracking and automated escalation, the design **targets an 80%+ reduction in RB-IOs SLA breaches**.

Those outcomes are particularly relevant against the wider grievance environment in which banks are operating. Complaint volumes are high, Ombudsman escalations are increasing, and the regulatory framework increasingly connects grievance performance with financial and supervisory consequences.

Samadhan was our attempt to address that problem across the full operating chain.

It brings fragmented grievance channels into a common layer, recognises duplicate complaints and wider incidents, uses agentic AI to investigate cases within explicit boundaries, grounds responses in regulatory evidence, supports multilingual communication, maintains a tamper-evident history and turns that history into the documentation needed for regulatory closure.

The result is an accelerator designed around a deceptively simple idea: a bank should spend its time resolving customer problems, not repeatedly rediscovering the same one.



That is what this project really solved.

A centralised, governed, searchable system – built to assist rather than obstruct.

www.joshsoftware.com | contact@joshsoftware.com | +91 7887889902