www.sevaka.ai

WHITEPAPER

Eliminating AI Hallucinations in Financial Services

How platform-led architecture mitigates hallucination

Sevaka

CONTENTS

Introduction

Artificial intelligence is rapidly entering financial services, promising efficiency and better member experiences. But in regulated environments, where accuracy, consistency, and auditability are essential, concerns about AI hallucinations remain high.

These concerns are valid. Language models generate responses probabilistically, which can introduce inconsistency if not properly governed.

This paper explains why hallucinations occur, where traditional approaches fall short, and how Sevaka's platform-led architecture enables reliable, policy-aligned AI suited to regulated financial services.

Contributors

This whitepaper is shaped by the Sevaka team, who work every day with retirement savings providers, advisers, and industry leaders. The contributors below bring experience across product, advice, technology, and AI, and have helped turn real industry challenges into practical guidance for this series.

Clive Fernandes - CEO

As founder of National Capital, New Zealand's largest KiwiSaver advice fintech, Clive has direct experience with the challenges KiwiSaver providers and advisers face.

Matthew Hare - CTO

As the former CTO of Instillery, an Enterprise DevOps company, Matthew led the organisation to win the prestigious IDC Digital Transformer of the Year award during his tenure. He ensures Sevaka is enterprise-ready.

Peter Hwang - HEAD OF AI

Pete has a track record of delivering real-world impact. He led the development of a flagship product through to acquisition, built a stock optimisation tool that saved over $1 million, and has co-authored research on fine-tuned LLMs.

Arran Cunningham - DESIGN, UI/UX

With over 25 years of experience in branding, design and UI/UX, Arran was part of National Capital's founding team as well as the lead designer/developer of mortgages.co.nz.

Mika Kato - OPERATIONS & STRATEGY

With a background in government, consulting, and digital reform, Mika specialises in improving systems and service delivery. She leads the operational delivery and research underpinning Sevaka's AI in Wealth Management Initiative.

ONE

What hallucinations are, and why they are a problem

In financial services, a confident but incorrect response carries consequences.

Hallucinations can lead to:

- references to policy clauses that do not exist
- omission of eligibility conditions
- different answers to similar member questions
- decisions that cannot be clearly explained to regulators.

Accuracy must be consistent, defensible, and repeatable. Systems that generate plausible language without controlled decision logic introduce avoidable risk.

In the context of AI, a hallucination occurs when a system produces information that appears coherent and confident, but is incorrect, incomplete, or made up.

Large Language Models (LLMs), such as ChatGPT, generate responses by predicting likely sequences of words based on patterns in data. They don't verify facts or apply rules in the way humans do. As a result, they can produce outputs that read as authoritative while still being wrong.

In financial services, hallucinations can appear in subtle ways, like:

- referencing policy clauses that do not exist
- omitting critical eligibility conditions
- combining rules from different contexts
- presenting estimates or interpretations as definitive guidance.

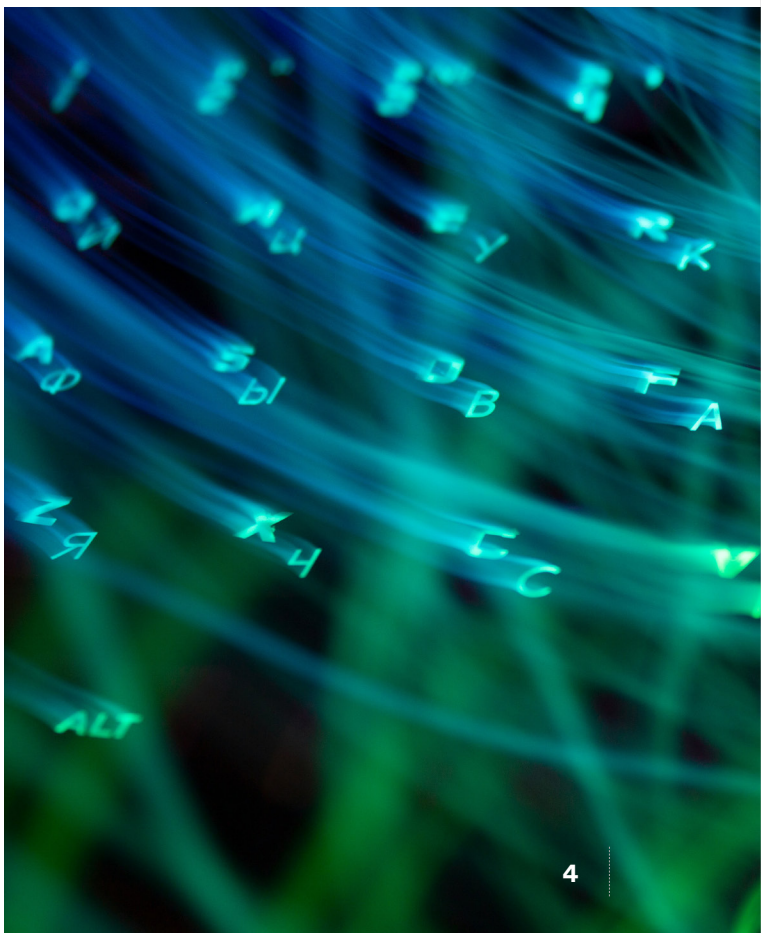
These errors can be hard to detect because the output is often well written and confident. At scale, they undermine trust in automated systems and increase operational and compliance risk.

Hallucinations are not limited to obvious factual errors. They also occur when AI systems attempt to reason through complex policy logic or make decisions that require strict rule adherence. In these cases, the problem can be more than just incorrect information, it can also be incorrect outcomes.

Addressing hallucinations doesn't just require better prompts or larger models, it requires a deliberate approach to how AI is used within regulated workflows, and clarity about where responsibility for decisions sits.

In regulated environments, accuracy is non-negotiable. A response that sounds reasonable but is factually wrong can expose organisations to compliance breaches, poor member outcomes, and loss of trust.

Members rely on information to make decisions about their money, and regulators expect firms to be able to explain how outputs are produced. As AI adoption increases, understanding how hallucinations arise and how they can be mitigated becomes essential.



TWO

Large Language Models vs platforms: an important distinction

To understand why hallucinations occur even when AI systems appear well designed, we must distinguish between language models and the platforms that govern how they are used.

An **AI model**, such as an LLM, is trained on large datasets to recognise patterns in language. Through training, the language model learns how words, phrases, and concepts tend to appear together. This allows it to interpret language and generate text, but always in a probabilistic way.

A **platform** governs how that language model is used. It defines what the language model is allowed to see, what tasks it is allowed to perform, and how its outputs are handled.

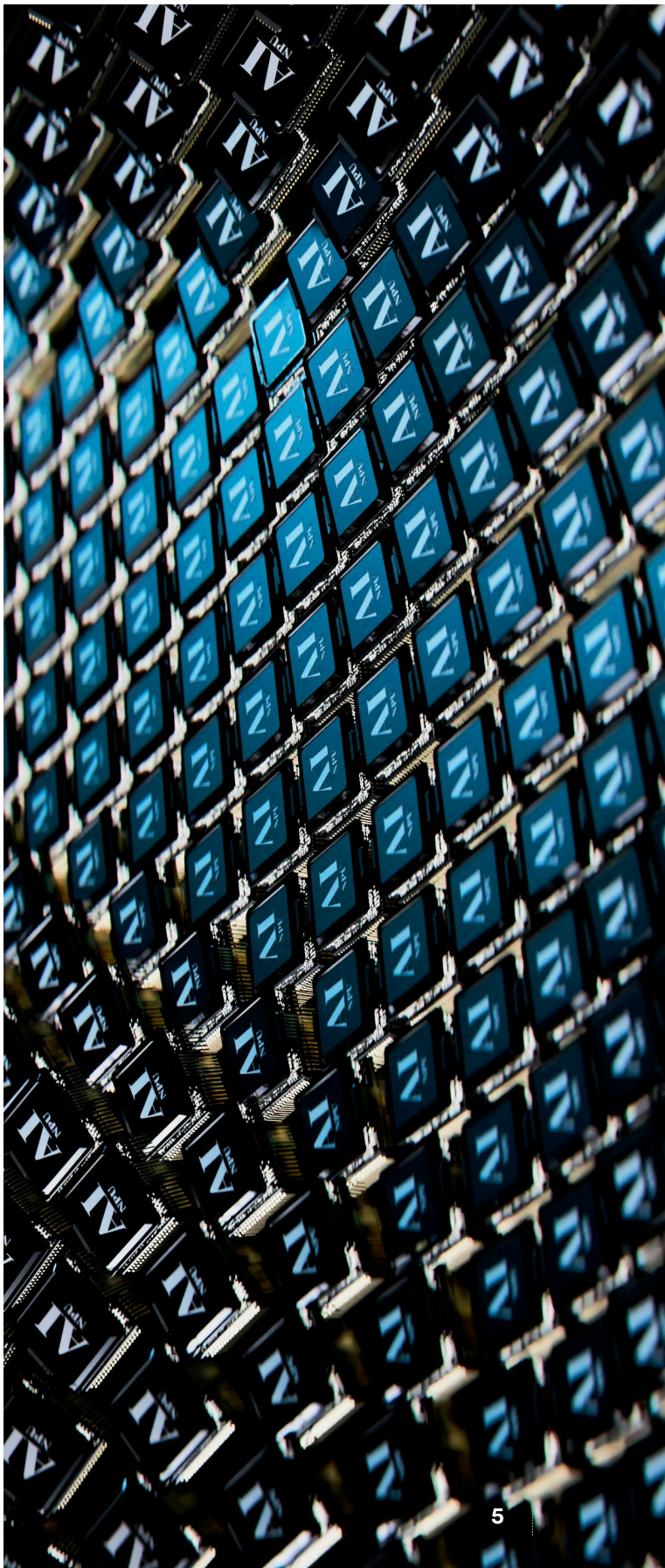
In simple terms:

- the **language model** learns how to recognise language patterns
- the **platform** defines what those patterns are allowed to mean and what actions can follow.

A platform can control:

- what information the language model is allowed to see
- what tasks it is allowed to perform
- how outputs are structured
- how decisions are logged, tested, and audited
- how and when humans intervene.

In regulated contexts, this distinction matters because the risk does not sit in recognising language. It sits in interpreting policy, applying rules, and producing outcomes that must be consistent, explainable, and defensible.



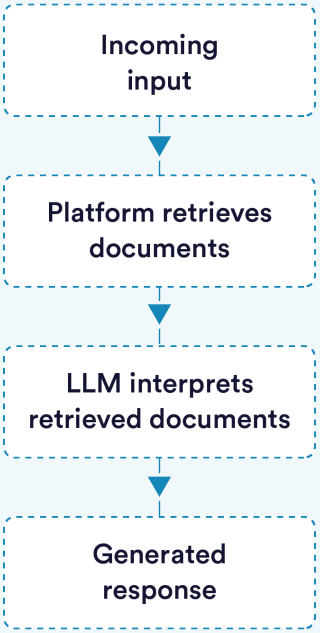
THREE

How most AI implementations try to reduce hallucinations: Retrieval-Augmented Generation

Most AI implementations follow a Retrieval-Augmented Generation (RAG) pattern:

Language Model-led Generation

(Traditional RAG)



Even with relevant document retrieval in place:

- the language model determines how retrieved content is interpreted
- the language model determines how rules are applied
- the language model determines how responses are composed.

A common approach to reducing hallucinations is retrieval-augmented generation (RAG).

In a RAG setup, the platform retrieves relevant documents at the moment a user asks a question. These documents are typically drawn from internal sources such as policy documents, manuals, or knowledge bases.

The retrieved material is then provided to the language model, which is prompted to generate a response using that information, instead of relying only on what it learned during its original training.

The logic is: if the language model can reference relevant, up-to-date documents before responding, it should be less likely to fabricate information.

In practice, a traditional RAG flow looks like this:

- a user asks a question
- the platform searches internal documents for relevant text
- selected snippets are passed to the language model
- the language model generates a natural-language response based on those snippets.

For many knowledge-based use cases, this approach works reasonably well. RAG can improve factual accuracy and reduce reliance on outdated training data, which is why it has become a common pattern in enterprise AI tools.

In regulated environments, however, RAG alone introduces limitations that are often underestimated.

FOUR

Why traditional RAG breaks down in financial services

Language models optimise for likely language patterns rather than policy certainty.

Because responses are generated dynamically:

- two similar requests can produce different answers
- subtle interpretation differences can change outcomes
- model upgrades can alter behaviour.

This architecture places decision control inside a system designed to optimise for likely language patterns rather than policy certainty.

The type of results this produces might be acceptable for search or drafting support, but it introduces significant risk in regulated workflows.

RAG improves what a language model can reference, but it doesn't guarantee correct outcomes. This becomes a problem in financial services for several reasons.

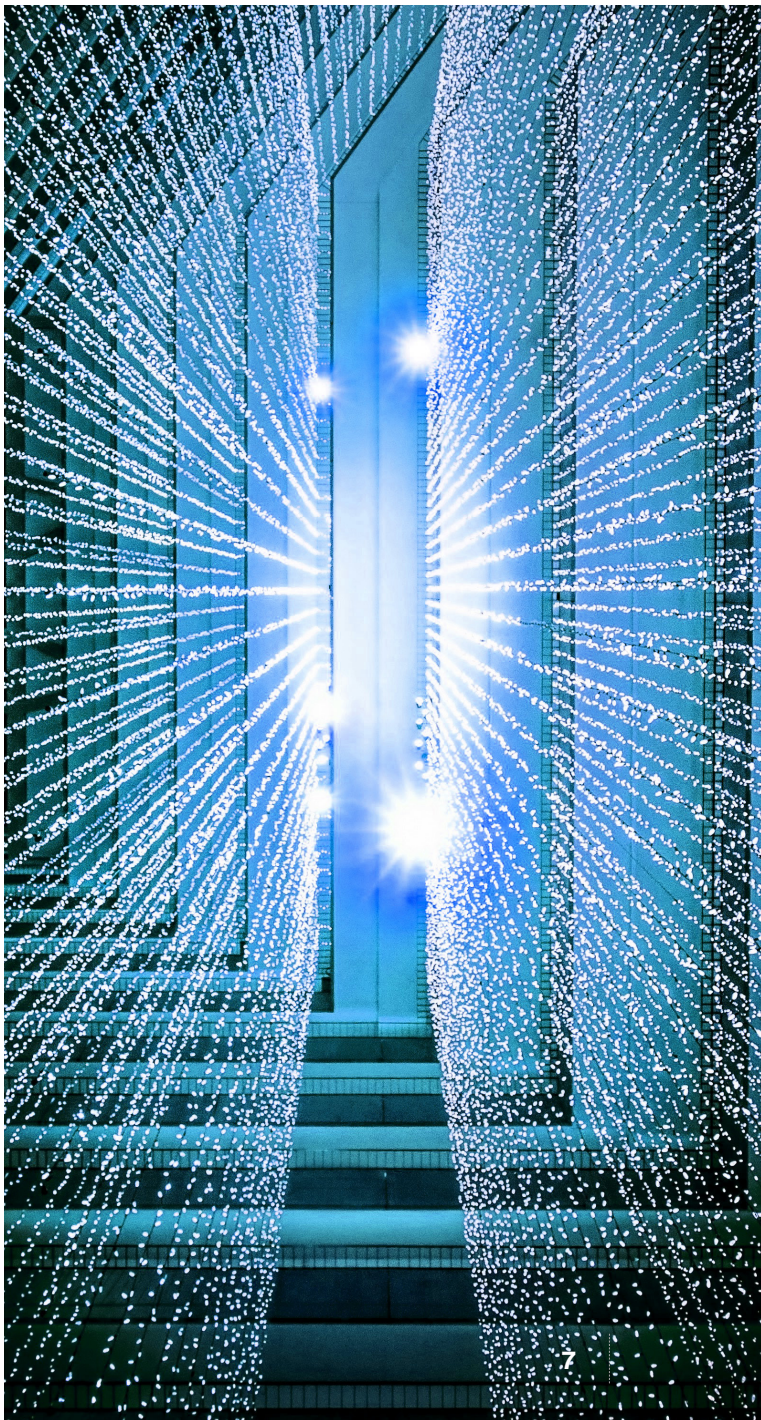
First, retrieval does not guarantee correct interpretation. Even when the right policy text is retrieved, a language model might misinterpret it, overlook exceptions, combine clauses incorrectly, or apply logic inconsistently. The response may sound confident while still being wrong.

Second, many financial services interactions require predictability. These interactions are about determining eligibility, required actions, or next steps. The decisions are governed by rules and thresholds, not written responses. Asking a language model to reason through these decisions without clear rules introduces unnecessary uncertainty.

Third, verification becomes complex. When something goes wrong, teams must determine whether the issue arose from retrieval, interpretation, generation, or prompting. This complicates auditability and regulatory explanation.

Finally, language model behaviour changes over time. As underlying language models evolve (e.g. from ChatGPT 4.0 to ChatGPT 5.0), their interpretation of retrieved information can drift. In RAG-only setups, this can lead to subtle changes in outcomes that are hard to detect and difficult to justify to regulators.

RAG is useful for answering questions, but it is far less reliable as the foundation for regulated decisions.

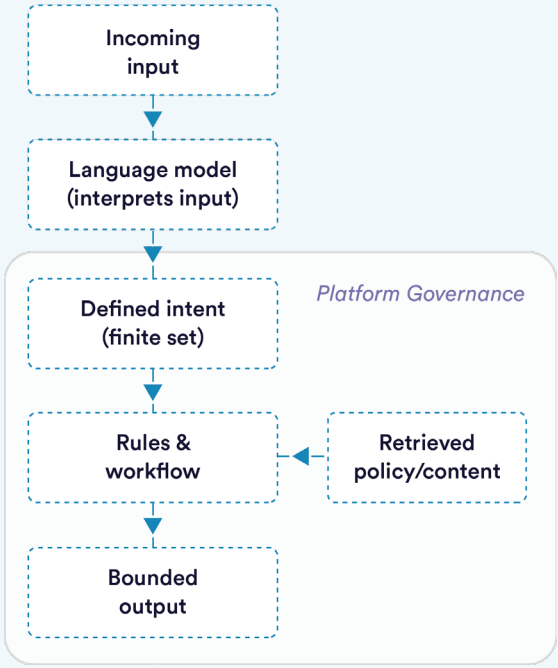


FIVE

Sevaka's approach to reducing hallucination risk

Sevaka's AI Financial Agent (AIFA) changes how the AI in language models is used by giving it structure and rules to work within.

Platform Led Governance



Retrieval of documents is governed by the workflow. Policy and content are retrieved to support rule execution, not to generate responses freely.

Sevaka takes a different approach by redesigning how AI is used within regulated financial services workflows.

Instead of relying on a language model to infer how financial services workflows should operate, the Sevaka team explicitly defines them within its platform.

The AIFA platform encodes financial services expertise directly into intent definitions, workflow logic, and decision rules. This determines what the system can interpret, which actions are permitted, and which outcomes are valid in a regulated context.

Sevaka's AI platform is called AI Financial Agent (**AIFA**). It sits above the underlying language model and governs how the model is used within financial services workflows.

Within the **AIFA context**, the language model is used primarily for language understanding. AIFA's role is to then interpret how members express their needs and classify those messages into defined intents. Decision-making is then also handled by the AIFA platform, through governed workflows that reflect financial services policy, regulation, and operational reality.

This approach is built on three principles:

1. Intent first

> The proprietary Intent matching framework is central to Sevaka's AIFA platform and a core part of its intellectual property.

The platform is configured with a finite set of valid member intents, rather than allowing open-ended interpretation. These intents reflect real financial services use cases and are defined by Sevaka's financial services specialists, and form the gateway into governed workflows.

The language model is used to interpret how a member expresses themselves. The platform then classifies the message into one of the predefined intents. The model does not create new intents or decide what actions should occur - all intent recognition occurs within the boundaries set by the platform.

2. Rules over reasoning

Once intent is identified, predefined workflows within the platform determine what happens next. These workflows have been designed and maintained by the Sevaka team to reflect the relevant policies, eligibility criteria, and operational constraints of financial services.

The underlying language model follows clearly specified pathways and does not reason about policy or make judgement calls.

3. Bounded outputs

Responses are produced through controlled structures aligned to each workflow in the platform. This limits open-ended generation and ensures outcomes remain consistent, explainable, and aligned with policy.

Because outcomes are governed by the AIFA platform and the domain expertise encoded within it, regulated behaviour remains stable even as underlying language models evolve. Improvements in language models enhance language understanding without changing decision logic.

By separating understanding from decision-making and building the system around clear, domain-specific rules, the platform reduces hallucination risk at the source. It steps in at the intent stage, stopping small misunderstandings in language from turning into bigger downstream errors.

Architectural Difference

Traditional RAG:

- Language model-led interpretation
- Language model-led rule application
- Language model-led response composition

Sevaka AIFA:

- Platform-defined intent taxonomy
- Policy-driven workflow execution
- Controlled output structures

In the platform led flow, the language model only interprets language, while the platform governs consistent decisions and outcomes.

SIX

Product example: how this works in practice

To make this approach concrete, consider a common client servicing scenario: responding to a member email.

Email response automation

In a traditional AI setup, a language model might be asked to read an email and draft a response directly. Even with RAG, the language model is still responsible for composing the answer and applying policy logic. This is where hallucination risk arises.

When responses are generated dynamically by the LLM at the time of interaction, two members asking the same question can receive different answers. In financial services, these inconsistencies can have big impacts. Small differences in wording or interpretation can lead to different outcomes, confusion for members, and challenges for compliance teams trying to explain decisions.

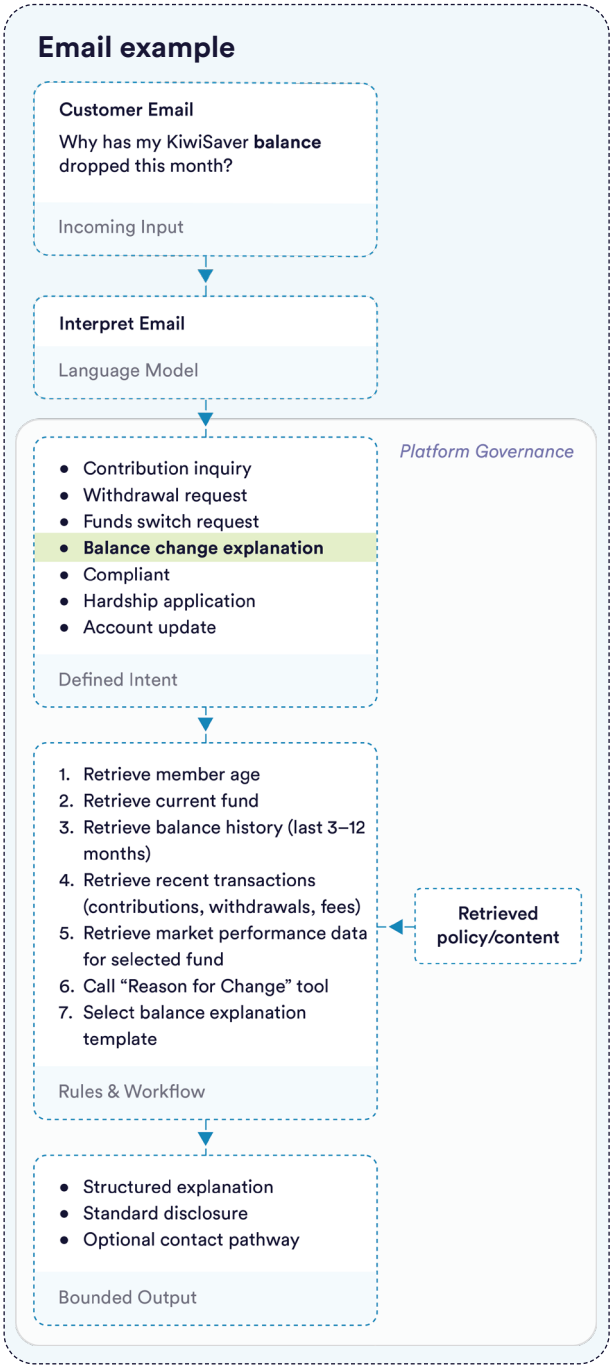
A customised platform like AIFA handles this differently.

When a member email arrives, the platform:

- uses the message itself as the primary input
- uses the language model to interpret what the member is trying to do, such as requesting information, initiating a process, or seeking clarification
- classifies the message into a defined intent category.

Once intent is identified, the platform moves the interaction into a governed workflow:

- predefined rules determine which response pathways are valid
- approved response structures are selected based on policy and operational requirements
- required information or next steps are determined deterministically.



Rather than asking the language model to draft a response, the platform assembles an outcome through controlled logic.

The result is a response that feels natural to the member, while remaining consistent, explainable, and aligned with policy.

SEVEN

Building trust through transparency and auditability

Trust in regulated AI requires four things

In financial services, trust in AI systems depends on:

- consistent treatment of similar scenarios
- clear alignment with policy and regulation
- traceable decision pathways
- stable behaviour, even as underlying models evolve.

Language models optimise for likely language patterns, but regulated institutions require structured, policy-aligned outcomes.

Sevaka's architecture bridges that gap by encoding institutional expertise into the platform and building transparency into every interaction.

How Sevaka embeds trust into the system

Sevaka builds trust by shifting decision control away from probabilistic language generation by language models, and into governed workflows controlled by the platform.

1. Institutional expertise is encoded directly into the platform

AIFA is configured around financial services policy and operational requirements. The platform encodes:

- financial services intent definitions
- policy logic and eligibility criteria
- workflow constraints
- operational compliance requirements
- client-specific rules.

These definitions are designed, reviewed, and maintained by financial services specialists within the Sevaka team, as opposed to generic templates, ensuring the system reflects how regulated institutions actually operate.

2. Transparency is built into every interaction

Transparency and auditability are architectural features.

Each interaction can be traced end-to-end, including:

- original user input
- detected intent
- selected workflow
- rules applied
- response pathway used.

Because decision logic sits outside the language model:

- outcomes can be tested and validated
- behaviour can be monitored over time
- model upgrades do not silently alter regulated outcomes.

This traceability enables institutions to clearly explain outcomes in language aligned with policy, operations, and regulatory expectations.

3. The result: structured, auditable AI

With Sevaka AIFA:

- similar scenarios follow the same governed pathway
- policy logic is applied through controlled workflows
- outputs remain structured and explainable
- governance remains stable even as language models evolve.

AI operates as a controlled operational capability within the institution, instead of as an unbounded response generator.

EIGHT

Eliminating hallucination risk through architecture

In financial services, hallucination risk is controlled through platform architecture, rather than bigger and better models.

Traditional AI implementations place interpretation, rule application, and response generation inside a probabilistic system optimised for language fluency.

Sevaka takes a different approach by:

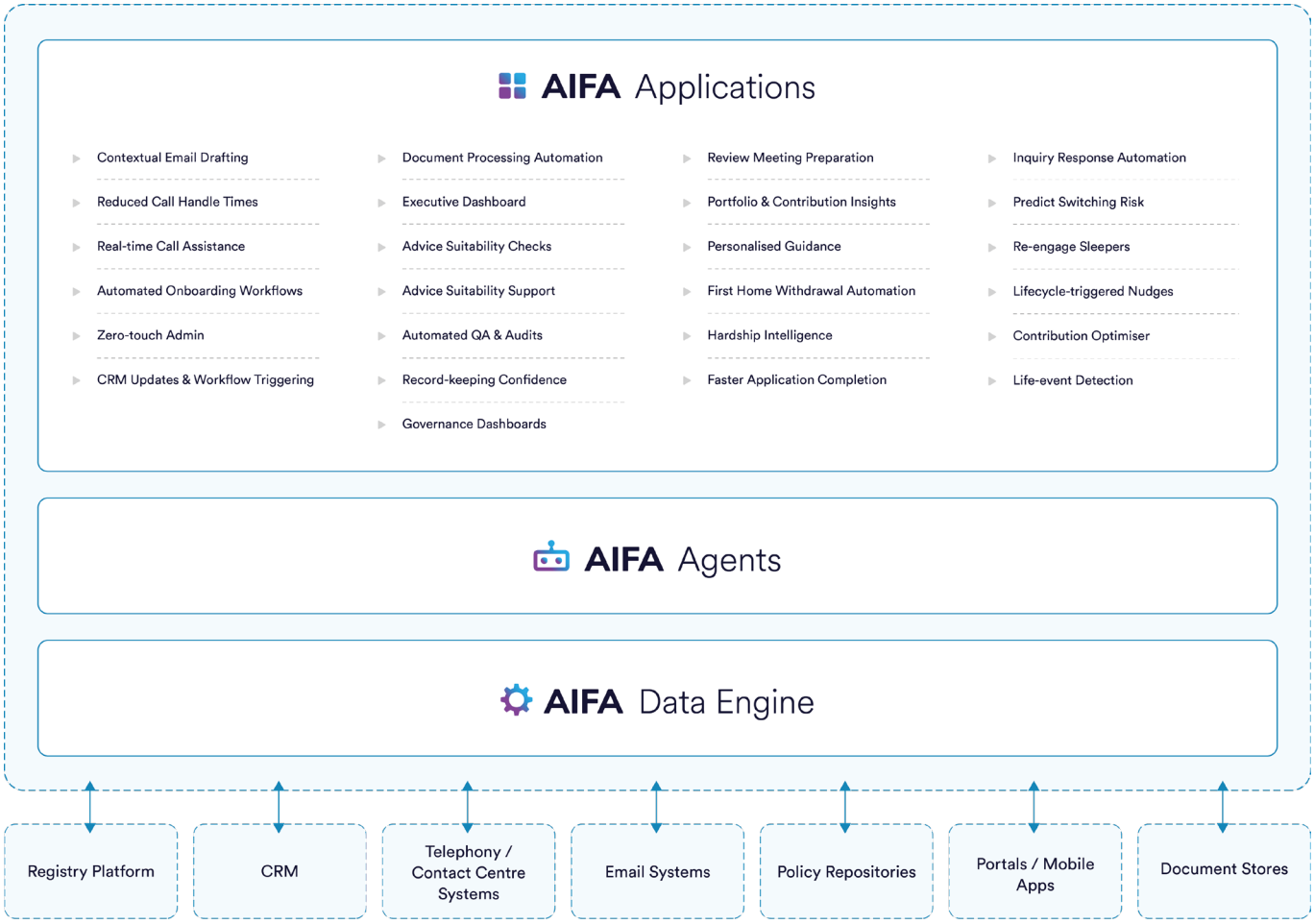
- separating language understanding from decision-making
- classifying interactions into a finite set of defined intents
- governing outcomes through deterministic workflows
- embedding financial services expertise directly into the platform.

Sevaka reduces hallucination risk at its source by placing decision logic inside governed workflows rather than inside the model.

With that structure in place, AI can scale operational capacity without weakening certainty, accountability, or transparency.

In regulated environments, trust depends on consistency, explainability, and control of AI decisions.

Sevaka's AIFA Platform



AIFA is an AI financial agent platform built by Sevaka to help Wealth Management and Retirement Savings Providers automate operations and improve member experience. It reduces pressure on teams, lifts service quality and gives members faster, clearer and more consistent support.

AIFA Data Engine

The AIFA Data Engine connects to the systems providers already use: registry and admin platforms, CRM, telephony, email, portals and policy libraries. It brings this information together so AIFA has the context it needs to act accurately.

AIFA Agents

The AIFA Agents analyse information and take action. They apply policy rules, detect patterns, prepare drafts, generate summaries, predict risks and move workflows forward. This is where the intelligence and automation sit.

AIFA Applications

AIFA Applications are use-case-specific implementations built to slot into existing processes to reduce costs, improve efficiency, and lift member outcomes across servicing, operations, compliance, and advice.

Architecture That Mitigates Hallucination

AIFA reduces hallucinations by placing control around the AI rather than relying solely on the model.

The language model is used to understand what the user is asking. The AIFA applications define what actions are permitted and which policy paths can be followed. The AIFA agents then run structured workflows.

Instead of producing open-ended answers, responses are built within approved structures. That means outputs are consistent, traceable, and aligned with policy.