

**Deloitte.**

**aws**



# Defensive UX for Generative AI

**A framework for responsible,  
auditable AI at scale**

*May 2026*

# Why scaling enterprise GenAI requires transparency, controls and *human accountability*

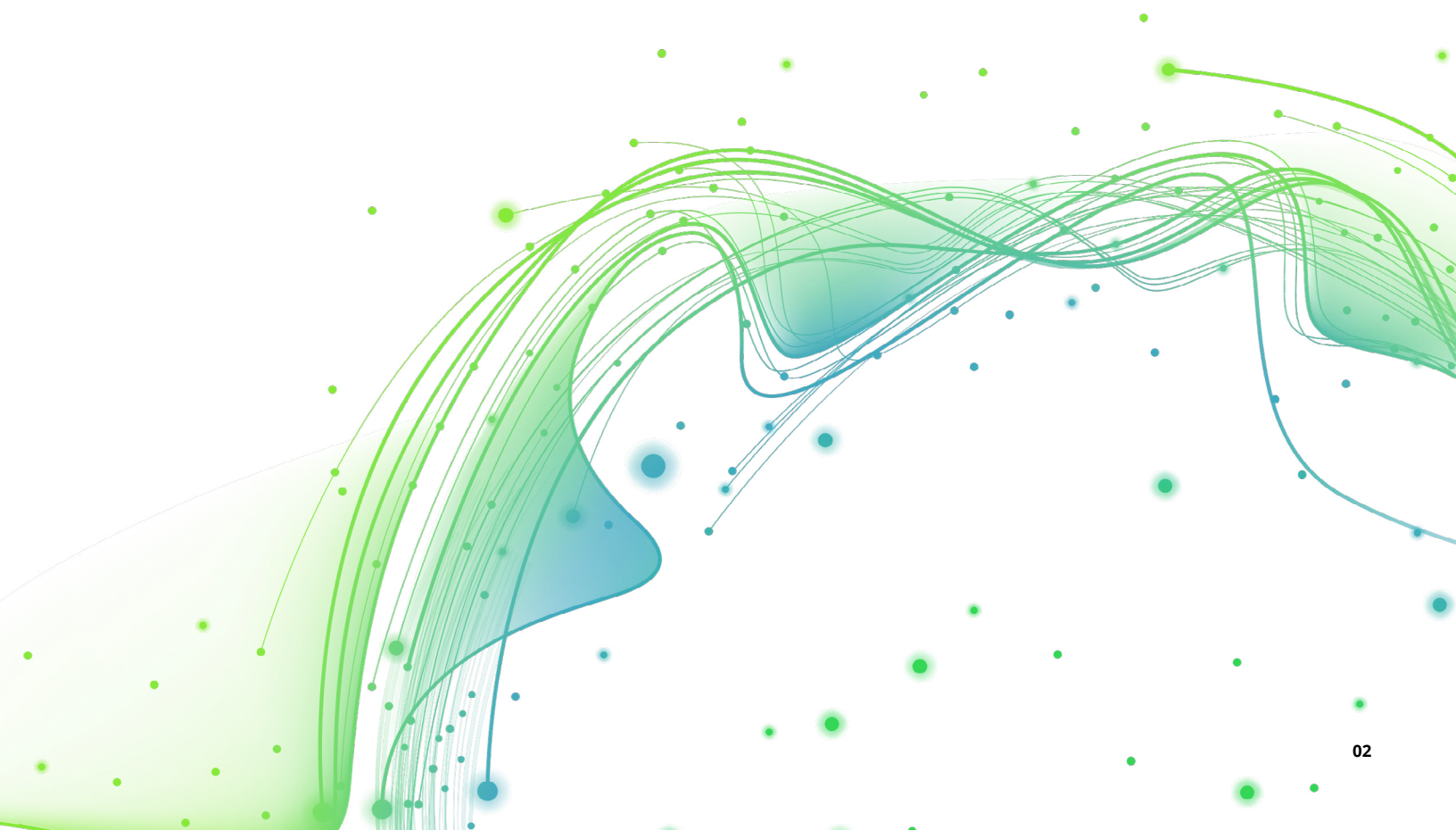
Scaling Generative AI (GenAI) applications in the enterprise, where GenAI governance and AI risk management are essential, demands a fundamentally new approach to human-computer interaction. Contemporary design philosophies such as invisible AI and natural UI prioritize seamless, unobtrusive experiences that make AI disappear into the background—most often at the expense of AI transparency and auditability.

**Defensive UX** takes the opposite approach. It consciously and deliberately makes users aware that responses are AI generated, prompting them to exercise appropriate caution before accepting outputs as ground truth.

This distinction is not academic.

In enterprise setting—especially in regulated industries—the stakes are materially different from consumer applications. Business users complete critical transactions, regulatory filings and client deliverables where errors carry real financial, legal and reputational consequences. The defensive UX framework, with its emphasis on human accountability and prompt governance, is designed precisely for this reality.

The defensive UX framework is a key feature of **10X Analyst**—an agentic AI platform built by Deloitte with Amazon Web Services (AWS) specifically for the Financial Services Industry.



# Defensive UX

Defensive UX is particularly important for B2B and employee-facing applications where **accountability cannot be delegated to technology**. Enterprise users remain accountable for outcomes and must defend their decisions to regulators, senior leaders, auditors and clients. Just as a finance analyst cannot excuse flawed analysis by blaming a junior colleague, they similarly cannot deflect responsibility by claiming “the AI did it.” The buck stops with the human decision-maker, regardless of the tools employed.

Furthermore, organizations must demonstrate to regulators and compliance personnel that adequate measures exist within each application to ensure the employee running the GenAI process retains accountability for results. The defensive UX framework provides the structural foundation for building this case.

## *Human-AI interaction design for GenAI:* From invisible AI to auditable workflows

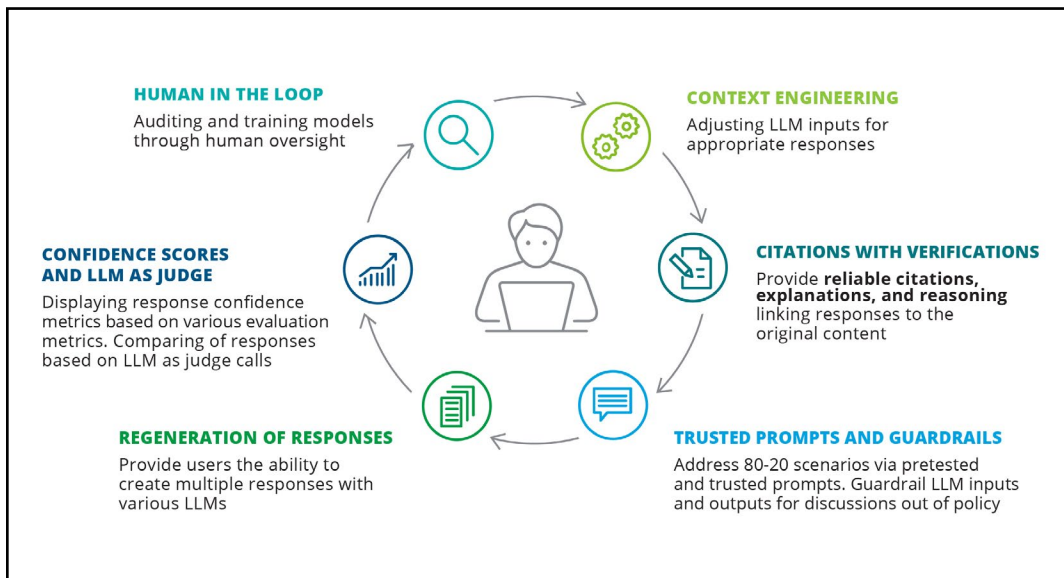
GenAI applications introduce a more nuanced, bidirectional relationship between humans and computers. This interaction operates at a granular level where humans depend on AI to accelerate and enhance their work, while AI systems require human feedback, context and validation to improve accuracy and relevance.

This symbiotic relationship demands a user experience design that acknowledges AI’s capabilities and limitations while holding humans ultimately accountable for final results. Defensive UX provides the design language for this new paradigm.

# The defensive UX framework: Six UX patterns for responsible enterprise GenAI

Based on our work with leading financial services and insurance organizations, Deloitte developed the defensive UX framework, a key feature of the 10X Analyst platform. The framework's comprehensive approach identifies six essential components for building responsible GenAI applications that balance innovation with risk management.

Figure 1. The six components of the defensive UX framework



These principles are embedded in Deloitte's agentic solution 10X Analyst, a reference implementation that brings the defensive UX framework to life through practical application design patterns and user interaction models. *The examples throughout this article are drawn from this platform built for the Financial Services Industry.*

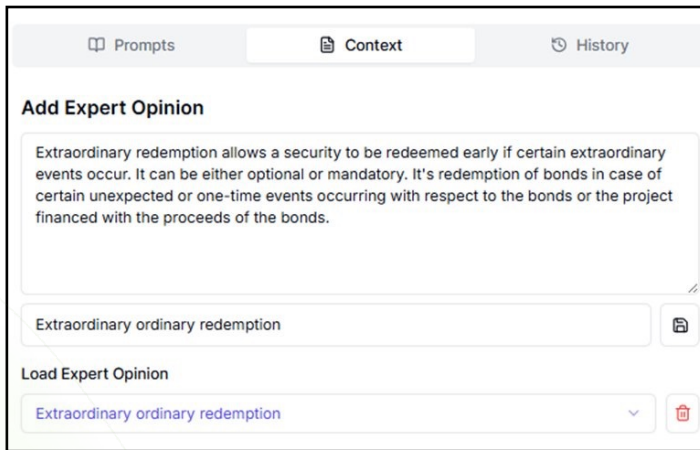
# 1. Context engineering

**Context is the foundation of accurate AI responses.** This component focuses on empowering both users and systems to adjust LLM inputs so the model has the right context to create optimal outputs. Rather than treating any form of AI as a black box, context engineering provides interactive refinement tools that allow users to adjust parameters, add relevant background information, and specify output requirements before submitting the starting prompt.

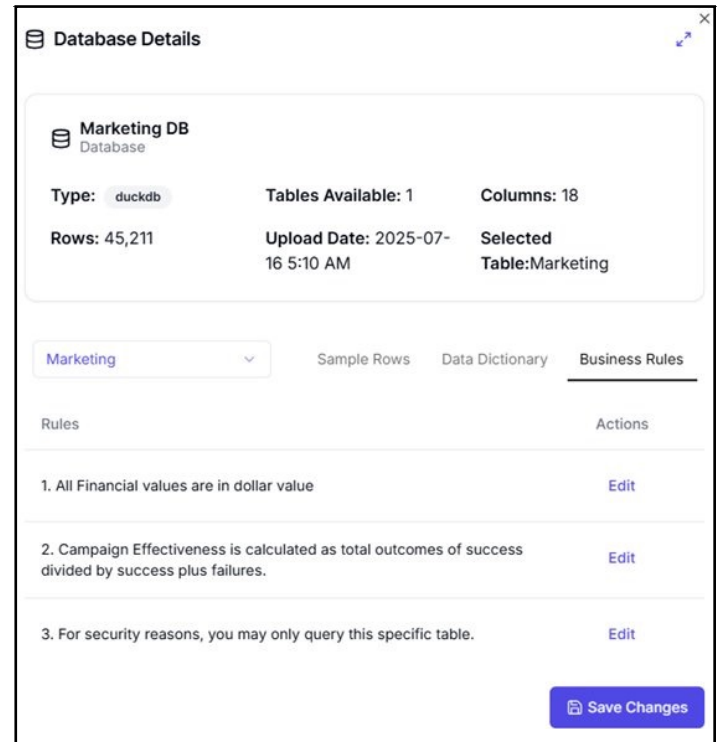
Equally important is intelligent context assembly, where the system automatically incorporates relevant organizational data, user history and domain-specific knowledge. Preprocessing capabilities help users structure their queries effectively, reducing ambiguity and improving response quality.

By making context visible and adjustable, users gain greater control over AI outputs while developing a deeper understanding of how their inputs influence results. In the 10X Analyst implementation, users can add expert opinions and load domain-specific context before submitting queries and can define business rules that constrain how the AI interprets data.

**Figure 2. Adding expert opinion to guide AI context in 10X Analyst**



**Figure 3. Database details with editable business rules that shape AI interpretation**



## 2. Citations with verifications

**Trust requires transparency.** This component ensures every AI-generated response can be traced back to its sources. Reliable citations link specific claims to original documents, datasets or knowledge bases. Clear reasoning chains explain how the AI arrived at its conclusions, making the logic transparent and auditable. Verification tools enable users to quickly validate sources and cross-reference information, while confidence indicators help assess the reliability of different information sources.

This approach transforms the AI from an oracle into a research assistant that shows its work, enabling users to make informed decisions about which outputs to trust. In practice, the system asks the LLM to explain its reasoning and provide references that substantiate each answer.

Figure 4. Source document references with factual accuracy assessment

The screenshot shows a window titled "Explanation" with a close button (X) in the top right corner. It has two tabs: "References" (selected) and "Reasoning".

**Source Documents Reviewed**

- [Redacted] Bond Call\_1.3.2024.pdf
- [Redacted] Rating Change\_9.8.2024.pdf
- [Redacted] Rating Change\_7.31.2024.pdf

**Factual Accuracy: Accurate**

- Evidence: The selected text accurately describes the January 3, 2024 notice of full optional redemption for City of [Redacted], [Redacted] Combination Tax and Surplus Revenue Certificates of Obligation.
- Page Reference: Page 1 of [Redacted] Bond Call\_1.3.2024.pdf
- Exact Quote: "NOTICE OF FULL OPTIONAL REDEMPTION CITY OF [Redacted], [Redacted] COMBINATION TAX AND SURPLUS REVENUE CERTIFICATES OF OBLIGATION, SERIES NOTICE IS HEREBY GIVEN pursuant to the terms of the governing document dated as of 07-May-2015, between [Redacted] [Redacted], as Issuer, and U.S. Bank Trust Company National Association, as Trustee or Agent, that the bonds listed below have been selected for Full Redemption on 15-Feb-2024 (the Redemption Date) at the price listed below of the principal amount (the Redemption Price) together with interest accrued to the Redemption Date."
- Relevance to Question: The original question asked for a chronological summary of bond rating changes for the City of [Redacted]. The selected text correctly identifies the January 3, 2024 document as a notice of redemption rather than a rating change, which is an important distinction when providing an accurate chronological summary of rating changes.
- Analysis: The document dated January 3, 2024 is indeed a notice of full optional redemption for multiple bond series to be redeemed on February 15, 2024. The document does not contain any information about rating changes - it simply announces that certain bonds will be redeemed. The selected text accurately states this fact and correctly notes that this document "did not include a rating change." This is important context for the chronological summary of rating changes, as it explains why this document, despite being the earliest in the timeline, does not represent the first rating change. The actual rating changes occurred later, with the first being Fitch's downgrade from 'AA+' to 'AA' on July 29, 2024 (documented in [Redacted] Rating Change\_7.31.2024.pdf), followed by an upgrade back to 'AA+' on September 5, 2024 (documented in [Redacted] Rating Change\_9.8.2024.pdf).

Figure 5. Step-by-step reasoning chain showing the AI's logic

**Explanation**

References Reasoning

### My Step-by-Step Logic for Including "January 3, 2024 Notice of Full Optional Redemption"

I approached the question about bond rating changes using the following logical process:

- Step One: Understanding the User's Question**  
The user asked for a **chronological summary of rating changes** for the City of [REDACTED], [REDACTED] bonds. This required me to identify all events related to bond ratings in the provided documents and organize them by date.
- Step Two: Reviewing All Available Documents**  
I examined all three provided documents:
  - "[REDACTED] Bond Call\_1.3.2024.pdf" (dated January 3, 2024)
  - "[REDACTED] Rating Change\_9.8.2024.pdf" (dated September 8, 2024)
  - "[REDACTED] Rating Change\_7.31.2024.pdf" (dated July 31, 2024)
- Step Three: Identifying Relevant Bond Events**  
I needed to determine which documents contained actual **rating changes** versus other bond-related events. The document "[REDACTED] Bond Call\_1.3.2024.pdf" contained information about a **redemption notice** rather than a rating change, while the other two documents explicitly mentioned rating changes.
- Step Four: Creating a Comprehensive Timeline**  
To provide a complete chronological summary, I included all bond-related events from the earliest date (January 3, 2024) to the latest (September 5, 2024). While the January document didn't contain a rating change, it was still a significant bond event (redemption) that occurred before the actual rating changes in July and September.
- Step Five: Distinguishing Between Rating Changes and Other Bond Events**  
For each event in the timeline, I needed to clearly indicate whether it was an actual **rating change** or another type of bond event. For the January 3 document, I specifically noted that it "did not include a rating change" to avoid any confusion while still providing a complete chronological record.

For data-driven use cases, the system can also expose the underlying SQL queries generated by the LLM. This allows the company's campaign analysts and data professionals to validate the logic of recommendations before taking them to the field.

Figure 6. SQL code generated by the LLM, enabling analysts to validate query logic

```
SQL Code

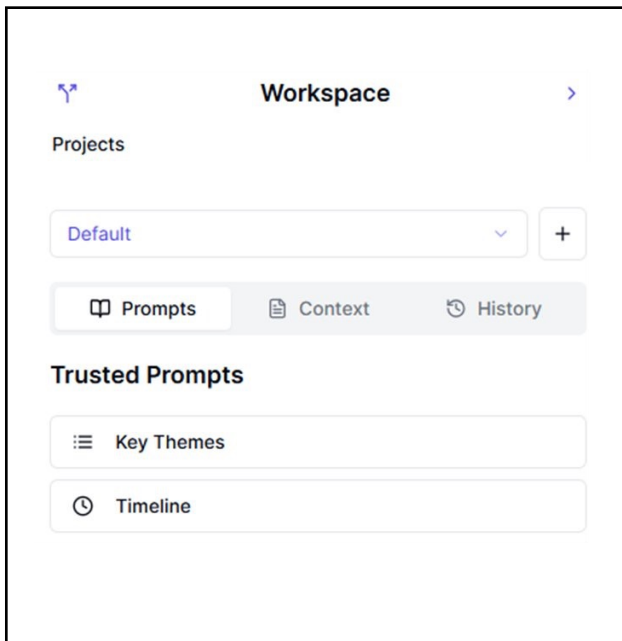
SELECT
  CASE
    WHEN balance < 0 THEN 'Negative Balance'
    WHEN balance = 0 THEN 'Zero Balance'
    WHEN balance BETWEEN 1 AND 500 THEN 'Low Balance (1-500)'
    WHEN balance BETWEEN 501 AND 2000 THEN 'Medium Balance (501-2000)'
    WHEN balance BETWEEN 2001 AND 5000 THEN 'High Balance (2001-5000)'
    ELSE 'Very High Balance (5000+)'
  END as balance_segment,
  COUNT(*) as customer_count,
  SUM(CASE WHEN term_deposit_sub = 'yes' THEN 1 ELSE 0 END) as conversions,
  ROUND(AVG(CASE WHEN term_deposit_sub = 'yes' THEN 1.0 ELSE 0.0 END) * 100, 2) as conversion_rate,
  ROUND(AVG(duration), 0) as avg_call_duration
FROM marketing
GROUP BY
  CASE
    WHEN balance < 0 THEN 'Negative Balance'
    WHEN balance = 0 THEN 'Zero Balance'
    WHEN balance BETWEEN 1 AND 500 THEN 'Low Balance (1-500)'
    WHEN balance BETWEEN 501 AND 2000 THEN 'Medium Balance (501-2000)'
    WHEN balance BETWEEN 2001 AND 5000 THEN 'High Balance (2001-5000)'
    ELSE 'Very High Balance (5000+)'
  END
ORDER BY conversion_rate DESC
```

### 3. Trusted prompts and guardrails

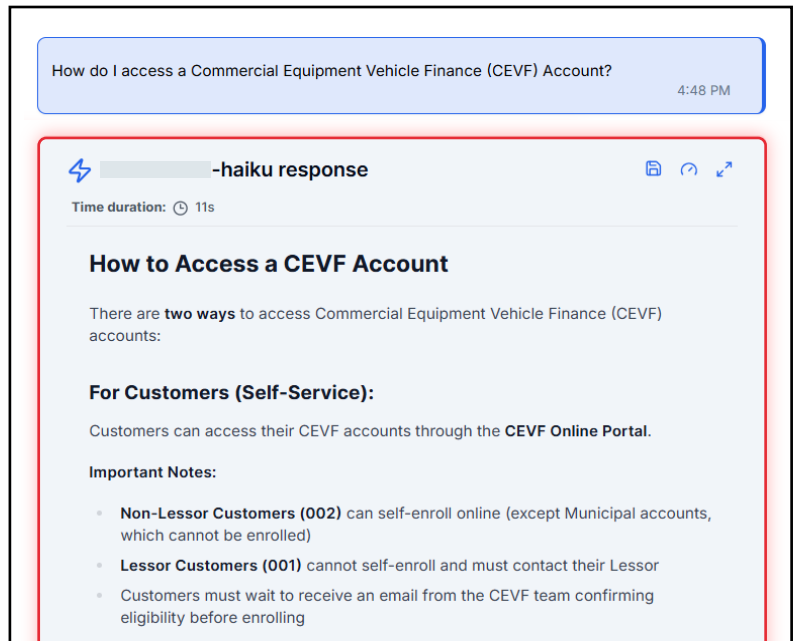
**Not all prompts are created equal.** This component establishes a tiered system that balances flexibility with control. Pretested prompt libraries cover approximately 80% of common use cases, providing consistent, high-quality outputs for routine tasks. Validated prompt templates have been rigorously tested and approved for specific business functions. Tiered access controls allow general users to access curated prompts while the company's power users can create custom queries, subject to validation.

Detailed guardrails detect and prevent out-of-policy discussions, potential regulatory violations, inappropriate use of sensitive data, and attempts to bypass security measures. The system provides visual cues to alert users when responses fall outside trusted boundaries.

**Figure 7. Trusted prompt library with curated, pretested queries**



**Figure 8. Visual cue (red border) alerting users to responses flagged as potentially inaccurate or incomplete**



This structured approach reduces errors while preserving the flexibility needed for innovation and edge cases. The visual differentiation between trusted and untrusted outputs helps ensure users apply the right level of scrutiny to each response.

## 4. Regeneration of responses

**One size rarely fits all in complex business environments.** This component acknowledges that AI responses often require iteration. Multiple response generations provide users with distinct options to choose from. Iterative refinement tools allow users to regenerate responses with additional context or modified parameters. Comparison interfaces highlight differences between response variations, and context preservation maintains conversation history while exploring alternatives.

A particularly powerful technique involves cross-model validation: If two separate LLMs from different vendors, trained on different datasets, produce the same answer to a question, users can have significantly higher confidence in that result. For example, when extracting filing dates from bond documents, agreement across models provides a strong signal of accuracy.

By treating AI outputs as starting points rather than final answers, this approach encourages critical evaluation and enables users to find the most appropriate solution for each of their specific needs.

**Figure 9. Model selection and additional context for response regeneration**

**Select Model for Regeneration** ✕

Choose a different language model to regenerate your response

 Claude-3.7-sonnet Balanced	 Claude-3-haiku Fast	 Llama3-70b-instruct Powerful
 Claude-4 Advanced	 Claude-4.5 Ultra	

**Additional Context for Generation**

Add specific instructions or information to guide the model's response...

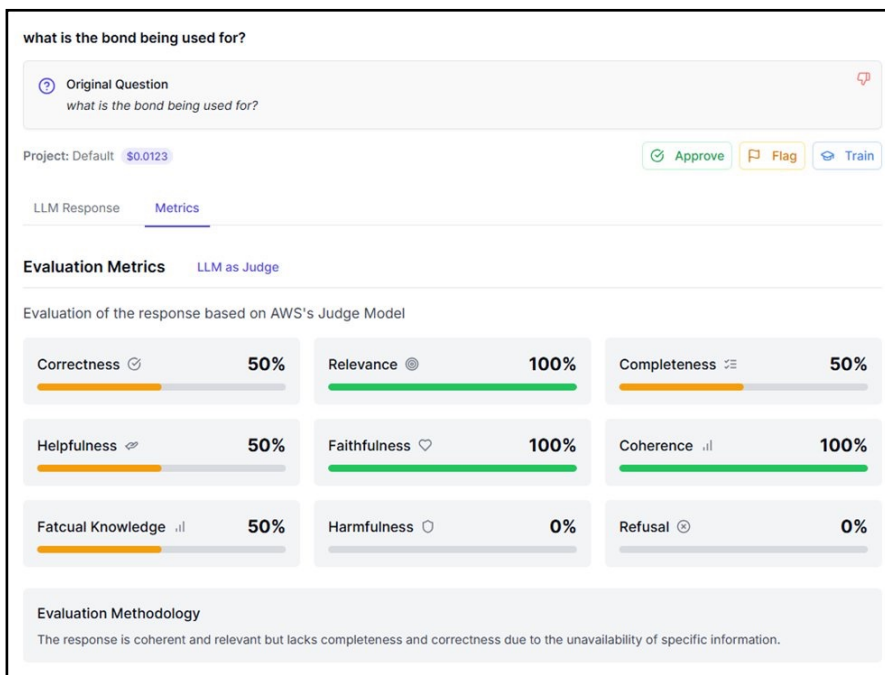
Cancel Regenerate

## 5. Confidence scores and LLM as judge

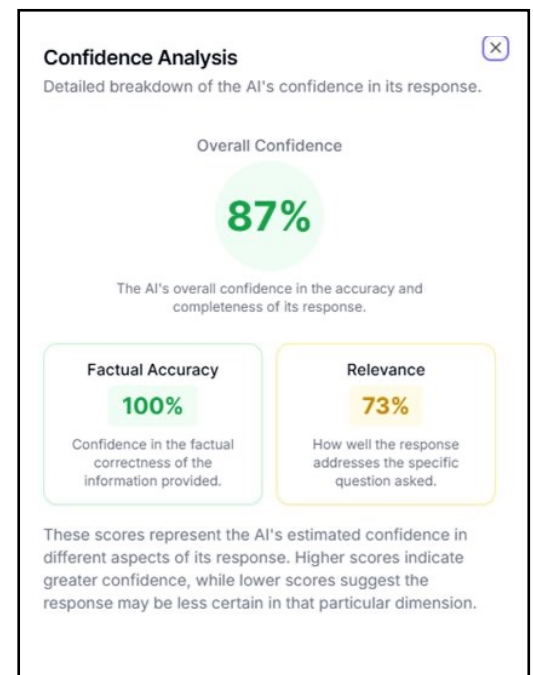
**Not all AI outputs are equally reliable.** This component provides sophisticated mechanisms for assessing response quality. Confidence metrics displayed alongside each response are derived from factors such as relevance to the original question, factual accuracy and completeness of the answer.

The LLM-as-judge functionality employs a secondary model (AWS judge model—an LLM that scores AI outputs) to evaluate the primary response for accuracy, completeness and potential issues. Calibration tracking monitors confidence score accuracy over time to improve reliability. These tools help users quickly identify which outputs require additional scrutiny versus those that can be trusted with higher confidence.

**Figure 10. Evaluation metrics from the LLM-as-Judge model across nine quality dimensions**



**Figure 11. On-demand confidence analysis with factual accuracy and relevance scores**



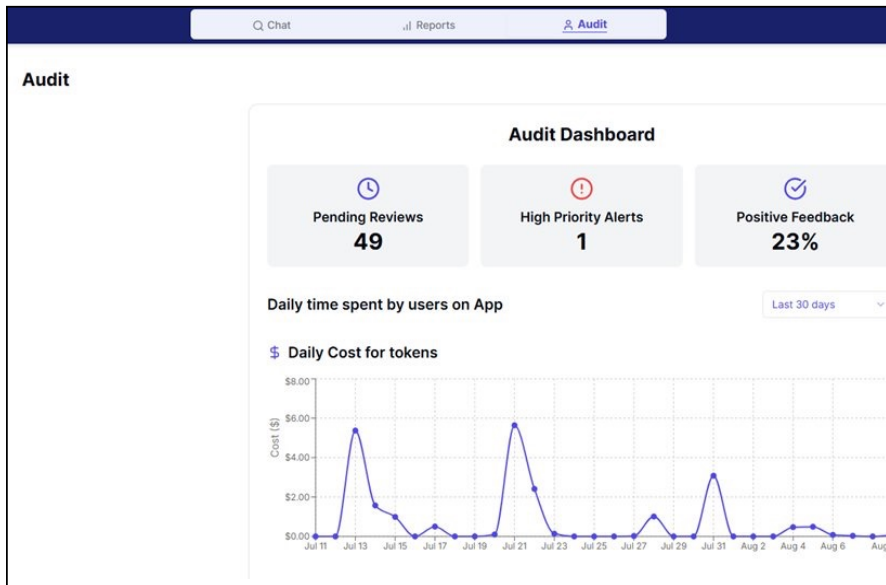
## 6. Human in the loop

**The ultimate safeguard is human judgment.** This component ensures meaningful human oversight at critical junctures. Supervisors and subject matter experts can review AI-generated responses before they impact business decisions, track usage patterns, identify potential misuse, and analyze response quality trends across teams and use cases.

Feedback mechanisms capture user corrections and validations to build organization-specific training datasets, improve model performance over time, and create institutional knowledge repositories. Escalation workflows automatically route uncertain or high-risk outputs to appropriate reviewers. Performance analytics help organizations understand where human intervention adds the most value.

This human-centered approach ensures that AI augments rather than replaces human expertise, maintaining accountability while leveraging the efficiency benefits of automation.

**Figure 12. Audit dashboard tracking pending reviews, alerts, usage patterns and costs**



**Figure 13. User feedback mechanism for flagging and rating AI responses**

The screenshot shows a 'Report an issue' form. It asks 'What's the issue with this response? (Select all that apply)'. The options are: Factually inaccurate, Irrelevant, and Offensive. There is also an 'Additional details (optional)' section with a text area for 'Please provide more details about the issue...'. At the bottom are 'Cancel' and 'Submit' buttons.

# *Putting Defensive UX into practice:* A responsible GenAI adoption blueprint

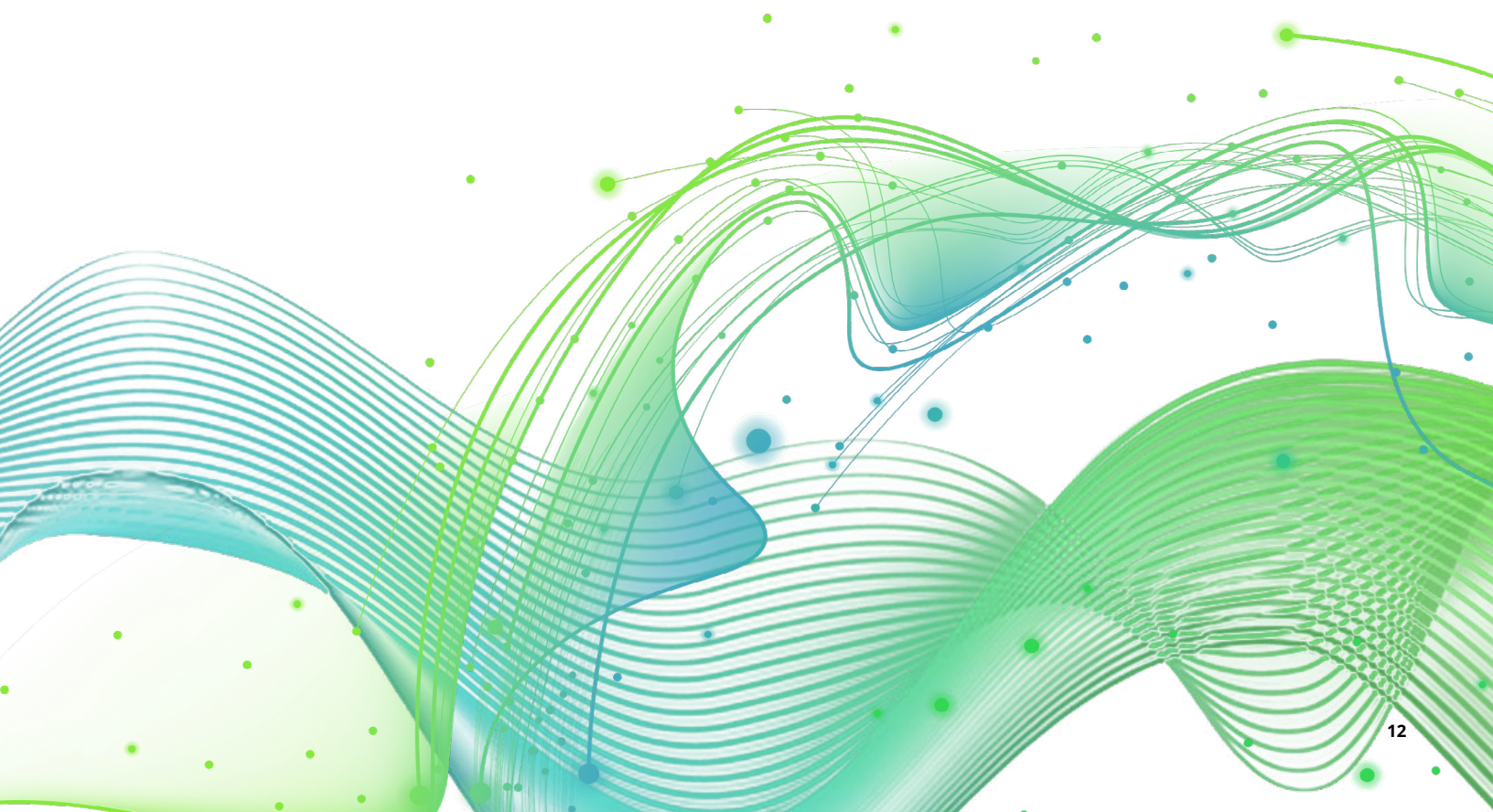
The defensive UX framework represents a deliberate counterpoint to the prevailing design philosophy that AI should be invisible. It reframes UX as a control layer for enterprise GenAI governance. In enterprise contexts, invisibility is a liability. Especially in regulated workflows, auditability is a requirement, not a “nice-to-have.” When the consequences of AI-generated outputs include regulatory penalties, reputational damage and flawed business decisions, organizations need users who are engaged, critical and accountable. Defensive UX operationalizes AI risk management at the point of use.

The six components of Defensive UX work in concert to create this environment: Context engineering gives users agency over inputs. Citations with verifications provides the evidentiary trail. Trusted prompts and guardrails establishes the boundaries. Regeneration of responses encourages exploration and cross-validation.

Confidence scores and LLM as judge quantifies reliability. And human in the loop helps ensure that technology serves human judgment, not the reverse. Together, these patterns can create traceable, reviewable decision workflows.

Deloitte's 10X Analyst, an enterprise platform developed with AWS, can help clients build—and use—safe, production-ready GenAI applications. The six-component framework can be treated as a reference architecture to adapt to specific domain and control requirements.

As organizations scale their GenAI initiatives, the question is not whether to adopt AI, but **how to adopt it responsibly**. Defensive UX provides an answer: It enables financial services enterprises to design experiences tailored to their requirements that can make humans smarter, more cautious and ultimately more accountable in their use of AI-generated insights.



# Authors



**Prakul Sharma**

Principal, AI & Data  
Deloitte Consulting LLP  
praksharma@deloitte.com



**Niraj Juneja**

Technology Fellow, AI & Data  
Deloitte Consulting LLP  
nirjuneja@deloitte.com

**Special thanks** to our AWS contributors:



**Simon Philips**

Principal, Global Partner Lead,  
Financial Services  
AWS



**Taehyun [TH] Her**

Principal Solutions Architect,  
Financial Services  
AWS

# Deloitte.

As used in this document, “Deloitte” means Deloitte Consulting LLP, a subsidiary of Deloitte LLP. Please see [www.deloitte.com/us/about](http://www.deloitte.com/us/about) for a detailed description of our legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting.

This publication contains general information only and Deloitte is not, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor. Deloitte shall not be responsible for any loss sustained by any person who relies on this publication.

Copyright © 2026 Deloitte Development LLC. All rights reserved.