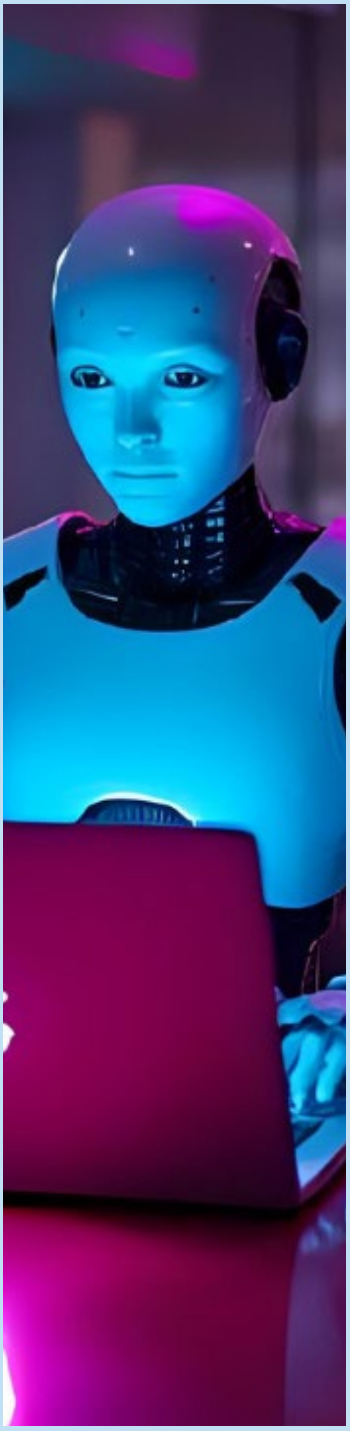
VIEW POINT



A NEW APPROACH TO ENSURING SAFE AGENTIC AI IN PUBLIC SECTOR APPLICATIONS

Abstract

As agentic AI systems become increasingly woven into critical government processes, from public service delivery to national defense and regulatory compliance, ensuring confidence in their results, maintaining citizen trust, and monitoring operational efficiency within government agencies requires high levels of reliability, security, and performance. Here we introduce a four-level framework designed to govern agentic AI systems at scale within the public sector.

Infosys® Public Services

The Challenges of Governing Agentic AI in the Public Sector

Agentic AI systems pursue complex goals and workflows with minimal human supervision. Unlike narrow AI models confined to single tasks, agentic AI exhibits decision-making and adaptive behavior similar to a human agent, enabling it to automate intricate public sector workflows such as permit approvals or disaster response logistics.

This greater autonomy brings immense potential for enhancing public sector efficiency and innovation. However, it also introduces risks if the agent behaves unexpectedly, potentially leading to incorrect decisions or disruptions in essential government services.

As agentic AI systems become more complex and widely adopted across government agencies, ensuring their reliability, efficiency, and security has become a critical challenge.

Effective monitoring and evaluation strategies are needed to detect failures, optimize performance, and improve AI decision-making through human feedback. Without proper oversight, an autonomous agent could stray from the intended public service goals, produce biased or incorrect outputs impacting citizens, or expose security vulnerabilities in critical infrastructure.

Monitoring and evaluating these agentic systems isn't straightforward, especially given the unique complexities of government operations. Agents are unpredictable and complex, exhibiting behavior that can be difficult to assess across diverse public sector contexts. This is compounded by the fact that there are no standardized evaluation metrics and benchmarks for agentic behavior, with many existing methodologies being too simplistic or context-dependent. This means that government agencies often cannot compare results across different programs or departments, and can inadvertently mask shortcomings in an agent's performance, which can erode public trust and accountability.

Further, ensuring evaluation covers ethical and safety dimensions is particularly challenging in the public sector, requiring careful monitoring of outputs and decisions to guarantee fairness and equitable service delivery for all citizens. There are also the integration and operational challenges inherent in complex government IT systems that make it hard to isolate the source of an error or performance issue. Finally, feedback and improvement loops are difficult to implement – human oversight from public servants is needed to review agent decisions and ensure input is systematically incorporated in a timely fashion.

The Infosys Multi-Level Framework for Government AI Governance

At Infosys Public Services, we use a four-level framework to evaluate AI agents, specifically tailored to ensure responsible deployment in the public sector:

1. **Infrastructure surveillance**
2. **Prompt or response evaluation**
3. **Performance monitoring**
4. **Feedback integration**

This framework provides a structured methodology to observe and control an agent's behavior at every level of an AI system, ensuring accountability and transparency in government operations. By combining telemetry, logging, automated metrics, and human oversight, the framework ensures continuous improvement and operational transparency in AI-driven public workflows.

1. Infrastructure Surveillance: The Foundation of Reliable Public Services AI

The foundation of the framework is a robust observability infrastructure to monitor the running environment of the AI agent. Even the most sophisticated agent deployed for the government is ultimately just software running on computers – thus traditional infrastructure monitoring is the first line of defense against operational disruptions and security breaches.

This layer involves tracking system metrics such as CPU and memory usage, network

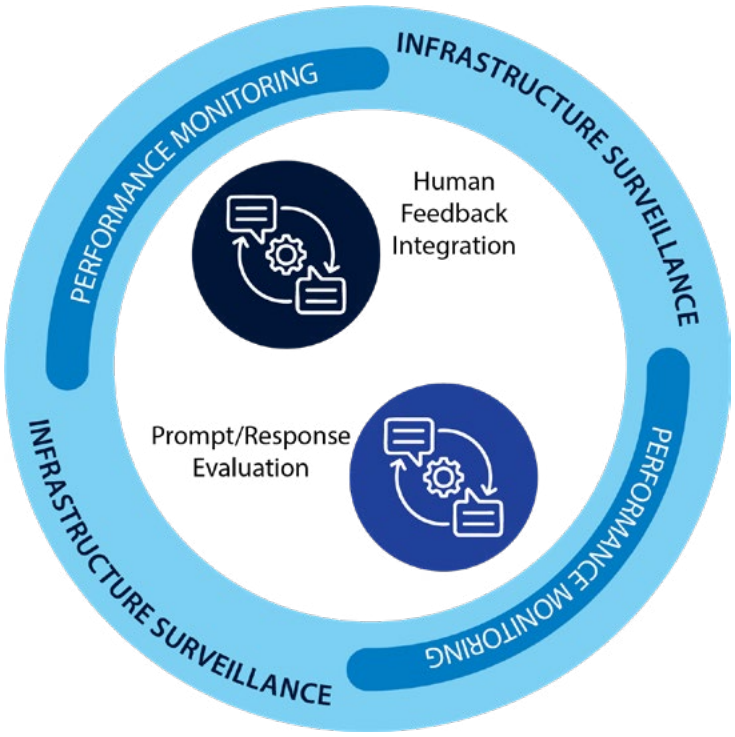


Figure 1. The Infosys Multi-Level Framework for Government AI Governance

I/O, request rates, error logs, and other low-level telemetry, all crucial for maintaining critical government systems.

A well-implemented observability system continuously monitors network traffic, request patterns, system health, and resource utilization to detect anomalies, such as unusual activity on citizen data portals or potential cyber threats. Dashboards display real-time agent logs and system metrics, enabling agency engineers to spot unusual spikes in activity or suspicious access attempts. If an autonomous agent is deployed via cloud microservices, tools like Prometheus and Grafana can be integrated to visualize these metrics and alert on out-of-bound values, ensuring the stability of public digital infrastructure.

2. Prompt/Response Evaluation: Ensuring Trustworthy Citizen Interactions

At the next level, the focus is on evaluating the content of the agent's interactions – namely the prompts it receives from citizens or government staff and the responses or actions it generates. Ensuring that AI-generated responses are accurate, appropriate, and aligned with expected behavior is critical for maintaining citizen trust, upholding public policy, and ensuring compliance with regulations.

This component introduces application-specific evaluation of the agent's decisions and outputs in the context of public service delivery. The best practice is to log all prompts given to the agent and the agent's responses, which are then compared against reference standards or metrics, such as established government guidelines or legal frameworks.

One effective technique is to compare AI outputs with human-generated outputs for the same or similar inputs. For instance, in deploying an AI assistant for public inquiries (e.g., permit applications, benefit eligibility), government agencies can log the AI's answer to a query and later compare it to how a human government agent answered that query (or an ideal answer from a public knowledge base).

Automated similarity measures like cosine similarity or BLEU scores can then quantify how close the AI's response is to the human answer. This helps in detecting omissions or inaccuracies that could lead to citizen frustration or incorrect information. Moreover, this evaluation is continuous: The system repeatedly evaluates its performance by examining new AI responses against the accumulating set of human-validated answers.

An important part of this evaluation layer is identifying malicious or problematic prompts and ensuring the agent's responses are robust against misuse. The monitoring framework flags any input that matches patterns of known prompt injection attacks or disallowed content, crucial for safeguarding government systems. If a user tries to prompt the agent to reveal confidential citizen information or produce discriminatory content, this should trigger an alert or a safe failure, upholding ethical AI principles in government.

On the output side, given that agents are typically embedded in key government processes, agencies must be able to guarantee response quality and safety. One novel approach is using a Large Language Model (LLM) "as a judge" to evaluate the agent's outputs (see Figure 1), ensuring outputs meet public sector standards. In other words, a separate AI model (or the agent itself in critique mode) can assess whether a given response is well-formed, helpful, and non-harmful. This ongoing content monitoring is crucial for detecting biases or hallucinations early and retraining or adjusting the agent accordingly to prevent inequities in public service.

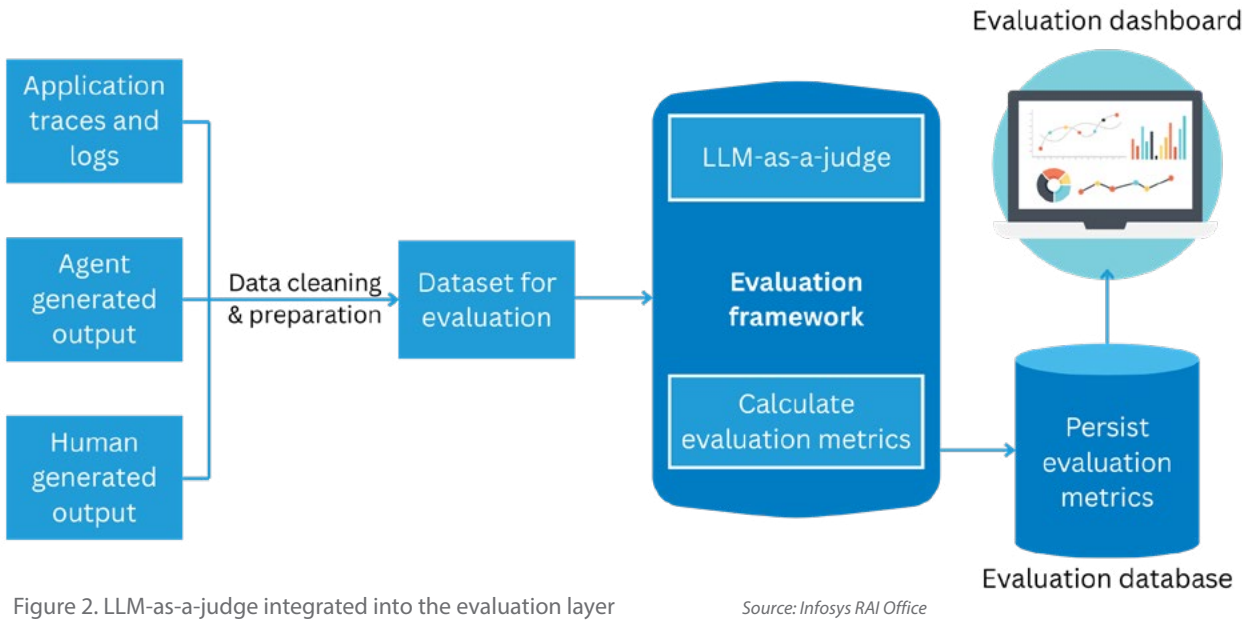


Figure 2. LLM-as-a-judge integrated into the evaluation layer

Source: Infosys RAI Office

3. Performance Monitoring: Optimizing Resource Utilization and Service Delivery

Beyond what the agent accomplishes, it's increasingly vital to measure how efficiently it performs—especially when managing taxpayer-funded operations—to ensure cost-effective service delivery. Performance monitoring is therefore the third pillar of the framework, focusing on timing, throughput, and resource usage at the task level for government applications. Optimizing the speed and efficiency of AI-driven public workflows is essential for scalability of services and citizen satisfaction. An otherwise correct agent could still fail if it is too slow or costly to be practical for widespread government adoption.

The framework measures execution times for each step the agent takes and for entire task flows, such as processing a social service application or routing a public safety request. Each agent task is logged with timestamps for key stages: When the task was initiated, when the agent began processing, and when the task was completed. From these, agencies can compute durations and latencies for every action an agent makes.

By aggregating this data, bottlenecks can be identified – perhaps a particular sub-module (like a knowledge retrieval step for regulations) consistently lags, or certain legacy tools the agent calls are slowing it down.

Further, visualizing performance metrics over time using control charts helps establish baselines and detect regressions. For example, if an "average response time per citizen query" increases significantly after a new agent version is released, the chart will show a clear deviation, prompting investigation to maintain critical service levels. At Infosys Public Services, we also monitor time-to-first response and throughput (tasks per minute) as key indicators of performance for government systems.

In summary, the performance monitoring layer treats the agent as a software system whose service level agreements (SLAs) for public service delivery must be met – ensuring it operates within acceptable time and resource limits, optimizing taxpayer investment.

4. Feedback Integration: Human Oversight as the Cornerstone of Public Sector AI

The final and arguably most critical layer of the framework is integrating human feedback from government employees and citizens to continually refine the agent. Autonomous does not mean unattended – human oversight is a cornerstone of safe agentic AI, particularly in the public sector where accountability is paramount. In this methodology, we capture explicit feedback from users (citizens, government staff) and human evaluators (subject matter experts, policy analysts) and feed it back into the improvement cycle.

One approach is to solicit user ratings or critiques after interactions. For instance, after an agent completes a task or answers a question regarding a public service, the citizen or a government moderator might rate whether the outcome was satisfactory. These ratings become another data point in the evaluation repository.

More formally, we can measure what is known as golden instruction adherence – whether the agent followed the key instructions or policies it was supposed to, such as specific government regulations or ethical guidelines, and exception F-scores – rates at which the agent triggers exceptions or falls back to human intervention.

By tracking how often and in what way the agent deviates from expected behavior (the "golden path" of compliant and ethical public service), organizations can improve the AI's decision-making over time. For example, if an AI assistant repeatedly asks for clarification on certain citizen requests, this might indicate unclear instructions in those scenarios, which means developers should adjust the agent's prompt or training data to improve clarity and efficiency for members of the public.

Human feedback is vital for future-proofing the AI system in face of changing conditions, such as new legislation or evolving public needs. Real users will inevitably use the agent in unanticipated ways. Monitoring how these real interactions differ from the training scenarios provides insight into where the AI might need adaptation for better public service.

Our framework recommends periodic review sessions where human reviewers from relevant agencies go through logs of agent decisions, particularly the borderline cases, either flagged by the system or sampled randomly. In these sessions, reviewers can label decisions as correct or flawed and provide explanations crucial for maintaining public trust and compliance. This creates high-quality data for fine-tuning the agent or updating its heuristics to align with public interest.

Some modern tools facilitate this loop; for instance, LangSmith, by LangChain, allows developers to collect traces of agent runs and attach feedback or ratings to each run, all within a single platform. Such platforms support LLM-native observability, meaning they are built to handle the nuances of language model outputs and chain-of-thought traces. They let human operators search and filter agent runs (e.g., finding all runs where the agent gave a low-quality answer) and then examine those in detail to debug or improve the logic for public sector applications.

Leveraging Specialized Tools to Accelerate Safe AI Deployment

Developing this framework requires dedicated effort, a strategic partner, and ideally the establishment of a platform engineering squad within government IT to offer disparate teams access to the technology in an automated fashion. However, there are tools available that can speed up the process of safeguarding agentic AI systems across the four layers of the framework.

LangSmith, a platform designed to support the development, testing, and optimization of LLM applications, provides features for debugging, tracing, and monitoring agentic systems. Using LangSmith,

Real-World Impact: Government Application in Practice

Case Study: Enhancing Citizen Services for IT Support

A Fortune 500 company, similar to a large government agency managing internal IT support or citizen help desks, deployed an agentic AI assistant to handle IT service deck requests such as password resets and troubleshooting. This is a complex, real-world workflow automation scenario directly transferable to public service. The company implemented a monitoring and evaluation scheme following the Infosys multi-level framework described. At the infrastructure level, they aggregated logs of all agent actions in their cloud environment. Early on, this helped catch a misconfiguration where the agent was inadvertently making an API call twice for each user request - logs showed an abnormal patter of duplicate requests, altering engineers to fix the logic, which could prevent unnecessary resource consumption in government IT. At the integration level, the quality of the agent's responses was evaluated by comparing them with the human IT support responses. For a period, the AI's answers and the human operators' answers to similar tickets were collected.

Using semantic similarity and manual review, the team identified areas where the AI's answers were lacking. For example, the AI often omitted an apology in responses when a user faced an inconvenience, whereas human agents always included a polite apology. This was flagged through content monitoring, and the prompt was adjusted to include an apology where necessary, highlighting the importance of empathetic interactions in public service. They also used an LLM-as-a-judge approach: for each resolved ticket, another language model rated whether the AI's resolution was satisfactory or if the user might need follow-up-up. These ratings surfaced a confusion - something the judge model could catch by "imagining" a user's perspective, critical for clear and helpful communication in government.

The outcome of this deployment was that the company managed to automate a significant portion of routine IT requests with the agent, while the monitoring framework provided confidence that any decline in performance or unexpected behavior would be quickly detected and addressed. This goes some way to highlight how multi-level evaluation in a real enterprise setting not only averts failures, but also guide the AI to a of performance and reliability that meets business requirements, demonstrating its direct applicability to public sector service delivery.

government developers can log each agent run with all intermediate steps and then use a dashboard to see aggregate statistics like latency, tokens used, cost, and user feedback over time. This makes it easier to spot outliers – for example, a particular day where the agent's error rate spiked – and drill down into what went wrong, ensuring consistent public service delivery.

Another relevant tool is **Galileo**, an AI evaluation and observability platform. Galileo's recently introduced Agentic Evaluations offers an end-to-end framework for evaluating AI agent performance, providing visibility into every action across entire workflows. It supports both system-level monitoring and step-by-step analysis, enabling government developers to build more reliable and trustworthy agents for public service. Platforms like this often include a guardrail metrics store and modules for prompt evaluation, fine-tuning, and monitoring in production.

For example, Galileo's Monitor module allows teams to set up custom metrics such as a hallucination rate or an accuracy score and track them in real time as the agent interacts with citizens or internal staff. It can automatically flag outputs that have a high likelihood of being hallucinated or harmful, using research-driven metrics, and thus helps in proactively catching errors before they impact public services or trust. Ensuring Stability, Safety, and Reliability of AI in Public Services

AI is increasingly used in key government processes, from logistics for public safety to citizen customer service. Agentic AI – one of our top 10 AI imperatives for 2025 – is a progression of generative systems towards autonomous entities that aims to make the public sector workforce more productive and hopefully less stressed, allowing human staff to focus on complex cases requiring nuanced judgment. However, this increased autonomy still needs structured oversight to ensure accountability and public trust.

Monitoring and evaluation must therefore be multi-layered, addressing challenges such as unpredictable AI behavior and the lack of standardized evaluation metrics across diverse government contexts.

A structured framework — incorporating infrastructure surveillance, prompt-response monitoring, performance tracking, and human feedback loops — ensures that government organizations maintain tight control over AI systems, allowing for continuous refinement and risk mitigation. By integrating these dimensions, AI-driven public workflows can achieve stability, safety, and reliability, preventing errors from escalating into significant public service disruptions or ethical failures. Doing so will ensure AI continues its march forward in the public sector, and help government agencies ensure that both employees and the public accept the results of these autonomous marvels, fostering trust in digital government initiatives.

Key Insights for Government Leaders

As agentic AI systems become more complex and widely adopted across federal, state, and local government operations, **ensuring their reliability, efficiency, and security has become a critical challenge** for public service delivery.

Many existing methodologies for monitoring and evaluating these systems are too simplistic and context-dependent, potentially masking shortcomings in an agent's performance and impacting citizen trust.

At Infosys Public Services, we use a **four-level framework to evaluate AI agents**, specifically designed to address the unique governance needs of the public sector, including infrastructure surveillance, prompt evaluation, performance monitoring, and feedback integration.

Specialized **tools are available to speed up this process**, including LangSmith and Galileo, an AI evaluation and observability platform, facilitating responsible AI adoption in government.

This approach isn't hypothetical. Infosys is successfully using it at our client organizations, demonstrating proven results such as reducing false escalations by **38%** and increasing output quality by **45%** in audit-like scenarios, which is directly applicable to improving government operations and public accountability.

To know more visit us at **www.infosyspublicservices.com** or email **Jonah Czerwinski.**

For more information, contact askus@infosyspublicservices.com

Infosys® Public Services

Infosyspublicservices.com

Stay Connected