COMPAiSS

Execution-Gated Inference

A Structural Solution to AI Hallucination in Regulated Institutional Environments

Frank P. Harvey, PhD
Senior Advisor, Dalhousie University
Founder, CEO and Chief Architect, COMPAiSS Inc.

May 2026

Patent Pending: CIPO 3,299,174 (Canada) / USPTO 19/455,963 (United States)
Contact: frank.harvey@dal.ca
compaiss.ca

Abstract

AI hallucination is universally acknowledged as a structural limitation of generative AI systems, not an incidental defect. Existing mitigation strategies - including Retrieval-Augmented Generation (RAG), post-generation filtering, confidence scoring, and human oversight - reduce hallucination rates but cannot eliminate them. Peer-reviewed empirical research confirms that even the most sophisticated RAG-based enterprise systems hallucinate between 17% and 33% of the time in high-stakes regulated environments. This paper argues that the persistence of hallucination under all current mitigation approaches is a direct consequence of a shared architectural assumption: that generative inference is always instantiated, and that safety is enforced afterward. COMPAiSS introduces an alternative architectural principle - execution-gated inference - in which a generative inference runtime is instantiated only when a pre-execution authorization gate confirms that verified institutional sources exist to support a response. When authorization fails, no generative computation occurs. There is no unauthorized inference from which hallucinations can arise. This paper describes the architecture, its theoretical basis, its empirical performance, its distinction from all prior art, and its application to regulated institutional environments including government service delivery, healthcare, legal services, and higher education. The COMPAiSS architecture is the subject of patent applications currently under examination in Canada and the United States.

1. The Hallucination Problem Is Structural, Not Incidental

Large language models (LLMs) generate text by predicting statistically likely continuations of input sequences. They are not truth machines. They produce plausible-sounding outputs calibrated to approximate the distribution of human language, not to represent verified facts. This architectural reality - training on next-token prediction without explicit negative labels - is the root cause of AI hallucination.

This is not a defect of any particular model or product. It is a property of the class of systems. OpenAI's own research explicitly describes hallucinations as "a fundamental challenge for all large language models," noting that "accuracy will never reach 100%" because some questions are inherently unanswerable or ambiguous, and because standard accuracy-based evaluations "reward guessing over acknowledging uncertainty."

The research consensus is unambiguous: hallucinations are an inherent and persistent limitation of current LLM architectures. A 2024-2025 comprehensive survey concludes that state-of-the-art detection and mitigation methods "fall short of fully eliminating hallucinations," characterizing this as "a key open challenge" and framing hallucinations as an ongoing research problem, not a solved issue. NIST's AI Risk Management Framework explicitly instructs organizations to document residual risk - the risk that remains after all mitigation measures have been applied - recognizing that "not all incidents and failures can be eliminated."

"Commercially-available RAG-based legal research tools still hallucinate. Nearly 1 in 5 of our queries caused the tools we tested to respond with misleading or false information." - Magesh et al. (Stanford / Yale), Journal of Empirical Legal Studies, 2025

The practical consequences of this residual risk depend entirely on the deployment context. For general-purpose consumer AI, occasional hallucinations are an accepted tradeoff for broad utility. For regulated institutional environments - government benefit programs, healthcare, legal services, higher education - the consequences are categorically different. A citizen who receives hallucinated information about their Employment Insurance eligibility, their Old Age Security entitlement, or their legal rights may act on that information with serious and irreversible consequences.

The field's response to this problem has been to invest in mitigation. RAG, guardrails, post-generation filtering, confidence scoring, human review workflows, and multi-layer moderation systems have all been developed, deployed, and refined. Each approach reduces hallucination rates. None eliminates them. This paper argues that the reason is architectural: all current mitigation approaches share the assumption that generative inference runs first, and that safety is enforced afterward. That assumption is the problem.

2. The Limits of Generation-First Architecture

Every major approach to AI hallucination mitigation in current deployment - RAG, guardrails, moderation, filtering, human review - operates on outputs that a generative model has already produced. The standard execution model is:

Standard AI Execution Lifecycle (all current systems)
Request received → Inference runtime instantiated → Generative model executes → Output produced → Safety / filtering / moderation applied → Response delivered

Authorization and safety mechanisms in this paradigm are admission controls and post-generation filters. They govern access to an already-instantiated inference service, or they evaluate outputs after the model has run. The generative computation - the stochastic process that can produce hallucinations - has already occurred regardless of whether any particular output ultimately reaches the user.

This is not an oversight. It reflects a deliberate and rational architectural decision for general-purpose systems. The generative runtime is persistent infrastructure. Safety is layered on top. This architecture is optimized for breadth, availability, and responsiveness.

But this same architecture carries a structural limitation that no amount of downstream safety investment can fully resolve: if the model generates hallucinated content, that content must be detected and filtered before delivery. Detection is incomplete. Even the best-resourced and most carefully designed detection systems - operating in narrow, well-controlled domains with curated data - produce detection F1 scores in the range of 60-65%, meaning a significant fraction of hallucinated content is not caught.

2.1 RAG Does Not Solve the Problem

Retrieval-Augmented Generation is the dominant industry response to hallucination. By connecting a generative model to an external knowledge base and requiring it to ground responses in retrieved documents, RAG substantially reduces hallucination rates relative to ungrounded generation - often by 30-50% relative to a baseline.

This improvement is real and valuable. But it is not sufficient for regulated institutional environments. The Stanford University and Yale Law School study that empirically tested the leading RAG-based legal AI platforms - Thomson Reuters (Westlaw) and LexisNexis - found hallucination rates of 17-33% in real-world legal research tasks. These are enterprise systems with substantial engineering investment, narrow domain scope, and curated legal corpora. They still hallucinate more than one time in six, and in the case of Thomson Reuters, more than one time in three.

RAG fails for multiple structural reasons that no tuning or configuration can fully address:

- Retrieval failure - the retriever may miss critical documents or retrieve partial evidence, causing the model to extrapolate from parametric memory.
- Noisy retrieval - retrieved passages may be only loosely related, causing models to synthesize plausible but incorrect claims.
- Context misuse - even with correct passages present, models sometimes produce content that is unsupported by or contradicts the retrieved references.
- Reasoning errors - even under perfect retrieval, models make logical mistakes: miscounting, misaggregating, or misinterpreting evidence.

The RAG limitation in one sentence
RAG substantially reduces 'I didn't know that' hallucinations while leaving intact hallucinations caused by reasoning, misinterpretation, or failure to faithfully track retrieved evidence.

2.2 Post-Generation Controls Cannot Prevent What Has Already Happened

Guardrails, moderation layers, output filtering, and human review are controls applied after the generative model has executed. They can catch a fraction of hallucinated outputs before delivery. They cannot un-generate content that has already been produced, and they cannot catch everything.

The mathematical formalization of safety in the prior art is revealing. Safety frameworks are defined as functions applied to model outputs:

$$y' = G(M(y|x))$$

Where M is the model, x is the input, y is the generated output, and G is the safety function. This formulation requires that M executes - that inference has already occurred - in order for G to have anything to operate on. Safety, in this paradigm, is definitionally post-generation.

This includes refusal. When a guarded system refuses to answer a question, that refusal is itself an output of generative inference. The model ran. It produced a refusal. This is architecturally different from a system in which the model never ran.

Cold-start literature further reinforces this paradigm's dominance. Inference runtimes across all major cloud platforms - AWS SageMaker, Google GKE, Azure ML - are provisioned as persistent or auto-scaled infrastructure whose existence is independent of individual authorization decisions. Per-request instantiation is treated as a performance anti-pattern to be minimized, not a safety mechanism to be embraced. High availability is an operational requirement. Non-existence of a runtime is an error condition, not a designed terminal state.

3. Execution-Gated Inference: The Architectural Alternative

COMPAiSS introduces a categorically different architectural principle. Rather than generating first and filtering after, COMPAiSS evaluates authorization before any generative inference is permitted to occur.

> **COMPAiSS Execution Lifecycle**
> Request received → Authorization gate evaluates institutional sources → IF authorized: Inference runtime instantiated → Response generated and delivered → IF not authorized: No inference runtime instantiated. No generative computation executed. Safe failure response delivered at zero marginal compute cost.

The pre-inference execution gate is not a guardrail, a filter, a moderation layer, or an access control applied to a running inference service. It is a condition that determines whether an inference runtime may exist at all for a given request. When authorization fails, no generative computation occurs. There is no inference to hallucinate from.

This distinction is not semantic. It reflects a fundamentally different placement of authority within the execution lifecycle:

- In generation-first systems: the model exists → authorization governs what it produces.
- In execution-gated systems: authorization determines whether the model exists → the model only exists when authorized sources confirm it may.

"COMPAiSS eliminates scope-violation hallucinations by design, and materially reduces - though does not eliminate - generation-quality risks within authorized scope through structured parsing and tightly bounded inference contexts." - COMPAiSS Technical Documentation

The diagram below illustrates the architectural distinction. Both execution paths begin with a client query. In generation-first systems, the model is always instantiated and safety is applied afterward. In execution-gated systems, the authorization gate determines whether the model exists at all for that request.

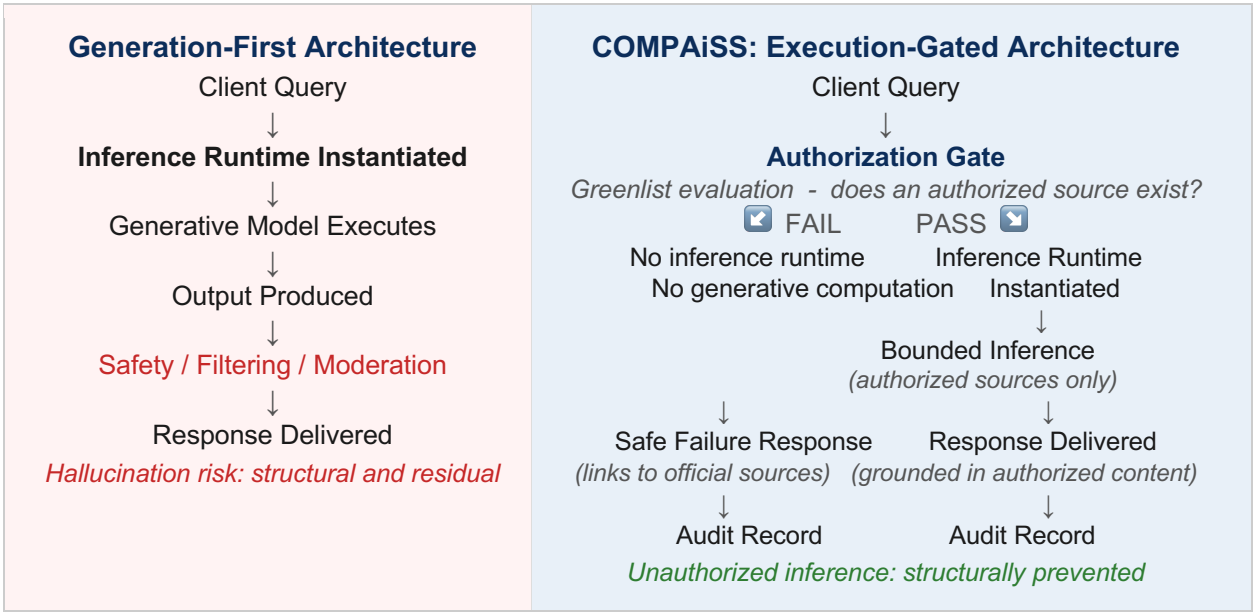

Figure 1: Architectural comparison of generation-first systems (left) and COMPAiSS execution-gated inference (right). In generation-first systems, a generative model always executes and safety is applied to its outputs. In the COMPAiSS architecture, the authorization gate determines whether a generative inference runtime may exist at all for a given request.

3.1 What the Gate Evaluates

The pre-inference execution gate performs automated discovery against an institution-approved source list - the greenlist - to determine whether authoritative institutional materials exist to support a response to the received query. Three conditions must be satisfied before inference is permitted:

- The query falls within the institution's defined scope.
- Authoritative, institution-approved materials exist to support an answer.
- Producing an answer would be defensible within the institution's governance framework.

If any condition is not met, the system enters a non-generative operational state that is terminal with respect to generative inference. No model weights are loaded. No inference execution

context is allocated. No stochastic generative computation is performed. The client receives a structured safe-failure response - direct links to authoritative institutional sources - at zero marginal compute cost.

If all conditions are met, a generative inference runtime is instantiated with a bounded epistemic environment: content sourced exclusively from the pre-authorized institutional pages. The model does not reason over a broad, unstructured corpus. It reasons over institution-approved materials only.

3.2 Defense in Depth

COMPAiSS does not claim that execution gating alone eliminates all possible generation errors. It claims something more precise: that execution gating eliminates hallucinations arising from unauthorized inference execution, and materially reduces generation-quality risks within authorized scope through the tight bounding of the inference context.

Within authorized scope, COMPAiSS employs a defense-in-depth architecture:

- Authorization gating prevents unauthorized primary model execution - eliminating hallucinations that arise from unauthorized inference execution on out-of-scope queries.
- Instruction-based constraints guide model behavior during authorized inference - reducing the probability of generation errors within scope.
- Post-generation URL validation ensures that all links in responses point to authorized institutional sources - preventing citation of unauthorized content.

This is not a sequence of filters applied after a generation-first model. It is a layered governance architecture in which the most fundamental constraint - whether the model runs at all - is the first and structurally most important control.

4. Architectural Distinction: Prior Art and Execution-Lifecycle Analysis

Four independent prior-art research probes - conducted using Perplexity, Claude, and Gemini across USPTO, WIPO/PCT, and Google Patents, supplemented by peer-reviewed systems literature - converged on a consistent finding. Across the surveyed record, we identified no prior-art reference within the surveyed literature and patent corpus that conditions the existence of a generative inference runtime on pre-execution authorization as an architectural design principle. This is an evidentiary finding within the scope of sources examined, not a claim of universal absence.

The dominant architecture in the entire prior-art corpus assumes inference as a persistent or auto-scaled service whose existence is independent of authorization or safety policy. Authorization determines whether a principal may invoke an already-running service. Safety determines what outputs may be delivered. Neither determines whether the service exists.

This assumption is not incidental. It is structurally embedded in the mathematics of safety frameworks, the design of cloud ML infrastructure, and the optimization goals of LLM serving systems. The prior-art record shows multiple forms of teaching away from execution-gated architecture:

Teaching-Away Type	Description
Mathematical	Safety frameworks define safety as y' = G(M(x)) - a function applied to model outputs. This requires M to execute, making non-execution incompatible with the standard safety model.
Architectural	Cloud ML platforms (SageMaker, GKE, Azure ML) provision inference as persistent or auto-scaled infrastructure. Authorization governs access; it does not control runtime existence.
Operational	Cold-start literature treats per-request instantiation as a performance anti-pattern. High availability is an SLA requirement. Non-existence of a runtime is an error condition, not a designed state.
Conceptual	Cloud platforms intentionally decouple authorization logic from infrastructure lifecycle control. This separation is a foundational design principle, not an accidental omission.
Systemic	Refusal frameworks - Constitutional AI, Learn-to-Refuse, Bedrock Guardrails - define refusal as an output of generative execution, not as non-execution. Safety requires the model to run.

Independent assessments by Claude, Gemini, and GPT-4 each confirmed these findings within the scope of that review. Within our review, no examined reference conditions inference existence on authorization as a designed architectural property. Within our review, no examined reference describes non-instantiation as a terminal designed system state. The combination of IAM/RBAC, autoscaling, and moderation - the most plausible combination from surveyed art - does not in the prior-art record lead to execution-gated inference, and is in several respects actively discouraged by the art's stated optimization goals.

The COMPAiSS architecture is the subject of patent applications under examination in Canada (CIPO 3,299,174) and the United States (USPTO 19/455,963). The independent claims describe a system in which a generative inference runtime is instantiated only when an authorization state exists, and in which when the authorization state does not exist, no generative inference runtime is instantiated, no executable inference context exists, and no stochastic generative computation is performed.

5. Empirical Performance: The Service Canada Deployment

The Service Canada AI Information Assistant (SCAI) - deployed at compaiss.ca/service-canada/ - is the first institutional implementation of the COMPAiSS execution-gated inference architecture in a federal government service delivery context. It provides natural language query capabilities for Service Canada programs including Employment Insurance, Canada Pension Plan, Old Age Security, Guaranteed Income Supplement, Social Insurance Number, and related services.

Empirical testing conducted May 26, 2026 produced the following results:

5.1 Bilingual Performance

A French-language query about Old Age Security eligibility and benefit amounts returned a complete, accurate response in French, grounded exclusively in French-language Government of Canada sources. Seven distinct /fr/ URLs were cited across the response, covering:

- securite-vieillesse (OAS main page)
- securite-vieillesse/admissibilite (eligibility)
- supplement-revenu-garanti (GIS)
- supplement-revenu-garanti/montant-prestation (GIS amounts)
- securite-vieillesse/demande (application)
- securite-vieillesse/paiements (payment schedule)

Quoted source material was drawn directly from French-language Government of Canada pages - not translated from English. This confirms that the system is not translating English responses into French, but is grounding French-language queries in French-language authoritative sources. This directly satisfies the Official Languages Act requirement for equivalent service quality.

5.2 Response Quality

Responses demonstrate structured, plain-language delivery appropriate for Service Canada's diverse client population, including:

- Clear program eligibility conditions drawn directly from authoritative source pages
- Explicit citation of Government of Canada source URLs in all responses
- TTY telephone numbers (for deaf and hard-of-hearing clients) included in every response containing contact information
- Actionable next steps and a follow-up input for personalized guidance
- Appropriate escalation to Service Canada staff for complex or sensitive queries

5.3 Phase 1 WCAG 2.1 AA Accessibility Audit - Completed

An automated accessibility audit (axe-core via Playwright headless Chromium) was run directly against the live interface source code. Seven violations were identified across two audit phases and all have been remediated:

- Submit button colour contrast - resolved by changing font-weight from 600 to 700, lowering the applicable WCAG threshold to 3.0:1. Service Canada brand red (#ff0000) preserved unchanged.
- Missing <main> landmark, missing H1, and content outside landmarks - resolved by two HTML tag changes.
- Placeholder text contrast - resolved by adding a ::placeholder CSS rule (colour #595959, ratio 6.89:1).
- ARIA live region added to response container - screen readers now announce loading state and response arrival.
- Decorative emoji in response output wrapped in aria-hidden spans - screen readers no longer announce emoji descriptions before headings.
- AI disclosure text added: 'This is an AI assistant. It provides information only and does not make decisions about your benefits.'

Fourteen accessibility rules passed in the original audit, including all form labelling, language attribution, zoom scaling, and keyboard accessibility checks. Phase 2 manual review - live

keyboard navigation, NVDA/VoiceOver screen reader testing, and mobile touch target verification - is the next scheduled accessibility milestone.

5.4 Multilingual Performance

Beyond official languages bilingual routing, the SCAI interface was tested with a natural-language query submitted in Tagalog - the language of over 900,000 Filipino Canadians, one of the fastest-growing immigrant communities in Canada and a significant Service Canada client population.

The query 'Paano ako mag-apply para sa Employment Insurance kung nawalan ako ng trabaho?' (How do I apply for Employment Insurance if I lost my job?) returned a complete, accurate response in Tagalog. The response included:

- Correct 2026 EI rates ($68,900 maximum insurable earnings, $729/week maximum benefit)
- Correct 55% calculation formula with worked examples
- Correct regional unemployment rate examples with actual hours requirements
- Correct waiting period waiver information (March 30 through April 11, 2026)
- Step-by-step application process in natural Tagalog
- TTY number (1-800-529-3742) for deaf and hard-of-hearing clients
- Links to Government of Canada source pages (English-language, as no Tagalog-language GOC pages exist)

The response demonstrates the multilingual translation pathway - distinct from the bilingual source-routing pathway used for French queries. For languages other than French, SCAI translates the query for governance evaluation, retrieves from English-language authorized sources, and delivers the response in the client's original language. The governance gate operates identically regardless of query language.

> **Multilingual governance significance**
> A Filipino-Canadian worker who loses their job and seeks EI guidance can receive accurate, current, source-grounded information in Tagalog at any hour without a 41-minute hold time. The system cannot hallucinate EI rates it has retrieved from the official 2026 rate card. The governance architecture makes no distinction between a query in English and a query in Tagalog - authorization is evaluated in English against verified sources, and the response is delivered in whichever language the client used.

5.5 What a COMPAiSS Deployment Feels Like Operationally

The architecture described in previous sections is most easily understood through the lens of how it operates from the perspectives of the three parties who interact with it: clients, institutional administrators, and the governance record.

The Client Experience

A client visits a Service Canada page and sees the query interface. They type a question in plain language - in any language. If their question falls within the greenlist scope, they receive a structured, plain-language response grounded exclusively in Government of Canada source

pages, with direct links to those sources and, where relevant, contact information including TTY numbers for clients who are deaf or hard of hearing. The experience is immediate, available around the clock, and requires no specialized knowledge of program names or policy terminology.

If their question falls outside scope - because no authorized source addresses it, or because it concerns a program not on the greenlist - they receive a clear, plain-language message directing them to Service Canada staff through existing channels. There is no ambiguous partial answer, no hallucinated content, and no confusing non-response. The failure mode is transparent and immediately actionable.

Greenlist Management: Institutional Control in Practice

The greenlist - the institution-approved source list that defines the system's authorized scope - is managed through a browser-based Greenlist Dashboard accessible to designated Service Canada administrators. Through the dashboard, administrators can add or remove authorized source URLs immediately, review the complete change history, and roll back to any previous version. Changes take effect within seconds. No developer involvement is required for routine scope updates.

When Government of Canada program pages are updated - new policy published, benefit rates revised, eligibility criteria changed - Service Canada administrators flag the change through the feedback process. The greenlist is updated accordingly. Because SCAI retrieves content dynamically from current Government of Canada web pages rather than maintaining a cached corpus, updated pages are reflected in responses immediately after the greenlist is confirmed current. If a source page is removed or its URL changes, the gate will fail to retrieve it and will not generate a response from that source: the client is directed to Service Canada staff rather than receiving stale information.

Governance and the Audit Trail

Every interaction generates a structured governance record capturing: the query received, the gate decision (authorized or denied), which sources were retrieved and parsed, which model version was used, and whether a response was delivered. This record is available to Service Canada administrators and provides a complete, auditable chain of accountability for every AI-mediated interaction.

For denied queries - approximately 40% in a typical institutional deployment - the record documents the gate failure reason and the safe-failure response delivered. For authorized queries, the record traces the response to the specific source pages consulted. At any point, an institutional administrator can review exactly what the system said and why. This is not a probabilistic or approximate audit capability. It is a deterministic record of every decision the system made.

Escalation and Human Oversight

SCAI is a supplementary channel, not a replacement for human service delivery. Every response directs clients to Service Canada staff for further assistance. Planned enhancements include a client-initiated connect-me-to-an-agent option and an automatic human handoff prompt after multiple consecutive exchanges, ensuring that clients who need human support are actively routed to it. A client who receives an incorrect or insufficient response - which is possible within authorized scope, as Section 6 discusses - can contact Service Canada staff immediately through existing channels, where staff can review the interaction using the audit record.

5.6 Additional Institutional Deployments

Beyond the Service Canada federal deployment, COMPAiSS is in active deployment at two Canadian research universities:

- Dalhousie University - student affairs and academic policy queries. Dalhousie conducted a formal institutional evaluation and selected COMPAiSS following a competitive assessment. The deployment covers student wellness and mental health support, academic standing, grade appeals, accommodation procedures, and related high-stakes student services queries.
- McGill University - undergraduate admissions and academic standing queries. The McGill deployment handles complex admissions deadline queries, faculty-specific academic probation conditions, transfer applicant requirements, and Ontario and Quebec high school applicant cutoff information. Testing confirmed that when authorized sources contain the answer, COMPAiSS cites them precisely; when they do not, the system explicitly acknowledges the gap and directs the student to the appropriate office - without hallucinating a plausible answer.

These deployments confirm that the execution-gated architecture generalizes across institutional contexts. The governance gate, greenlist management, audit trail, and safe-failure behavior operate identically whether the deployment serves federal benefit program clients, graduate students navigating academic misconduct procedures, or undergraduate applicants with time-sensitive admissions questions.

The McGill boundary behaviour test
When asked for the supporting documents deadline for transfer applicants from universities outside Canada for Fall 2026, the McGill deployment gave a precise answer where the authorized source confirmed one - April 30, 2026, cited directly from the official McGill Ontario Admissions Cutoffs page with the institution's own caveat that these are historical reference points, not guaranteed minimums. When asked a related question where the authorized source did not specify a deadline, the system said so explicitly and directed the student to the Applicant Portal and Admissions office. No date was fabricated. The boundary between what is authorized and what is not is enforced structurally, not probabilistically.

6. The Case for Regulated Institutional Environments

The standard industry argument for managed hallucination risk - that RAG plus guardrails plus human oversight is sufficient - rests on a tolerable-error assumption: that some residual rate of incorrect outputs is acceptable given the benefits of deployment.

This assumption is defensible for consumer-facing general-purpose AI. It is not defensible for regulated institutional environments where the outputs of AI systems affect citizens' access to benefits, legal rights, healthcare entitlements, or educational standing.

The Government of Quebec's own regulatory guidance illustrates the stakes: citizens who rely on AI-generated information to make financial or healthcare decisions may face serious consequences when that information is incorrect. French courts issued formal judicial warnings in late 2025 about AI-hallucinated legal content reaching judicial proceedings. The US GAO documented that agencies' AI use cases have increased ninefold from 2023 to 2024, with more than 85% of high-impact deployed use cases lacking required risk mitigation documentation.

This argument is not a dismissal of RAG or generation-first architectures. For general-purpose AI - enterprise knowledge search, content generation, broad research tools, customer support across open domains - those architectures serve their intended purposes well. Engineers who have built careful, well-governed RAG systems have done genuinely useful work. The question is not whether those systems are good at what they are designed for. The question is whether they meet the specific governance obligations that regulated institutional contexts impose: obligations to provide authoritative, traceable, and bounded information to citizens whose entitlements, health, or legal standing may depend on what they are told.

In this context, the managed-risk approach to hallucination creates institutional and legal exposure that the institutions themselves cannot easily bear. A government agency that deploys a RAG-based AI system and documents a 17-33% residual hallucination rate - the rate empirically confirmed for the leading enterprise legal AI systems - cannot represent that system as providing accurate and authoritative information to the public.

Execution-gated inference changes this calculus. By eliminating the structural condition under which hallucinations from unauthorized inference occur, COMPAiSS allows institutions to represent their AI systems as bounded, auditable, and governable in ways that generation-first systems cannot match:

- Every response is traceable to specific authorized institutional sources.
- Unauthorized queries produce no AI-generated content - only links to official sources.
- The governance architecture is deterministic and documentable: the institution controls which sources are authoritative, defines the scope boundary, and the system enforces both structurally.
- The audit trail captures every query, every gate decision, and every source consulted.
- No personal data is collected, processed, or retained.

6.1 Addressing Within-Scope Generation Errors

A sophisticated reviewer will correctly observe that execution gating in COMPAiSS prevents out-of-scope hallucinations but does not eliminate all generation errors within the authorized scope. This observation is accurate. However, the nature of within-scope errors is categorically different from the hallucination risk in ungated systems, and the relevant comparison is not between COMPAiSS and a perfect information system.

Within the authorized scope, every COMPAiSS response is drawn exclusively from verified Government of Canada institutional source pages. The pre-inference execution gate prevents any response from being generated unless at least one verified source is available. As a result, the system cannot fabricate unauthorized sources or invent program details. Within-scope errors are therefore primarily navigational or interpretive, not fabricational or decisional. In the worst case, a client may be directed to a less relevant authorized Government of Canada page rather than the most directly applicable one. Such errors are low-impact and readily correctable: clients can refine their query, follow alternative linked sources, or contact Service Canada staff through existing channels. Importantly, SCAI has no decision authority: it does not determine eligibility, calculate benefit amounts, or adjudicate claims, and its outputs do not create or alter client rights.

The worst-case outcome of an incorrect SCAI response is navigational or interpretive, not fabricational or decisional: a client may be directed to a

less relevant authorized Government of Canada page. The worst-case outcome of an ungated AI response is a confident, specific, wrong answer - a hallucinated dollar amount, eligibility rule, or deadline - that a vulnerable client acts on without knowing it was false.

This worst-case profile must be evaluated against the realistic alternative - what occurs routinely when clients cannot obtain immediate answers from Service Canada and turn to general-purpose AI systems such as ChatGPT, Gemini, or Copilot. These systems are not constrained to verified Government of Canada sources. They are empirically and architecturally documented to hallucinate program-specific details, including eligibility rules, benefit amounts, and application deadlines, producing confident and specific but incorrect answers that vulnerable clients may act upon without knowing the information is false. This is not a theoretical worst case. It is a routine consequence of ungated inference applied to institutional questions.

The comparison is therefore not between COMPAiSS and perfection. It is between two distinct failure profiles:

- A system whose worst case is a less relevant authorized link - immediately or readily correctable, carrying no decision authority and introducing no new decision risk; and
- A system whose worst case is a hallucinated dollar amount, eligibility rule, or deadline that a citizen uses to make a consequential financial or administrative decision.

A third consideration reinforces this asymmetry. By resolving routine informational queries accurately and immediately, SCAI frees Service Canada agents and advisors to focus on complex cases, sensitive client situations, and appeals - the interactions where human judgment is not merely preferable but essential. This reallocation increases the probability of correct outcomes across the entire Service Canada client population, not only for clients who use SCAI. The governance benefit of execution-gated inference thus extends beyond the interactions it handles directly to the interactions it enables human staff to handle more effectively.

7. Tradeoffs, Limitations, and Design Choices

Execution-gated inference is an architecture optimized for a specific set of constraints. Like all architectural choices, it involves tradeoffs. Understanding what it sacrifices - and why those sacrifices are acceptable in the target deployment context - is essential to evaluating whether it is the right choice for a given institution.

7.1 Scope Boundary Decisions Require Ongoing Institutional Judgment

The greenlist is the mechanism through which an institution defines what the system may answer. This is a governance strength - the institution retains complete control over scope - but it is also an ongoing operational responsibility. Greenlists must be actively maintained as programs evolve, policies change, and new services are added. A query that falls outside the current greenlist scope will not receive an AI-generated response, even if the answer is available on a Government of Canada page that has not yet been added.

Institutions that deploy COMPAiSS must designate administrators responsible for greenlist maintenance and build processes for flagging scope gaps as they emerge. The Greenlist Dashboard makes this operationally straightforward, but the institutional judgment about what belongs in scope - and what does not - remains a human responsibility that cannot be delegated to the system itself.

7.2 Within-Scope Generation Quality Is Not Guaranteed

Execution gating eliminates scope-violation hallucinations. It does not eliminate generation-quality risks within authorized scope. When the gate passes a query and the system generates a response from authorized source content, the underlying generative model can still misinterpret evidence, misaggregating information from multiple sources, or produce responses that are technically grounded but practically misleading. This is why COMPAiSS employs a defense-in-depth architecture within authorized scope: instruction-based constraints, structured source parsing, and post-generation URL validation all reduce - but do not eliminate - within-scope generation risk.

Institutions that deploy COMPAiSS should maintain a human escalation path, monitor response quality through the audit record, and update greenlist configurations and prompt architecture when systematic quality issues are identified. The system is designed to support this governance loop, not to replace it.

7.3 Bounded Scope Is a Feature, Not a Limitation

The most common objection to execution-gated inference is that it cannot answer questions outside its authorized scope. This is correct. It is also the point. COMPAiSS is not designed to answer any question. It is designed to answer questions that an institution has authorized it to answer, from sources the institution has approved, in ways the institution can audit and defend.

For institutions whose AI governance obligations require exactly that bounded, auditable behavior - and where the consequences of unauthorized answers are borne by citizens - this is not a limitation. It is the design specification. The 40% of queries that are denied inference entirely do not represent system failures. They represent the governance gate working as intended: routing clients to human channels for questions that fall outside the institution's approved information scope.

7.4 Latency Considerations

The pre-inference authorization step adds a small amount of latency relative to unguarded inference. In the Service Canada deployment, this includes a greenlist scoring pass against authorized source URLs and a Google Custom Search Engine query to identify relevant source pages. In practice, the total response time for authorized queries is comparable to other AI assistant deployments because the authorization and retrieval steps occur in parallel with user-perceived processing time. Denied queries return faster than authorized queries, because no inference is executed.

For applications requiring sub-second response times at very high volume, the architecture remains applicable, but deployment configuration - greenlist size, retrieval scope, caching strategy - should be optimized accordingly. This is a configuration question, not an architectural constraint.

7.5 The Core Tradeoff

Why the tradeoff is worthwhile for regulated institutional contexts

Execution-gated inference sacrifices open-domain breadth in exchange for bounded, auditable, governable behavior. For general-purpose AI, that is the wrong tradeoff. For regulated institutional environments where citizens rely on AI-mediated information to make decisions about their benefits, health, legal standing, or educational entitlements, it is the only tradeoff that meets the institution's governance obligations.

8. Comparison with Existing Approaches

Criterion	General-Purpose LLM	RAG System	Guardrails / Filtering	COMPAiSS (Exec-Gated)
Hallucination prevention	None - generates freely	Reduces by 30-50%	Detects/filters post-gen	**Eliminates hallucinations from unauthorized inference**
Residual hallucination rate	High	17-33% (enterprise)	Depends on detector quality	**Zero for out-of-scope**
Governance auditability	Low	Moderate	Moderate	**High - every decision logged**
Infrastructure cost	Low-moderate	$90K-$200K/yr (100K queries)[1]	Adds to RAG cost	**$15K-$30K/yr (100K queries)[1]**
Institutional control	None	Partial	Partial	**Complete - greenlist owned by institution**
Official languages	Translation only	Varies	Varies	**Verified bilingual source routing**
Privacy - personal data	Often retained	Often retained	Depends on system	**Stateless - nothing retained**
Patent status	N/A	N/A	N/A	**CIPO 3,299,174 / USPTO 19/455,963**

[1] Infrastructure cost estimates assume approximately 100,000 AI-assisted queries annually. Conventional RAG figures reflect total annual operating costs including inference compute, persistent vector database infrastructure, and compensatory governance tooling, based on published enterprise RAG deployment benchmarks. COMPAiSS figures reflect inference API costs at current model pricing, greenlist retrieval, and platform hosting, with approximately 40% of queries denied inference entirely at zero marginal compute cost. Actual costs vary with query volume, model selection, and cloud configuration. These figures are indicative and directional, not contractual.

9. Conclusion

The most important insight in this paper is not that hallucination rates can be reduced. It is that the existence of a generative inference runtime itself can become governance-controlled. That shift - from managing what a model produces to governing whether a model runs - is the foundational distinction between hallucination mitigation and hallucination prevention by architectural design.

The persistence of AI hallucination across all current mitigation approaches is not a failure of engineering effort. It is a consequence of a shared architectural assumption: that generative inference runs first, and safety is enforced afterward. Every RAG system, every guardrail, every moderation layer, every human review workflow operates on outputs that the generative model has already produced. Hallucinations arising from unauthorized inference execution are structurally unavoidable under this paradigm. Detection and filtering catch a fraction of them.

Execution-gated inference addresses this at the architectural level. By conditioning the existence of a generative inference runtime on the prior confirmation of authorized institutional sources, COMPAiSS prevents the structural condition under which hallucinations from unauthorized inference execution occur. Within authorized scope, generation-quality risks remain and are managed through defense-in-depth controls. That is the honest and precise claim.

This approach has no prior-art reference identified in the surveyed literature and patent corpus that discloses an equivalent execution-lifecycle architecture. It is under patent examination in Canada and the United States. It has been empirically validated across three Canadian institutional deployments - a federal government service delivery context serving Canadians in both official languages, and two research universities operating across high-stakes student services and admissions domains - and listed on the Government of Canada Artificial Intelligence Source List (CanadaBuys).

For regulated institutional environments - governments, hospitals, legal services, universities - where the consequences of AI-generated misinformation are borne by citizens rather than by systems, the question is not whether execution-gated inference is preferable. The question is whether the governance obligations of those institutions can be met without it.

References

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. Journal of Empirical Legal Studies. doi:10.1111/jels.12413

OpenAI. (2024). Why language models hallucinate. openai.com/index/why-language-models-hallucinate/

National Institute of Standards and Technology. (2023). AI Risk Management Framework (AI RMF 1.0). NIST AI 100-1.

Huang, L., et al. (2024). A Survey on Hallucination in Large Language Models. ACM Computing Surveys.

Shuster, K., et al. (2023). Retrieval Augmentation Reduces Hallucination in Conversation. arXiv:2104.07567.

Harvey, F. (2026). COMPAiSS: Execution-Gated Inference Architecture. Patent applications: CIPO 3,299,174 (Canada); USPTO 19/455,963 (United States).

Gouvernement du Québec. (2025). Risques liés à l'intelligence artificielle. quebec.ca.

US Government Accountability Office. (2025). Artificial Intelligence: Generative AI Use and Management at Federal Agencies. GAO-25-107653.

Prevail AI. (2024). Stanford Study Reveals Challenges of RAG in Legal AI Tools. blog.prevail.ai.

Treasury Board of Canada Secretariat. (2019, amended 2023). Directive on Automated Decision-Making. Canada.ca.