COMPAiSS

Breaking Out of the Matrix

A New Opportunity for AI Consulting Firms in Regulated Industries

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

Abstract

This white paper is addressed to the major consulting firms that advise regulated institutions on AI governance. It documents a cross-model empirical assessment confirming that the leading consulting firms operate within a generation-first paradigm in which hallucinations are treated as inevitable and managed after inference occurs. It argues that COMPAiSS introduces an alternative architectural principle, execution-gated inference, that removes the structural conditions under which out-of-scope hallucinations arise. It does not claim a perfect system. It claims something more precise: that the errors an execution-gated system makes are qualitatively different from those of ungated systems in ways that regulated institutions can govern, audit, and stand behind. Active deployments at Service Canada, Dalhousie University, and McGill University provide empirical validation. The architecture is the subject of patent applications under examination in Canada and the United States.

1. The Dominant Paradigm

1.1 Why the Generation-First Paradigm Persists

The generation-first paradigm is the product of how the AI industry evolved. Large language models were built for breadth: the ability to reason across domains, infer context, and produce fluent responses on virtually any topic. For consumer applications, content generation, internal productivity, and open-domain research assistance, this capability is genuinely valuable.

Enterprise AI consulting developed within this paradigm. The standard advisory approach connects a general-purpose model to organizational data through RAG, adds governance layers, builds compliance dashboards, and implements human review workflows. This is a coherent and

defensible response to a real challenge, and the practices built around it reflect substantial investment and expertise.

The generation-first paradigm also has structural economic logic. Because inference is always permitted to run, institutions must continuously invest in the compensatory infrastructure that manages its outputs. Moderation tooling, safety layers, audit systems, and human review workflows are genuine operational requirements, not invented complexities. They are the necessary cost of managing risk in a system where thinking always happens first. Consulting firms that built practices around this infrastructure did so because the infrastructure was, and remains, necessary for generation-first deployments.

This paper does not argue that those practices were wrongly built. It argues that for a specific class of institution, in a specific deployment context, a different architectural starting point is available, and that consulting firms are well positioned to lead their clients toward it.

1.2 What Four Independent AI Models Found

COMPAiSS conducted a cross-model assessment of the publicly available AI governance recommendations delivered by major consulting firms to regulated institutions. Four independent large language models, GPT, Gemini, Copilot, and Perplexity, were each provided with an identical prompt focused on how these firms conceptualize and control AI hallucinations. Each model produced a standalone assessment. The four assessments were then reviewed for consensus and divergence.

The methodological design is worth stating clearly. Each model was provided the same tightly scoped prompt and restricted to publicly available product documentation, governance frameworks, and architectural descriptions. No model was given the others' outputs during the assessment phase. The consensus finding therefore reflects independent convergence, not circular reasoning.

> **Cross-Model Consensus Finding (GPT, Gemini, Copilot, Perplexity)**
>
> All four models independently concluded that the major consulting firms operate under a generation-first, post-hoc mitigation paradigm. Hallucinations are treated as inherent and probabilistic properties of large language models, to be managed through governance, RAG, and human oversight after inference has already occurred.
>
> No firm was found to advocate for pre-inference execution denial or structural prevention of hallucination. In every framework examined, refusal is itself an output of generative execution. The model ran, and produced a refusal. That is architecturally distinct from a system in which the model never ran.

Firm	What the Framework Addresses	What the Cross-Model Assessment Found
Deloitte	Governance frameworks, QA, and post-generation human review	Hallucinations treated as inevitable reliability risk; no structural prevention mechanism identified across any product or framework
PwC	RAG and prompt discipline to reduce plausible-but-baseless outputs	Generation always occurs first; mitigation is applied during and after, not before

KPMG	Auditability, monitoring, and post-generation governance controls	Hallucinations treated as endemic; addressed reactively through oversight frameworks
Microsoft	Grounding, correction loops, and safety platform tooling (Bedrock Guardrails, Azure AI Content Safety)	Hallucinations defined as ungrounded content; reduced through grounding, not prevented; inference always runs

> *GPT finds no evidence of pre-inference execution denial or structural prevention across any firm. Gemini explicitly states that none of the four firms challenge the generation-first assumption. Copilot finds no advocacy of pre-generation blocking; refusal is treated as a configured behavior, not a default architecture. Perplexity explicitly notes that RAG is treated as a compensatory mechanism, not a hard gate preventing inference. Source: COMPAiSS, Hallucinations by Design: A Cross-Model Assessment, 2026*

1.3 The Structural Hallucination Problem

The persistence of 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.

RAG is the dominant consulting recommendation for regulated institutions. The intuition behind it is sound: if the AI only reads approved documents, hallucinations can be controlled. The limitation is that RAG controls what the AI reads, not whether or how the AI is permitted to reason. Even with a well-curated retrieval corpus, the underlying generative model retains the capacity to extrapolate from parametric memory, misinterpret retrieved content, and produce plausible outputs that are not grounded in what was retrieved.

Documented Hallucination Rates in Enterprise AI Systems (2023 to 2025)

- 17 to 34%: Hallucination rate for Westlaw AI-Assisted Research and Lexis+ AI on real-world legal queries. These are purpose-built RAG platforms marketed as hallucination-free by Thomson Reuters and LexisNexis. (Magesh et al., Stanford RegLab / HAI, 2024)
- 50 to 83%: Adversarial hallucination rates across six major AI models in clinical vignette testing using physician-validated scenarios. (Omar et al., Nature Medicine / Communications Medicine, 2025)
- 33 to 51%: Hallucination rates for OpenAI o3 reasoning models on open-domain factual benchmarks, higher than prior-generation models despite greater capability. (Graffius, 2026)
- 6%: Residual hallucination rate under optimal RAG conditions in tightly constrained grounded summarization tasks. This is the best-case scenario under controlled conditions. (Nishisako, Higashi and Wakao, 2025)

The central finding across 2023 to 2025 research: hallucination is not a temporary limitation that improving models will eliminate. It is a persistent, structural property of inference-first architectures. More capable models do not automatically hallucinate less. On complex reasoning tasks, some hallucinate more.

2. The Architectural Alternative

2.1 Execution-Gated Inference

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

Architectural Comparison

Generation-First (All Current Enterprise Systems):

Request received > Inference runtime instantiated > Generative model executes > Output produced > Safety / filtering / moderation applied > Response delivered

COMPAiSS: Execution-Gated Inference:

Request received > Authorization gate evaluates institutional sources > IF authorized: Inference runtime instantiated > Response generated from approved sources > Response 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, or a moderation layer 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 because no institutional source supports a response, no generative computation occurs. There is no unauthorized inference from which out-of-scope hallucinations can arise.

This distinction is not semantic. In generation-first systems: *the model exists and authorization governs what it produces.* In execution-gated systems: *authorization determines whether the model exists.* The model only runs when authorized sources confirm it may.

Within authorized scope, COMPAiSS employs a defense-in-depth architecture. Instruction-based constraints guide model behavior during authorized inference. Post-generation URL validation ensures all links point to authorized institutional sources. These controls reduce within-scope generation risk but do not eliminate it. Section 2.2 addresses that directly.

2.2 What COMPAiSS Does Not Claim

Some reviewers of an earlier draft of this paper noted that certain phrasings implied execution gating eliminates hallucination broadly. That is not the claim, and the distinction matters for a governance audience.

COMPAiSS makes a precise claim, drawn from its own technical documentation:

> *"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."*
> *Source: COMPAiSS Technical Documentation*

Within authorized scope, the underlying generative model can still misinterpret evidence, mis-aggregate information across sources, or produce responses that are technically grounded but

practically misleading. These residual risks are real and acknowledged. They are also categorically different from the hallucination risks in ungated systems, and that difference determines whether regulated institutions can govern and stand behind their AI deployments.

The relevant question is never whether a system is perfect. No AI system is. The relevant question is: what kind of errors does this system make, and can the institution detect, govern, and stand behind them? The table below addresses that question directly.

2.3 Comparing Failure Profiles

The case for execution-gated inference does not rest on a claim that COMPAiSS never makes mistakes. It rests on a claim that the mistakes it makes are qualitatively different from those in ungated systems in ways that matter specifically for regulated institutional governance.

Dimension	COMPAiSS: Within-Scope Errors (Residual)	Generation-First AI: Hallucination Risk (Structural)
Error class	Navigational or interpretive: client may be directed to a less relevant authorized institutional page	Fabricational or decisional: hallucinated eligibility rule, benefit amount, deadline, or policy detail presented with full confidence
Source grounding	Every response traces to at least one verified, institution-approved source URL	Response may cite no source, a fabricated source, or a real source that does not support the claim made
Decision authority	None: the system provides information only and does not determine eligibility, calculate benefits, or adjudicate claims	None formally, but confident wrong answers are routinely acted upon as authoritative by users
Correctability	Immediately correctable: client can refine query, follow linked source, or contact staff	Often not recognized as incorrect by the user; error may surface only after consequential action is taken
Frequency under RAG (enterprise)	Within-scope navigational errors: not independently benchmarked; audit record enables detection	17 to 34% on real-world queries in purpose-built legal RAG platforms (Magesh et al., Stanford/Yale, 2025)
Worst-case consequence	User contacts staff for clarification; no change to rights or entitlements	Vulnerable citizen acts on hallucinated benefit amount, missed deadline, or false eligibility rule with potentially irreversible consequences

Note: Within-scope navigational errors in COMPAiSS are not independently benchmarked against a public dataset. The audit record available to institutional administrators enables detection and correction. The generation-first hallucination rates cited reflect independently published peer-reviewed research on enterprise systems in regulated domains.

The asymmetry in the table above is the governance argument. The errors COMPAiSS produces are bounded, source-traceable, correctable, and carry no decision authority. The errors that generation-first systems produce in regulated environments are documented at 17 to 34% in purpose-built enterprise platforms, structurally inevitable, and in the worst case are acted upon by vulnerable people who have no way to know the information they received was fabricated.

> *"The worst-case outcome of an incorrect COMPAiSS response is navigational or interpretive, not fabricational or decisional: a client may be directed to a less relevant authorized 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."*
> *Source: COMPAiSS White Paper, Execution-Gated Inference, 2026*

2.4 Two Entry Points

COMPAiSS operates at two natural entry points within an institution's AI journey.

For institutions already invested in RAG: COMPAiSS operates as a pre-inference governance layer that addresses the residual hallucination rates RAG alone cannot resolve. This is not a recommendation to replace existing infrastructure. It is a governance control that closes the structural gap the generation-first architecture leaves open. Consulting firms can position this as a governance maturity upgrade, helping institutions that have invested in RAG infrastructure reach the accountability threshold that regulated environments require.

For institutions without RAG investment: Whether cost-constrained or troubled by the documented failure rates under optimal RAG conditions, these institutions have a direct alternative. COMPAiSS provides a complete governed AI architecture that enforces scope, source authority, and institutional policy without requiring enterprise infrastructure investment. For these clients, consulting firms can position COMPAiSS as the governance-first path, a way to achieve institutional AI deployment without entering the remediation cycle at all.

2.5 Cost Structure

The financial case follows directly from the architecture. Because approximately 40% of queries are denied inference entirely at zero marginal compute cost, and because authorized queries do not insert retrieved documents into the model context, COMPAiSS consumes materially fewer compute resources than conventional RAG deployments.

Cost Component	Conventional RAG	COMPAiSS
Inference (tokens)	$10k to $30k	$7k to $18k
RAG infrastructure	$30k to $60k	$0
Safety and moderation tooling	$30k to $75k	$0
Monitoring and compliance	$10k to $20k	$5k to $15k
TOTAL ANNUAL COST	**$90k to $200k**	**$15k to $30k**

Assumptions: approximately 100,000 annual queries, approximately 60% authorization rate. 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. Actual costs vary with query volume, model selection, and cloud configuration. These figures are indicative and directional, not contractual.

The 6x to 12x cost differential is not achieved through reduced capability. It is a structural consequence of the execution-gated architecture: cost is avoided because unauthorized queries do not generate any AI output at all.

2.6 Deployed and Validated

> **Active Institutional Deployments and Beta Testing (May 2026)**

Service Canada (Federal Government)

AI Information Assistant covering Employment Insurance, Canada Pension Plan, Old Age Security, Guaranteed Income Supplement, and Social Insurance Number services. Fully bilingual with French-language source routing confirmed. Multilingual query support tested including Tagalog, with accurate current-year benefit rates retrieved from official Government of Canada sources. WCAG 2.1 AA accessibility audit Phase 1 completed; phase two in progress. Listed on the Government of Canada Artificial Intelligence Source List (CanadaBuys).

Dalhousie University

Student affairs and academic policy queries covering student wellness, academic standing, grade appeals, and accommodation procedures. Dalhousie conducted a formal institutional evaluation and selected COMPAiSS following a competitive assessment that included an established generation-first higher education AI vendor.

McGill University

Undergraduate admissions and academic standing queries. Boundary testing confirmed the system cites sources precisely when authorized sources contain an answer, and acknowledges gaps explicitly when they do not, directing students to the appropriate office rather than producing a plausible fabricated response. No date was generated where no date existed in the authorized source. Currently in beta testing.

3. The Consulting Firm Advantage

3.1 The Advisory Opportunity

Regulated institutions are under increasing governance pressure from several directions simultaneously. The US GAO documented that federal agencies' AI use cases increased ninefold from 2023 to 2024, with more than 85% of high-impact deployed use cases lacking required risk mitigation documentation. French courts issued judicial warnings about AI-hallucinated legal content reaching judicial proceedings in late 2025. The Government of Quebec issued formal regulatory guidance warning of consequences for citizens who rely on AI-generated information for financial or healthcare decisions. NIST's AI Risk Management Framework requires documentation of residual risk, recognizing that not all incidents and failures can be eliminated under current architectures.

Institutions responding to this pressure need advisors who can offer more than another governance layer applied to a generation-first system. The consulting firm that can explain execution-gated inference to a hospital board, a university senate, or a government CIO, and help that institution assess whether its governance obligations are compatible with a residual hallucination rate, is offering something qualitatively different from what is currently available.

3.2 Three Advisory Service Lines

Advisory Service Opportunities

1. Governance Architecture Assessment

Help regulated institutions assess whether their current or planned AI architecture meets their governance obligations. The central question is not which product to buy. It is whether the institution's accountability framework is compatible with a residual hallucination rate, and if not, what architectural alternative meets their obligations.

2. Implementation and Scope Design

For institutions that adopt execution-gated inference, the implementation engagement is substantive: institutional scope definition, greenlist architecture, integration with existing systems, and staff readiness. This implementation work benefits from advisory experience in regulated environments in ways a technology vendor alone cannot replicate.

3. Ongoing Governance Advisory

Execution-gated AI requires ongoing institutional judgment: greenlist maintenance as policies evolve, scope boundary decisions as new programs are added, audit record review, and governance reporting. This advisory layer is tied directly to the institution's governance obligations rather than to a specific product.

3.3 Addressing the Internal Objection

The candid internal objection at any consulting firm reading this paper is straightforward: if execution-gated inference is adopted at scale, the remediation advisory practice shrinks. That is a fair observation and deserves a direct response.

The remediation model faces its own headwinds. As hallucination failures become more public and more legally consequential, the consulting firm associated with generation-first governance faces reputational exposure that a firm advising on governance-first alternatives does not. The institutions that are already making headlines for AI governance failures deployed generation-first systems with standard advisory support. The question is not whether the paradigm will eventually shift under regulatory and institutional pressure. The question is whether a firm leads that shift or responds to it after a competitor has.

Many consulting partners may find it most comfortable to position execution-gated inference initially as a governance-first control architecture that can sit alongside or upstream of generation-first systems, rather than as a replacement. That framing is accurate, commercially sensible, and supported by the two-entry-point structure described in Section 2.4.

3.4 The Regulatory Tailwind

- The US GAO documented that more than 85% of high-impact federal AI use cases lacked required risk mitigation documentation as of 2024, representing both a compliance gap and an advisory opportunity.
- The EU AI Act classifies AI systems used in education, employment, healthcare, and critical public services as high-risk, with requirements for human oversight, auditability, and transparency that generation-first systems with residual hallucination rates are challenged to satisfy.
- Canada's Directive on Automated Decision-Making creates accountability obligations for federal agencies that are most cleanly satisfied by systems whose outputs are deterministically traceable to authorized sources.
- The Government of Quebec's AI risk guidance explicitly warns that citizens who rely on AI-generated information for financial or healthcare decisions may face serious consequences when that information is incorrect.

Each of these developments creates a governance gap that regulated institutions need help closing. The consulting firm that can offer an architectural response, rather than another framework for managing inevitable risk, is the firm best positioned to capture that mandate.

3.5 A Note on Tone

This paper uses the Matrix and Road Less Travelled analogies because they are the author's own published framing and because they communicate a genuine architectural distinction efficiently. But it is worth being explicit about what those analogies are and are not.

They are not arguments that consulting firms have been doing AI governance wrong. The generation-first paradigm is rational for the use cases it was designed for. They are not arguments that COMPAiSS is the only defensible governance architecture. For general-purpose AI, enterprise knowledge search, and open-domain applications, generation-first systems serve their purposes well.

They are arguments that for a specific class of institution, in a specific deployment context, where citizens rely on AI-mediated information to make decisions about their benefits, health, legal standing, or educational entitlements, the architectural starting point matters in ways that no downstream remediation can fully address. That is a narrow, precise, and empirically supported claim. It is the claim this paper makes.

Conclusion

The most important architectural 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 structural prevention of out-of-scope hallucination.

COMPAiSS does not produce a perfect system. Within authorized scope, generation-quality risks remain and are managed through defense-in-depth controls. The honest and precise claim is this: execution gating removes the structural conditions under which out-of-scope hallucinations arise, and the errors that remain within authorized scope are qualitatively different from those in ungated systems in ways that regulated institutions can govern, audit, and stand behind.

For consulting firms advising regulated institutions, the practical implication is clear. Institutions are asking, with increasing urgency, whether their AI governance obligations are compatible with a residual hallucination rate. The firm that can answer that question architecturally, rather than with another layer of remediation, is offering something the current market does not yet provide at scale.

COMPAiSS is in active deployment at a federal government service delivery agency, a research university that selected it over an established generation-first competitor, and a second research university operating across high-stakes student services and admissions domains. The architecture is under patent examination in Canada and the United States. The evidence base is peer-reviewed and publicly documented.

The road less travelled is not a critique of the main road. It is an observation that for some institutions, with specific governance obligations, a different road is available. The consulting firms that can explain the difference, and guide the institutions for whom it matters toward the right choice, will lead the next chapter of AI governance advisory in regulated industries.

For further information or to discuss a consulting partnership:

Frank P. Harvey, PhD | frank.harvey@dal.ca | compaiss.ca

References

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., and 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

Omar, M., et al. (2025). Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Communications Medicine (Nature Publishing Group). PMCID: PMC12318031.

Graffius, S.M. (2026). Are AI Hallucinations Getting Better or Worse? We Analyzed the Data. DOI: 10.13140/RG.2.2.33179.53285.

Nishisako, T., Higashi, T., and Wakao, R. (2025). Hallucination rates in constrained RAG summarization. Cited in COMPAiSS White Paper, 2026.

Sun, et al. (2025). ReDeEP: Detecting Hallucination in LLMs via Decoupled Representation of Retrieval and Encoding. ICLR 2025.

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

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

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

Gouvernement du Quebec. (2025). Risques lies a l'intelligence artificielle. quebec.ca.

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

COMPAiSS Inc. (2026). Hallucinations by Design: A Cross-Model Assessment. compaiss.ca.

COMPAiSS Inc. (2026). The Road Less Travelled: Rethinking Generative AI Costs, Safety, and Trust in Regulated Institutions. compaiss.ca.

COMPAiSS Inc. (2026). Cost Savings by Design: Why COMPAiSS Costs Less Without Sacrificing Accuracy. compaiss.ca.

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