Authorizing Evidence Before Inference

Positioning COMPAiSS Within the Emerging Landscape of Governance-Oriented AI Architectures

DRAFT — Positioning and Differentiation Paper

[Author name] · COMPAiSS · July 2026 · compaiss.ca

Abstract

Institutions increasingly face a structural problem: the public asks artificial-intelligence systems questions about them, and nothing constrains what those systems use as evidence when answering. A growing family of architectures responds by inserting governance into the AI execution pipeline rather than relying on model behaviour alone. This paper positions COMPAiSS — a deployed institutional assistant in which the deploying institution authorizes the evidence sources an answer may be grounded in before inference begins — within that family. We organize the landscape along two axes: when control is applied (training time, before inference, during generation, or after generation) and what is authorized (values, content categories, dialogue flows, actions and capabilities, or evidence). Against the nearest architectural neighbours — constitutional training methods, input/output safeguard classifiers, programmable runtime guardrails, and capability-based control-flow designs such as CaMeL — COMPAiSS occupies a distinct position defined by a specific combination of properties: to our knowledge, it is the deployed implementation in which the object of pre-inference authorization is institutional evidence, the authorizing party is the institution itself, enforcement is deterministic and fail-closed, and answers preserve per-claim traceability. We do not claim to be the only effort moving governance upstream; convergent terminology has appeared independently in recent self-published technical reports. Rather, we characterize COMPAiSS as one implementation of an authorization-first architectural approach, distinguished by institution-authorized evidence before inference, deterministic execution gating with fail-closed governed refusal, per-claim traceability, and a body of comparative deployment evaluations.

1. Introduction and Purpose

General-purpose AI assistants now answer institution-specific questions — hospital preparation instructions, benefit eligibility rules, academic policies — for millions of people who never visit the institution's own website. The answers are frequently fluent, frequently useful, and, a material fraction of the time, wrong in ways the institution never sees and cannot correct. The institutional risk is not abstract: a fabricated appeal body, a wrong entrance, or an outdated eligibility rule carries operational and legal consequences that land on the institution being described, not on the system doing the describing.

A rapidly developing body of work responds to this class of problem by adding control layers to AI systems. These efforts differ widely in where control is applied and what, exactly, is being controlled. The purpose of this paper is to position one deployed system — COMPAiSS — within that landscape precisely: to state its architectural claim in testable terms, to identify its nearest neighbours honestly, and to specify what distinguishes it without overclaiming novelty.

The positioning claim of this paper is deliberately modest in form and specific in content: ***COMPAiSS is one implementation of an authorization-first architectural approach, distinguished by its emphasis on institution-authorized evidence before inference, deterministic execution gating, and evidence-preservation methodology.***

2. The Architectural Claim, Stated Precisely

COMPAiSS is built on a single organizing principle: the institution determines what the AI is authorized to use as evidence before inference begins, and inference occurs only within that authorized evidence boundary. Concretely, the architecture has five load-bearing properties:

- Evidence authorization precedes inference. An execution gate checks, before any generation occurs, whether authorized institutional evidence exists for the question. The gate is a structural property of the pipeline, not an instruction to the model.
- The institution is the authorizing party. Authorized sources are defined by a greenlist the institution controls. Governance authority remains with the deployed-to organization, not the model developer or platform vendor.
- Inference is confined to authorized evidence. The model synthesizes and reasons — it is not a retrieval index — but the material it may reason over is bounded ex ante. Training knowledge cannot substitute for missing institutional evidence.
- Failure is fail-closed and governed. Where authorized evidence is insufficient, the system declares the boundary and routes to a human channel rather than generating from general knowledge. Refusal is a designed, legitimate outcome.
- Answers are traceable per claim. Every institutional claim in an answer cites the authorized source it derives from, making errors identifiable and correctable at the source.

This claim is narrower than a general appeal to "governance-first" design, and it is testable: for any given answer, one can ask whether every institutional claim traces to a source on the institution's greenlist, and whether the system declines when no such source exists. Comparative evaluations applying exactly this test are summarized in Section 6.

3. A Taxonomy of Control Points

Governance-oriented AI architectures are best distinguished along two axes. The first is timing: when, relative to inference, does control apply? Four positions recur in the literature and in deployed systems: (i) training time, where control is baked into model weights; (ii) before inference, where a condition must be satisfied for generation to proceed at all; (iii) during generation, where rails shape the dialogue, topics, or tool calls as they happen; and (iv) after generation, where outputs are classified, filtered, or corrected.

The second axis — less often made explicit, and in our view the more consequential one — is the object of authorization: what is the thing being permitted or denied? Candidate objects include the model's values and dispositions; content categories (safety taxonomies); dialogue flows and topics; tools, actions, capabilities, and data flows; and evidence — the knowledge an answer is permitted to be grounded in. Two systems can both claim "authorization before execution" while authorizing entirely different things,

with entirely different institutional consequences. The analysis that follows applies both axes to the nearest neighbours identified in a verified prior-art review.

4. The Architectural Neighbours

4.1 Training-time value alignment: Constitutional AI

Constitutional AI (Bai et al., 2022) trains models against an explicit set of principles, using AI feedback to shape harmlessness without extensive human labelling. It is governance located in the weights: the constitution influences the model's disposition on every subsequent inference. Its strengths are generality and scale; its structural limitation, for institutional purposes, is that disposition is not enforcement. A constitutionally trained model still answers institution-specific questions from whatever its training distribution contains, and no institution outside the model developer holds the authoring pen. Constitutional AI and evidence authorization are complementary rather than competing: COMPAiSS can operate over constitutionally trained models while adding the institutional evidence boundary those models lack.

4.2 Boundary classification: Llama Guard

Llama Guard (Inan et al., 2023) is an LLM-based input/output safeguard: a classifier applied at the conversational boundary to detect content that violates a safety taxonomy. It exemplifies the post-hoc and boundary-filtering family — governance as inspection of what enters and leaves. Two properties separate it from evidence authorization. First, its object is content category (violence, self-harm, and so on), not evidential provenance; a fluent, policy-clean answer containing a fabricated institutional fact passes untouched. Second, it is itself a model, and therefore probabilistic: its judgments are subject to the same distributional behaviour it is deployed to police.

4.3 Programmable runtime rails: NeMo Guardrails and runtime governance

NVIDIA's NeMo Guardrails (2023–present) and its more recent runtime-governance extensions provide programmable rails around LLM applications: topic control, input/output checks, retrieval constraints, and tool-invocation control, expressed in an application-level configuration. This family moves governance meaningfully upstream of pure output filtering and gives application developers real control surfaces — including retrieval rails that constrain which documents enter context. It is the closest widely deployed toolkit family to COMPAiSS in spirit. The differences are locus and default: guardrails are a toolkit a developer may configure (including permissively), enforcement is a mixture of deterministic rules and model-based checks, and the framework is agnostic about who holds authority. COMPAiSS is a purpose-built institutional architecture in which evidence authorization is the non-negotiable default, the institution is the defined authority, and refusal on missing evidence is a designed outcome rather than a configuration choice.

4.4 Capability-based control flow: CaMeL — the closest architectural neighbour

CaMeL (Debenedetti et al., 2025) is, in our assessment and in the prior-art review, the closest existing architecture. Responding to prompt-injection attacks on LLM agents, CaMeL separates control flow from

data flow and enforces capability-based authorization: what an agent may do, and what data may flow where, is determined by explicit security policy before execution, not by the model's judgment. It shares with COMPAiSS the two deepest commitments — authorization before execution, and deterministic enforcement located outside the model.

The distinction is the object of authorization. CaMeL authorizes actions and data flows; its threat model is the injected instruction that makes an agent do something unauthorized. COMPAiSS authorizes evidence; its threat model is the fluent answer grounded in something the institution never said. Both remove a critical decision from the model's discretion, but they remove different decisions. A hospital deploying an agent needs CaMeL-style capability control so the agent cannot act wrongly; the same hospital deploying a public answer service needs evidence control so the answer cannot claim wrongly. The two are architecturally compatible, and a mature institutional stack plausibly contains both.

4.5 Convergent terminology: the pre-inference governance reports

A series of self-published technical reports (Akarkach, 2026a, 2026b, 2026c; Zenodo, non-peer-reviewed) defines "Pre-Inference Governance" as a paradigm in which inference is a conditional capability requiring prior authorization, with fail-closed non-execution as a legitimate state. We cite these reports for one specific reason: they document that the vocabulary of authorization-before-inference has appeared independently in multiple places, which supports the view that governance is moving upstream as a field-wide direction. Two boundaries govern the citation. First, the reports are conceptual rather than operational: by their own description they define a principle and provide no implementation, deployed architecture, or evaluation. Second, COMPAiSS was developed and deployed independently of this work and is described throughout this paper in its own established vocabulary — evidence authorization, execution gate, governed refusal. The comparison is included only to document convergence in vocabulary; the architectural substance on which COMPAiSS rests — a concrete, deployed mechanism in which the authorized object is institutional evidence — is established by the properties and evaluations of Sections 2 and 6, not by this convergence.

5. Comparative Matrix

The matrix below applies the two axes of Section 3 to the verified neighbours. Cells describe each approach as documented in its primary sources; the pre-inference governance reports are included as a stated principle rather than an implemented system, per their own self-description.

Approach	Control point (timing)	Object of authorization	Enforcement character	Primary failure mode addressed	Locus of authority
Constitutional AI (Bai et al., 2022)	Training time	Model values / behavioural disposition	Probabilistic (learned)	Harmful or unethical outputs	Model developer
Llama Guard (Inan et al., 2023)	Around inference (input/output)	Content categories (safety taxonomy)	Probabilistic (classifier)	Unsafe content passing the boundary	Model/platform developer
NeMo Guardrails / Relay (NVIDIA, 2023–2026)	Runtime, around generation and tool calls	Dialogue flows, topics, tool invocation	Programmable rails; mixed deterministic and model-based	Off-topic, unsafe, or unauthorized tool behaviour	Application developer

CaMeL (Debenedetti et al., 2025)	Before execution (control/data flow)	Actions, capabilities, and data flows of an agent	Deterministic (capability-based)	Prompt injection; unauthorized actions	System designer (security policy)
Pre-Inference Governance reports (Akarkach, 2026)	Before inference (as stated principle)	Legitimacy of computation itself (abstract)	Definitional; explicitly non-operational	Unauthorized inference as such	External accountable authority (abstract)
COMPAiSS	Before inference	Evidence — the institutional sources permitted to ground an answer	Deterministic execution gate; fail-closed governed refusal	Institutional misattribution and hallucinated institutional fact	The deploying institution (greenlist)

Table 1. Positioning by control point, object of authorization, enforcement character, failure mode, and locus of authority. Sources: primary documents listed in References; characterizations verified against live records in July 2026.

6. What Distinguishes COMPAiSS

Within this landscape, COMPAiSS is distinguished by the conjunction of five properties. None is individually unprecedented in concept; to our knowledge, their combination in a deployed, evaluated system is.

- Evidence as the object of authorization. Where the nearest deterministic neighbour (CaMeL) authorizes actions and data flows, COMPAiSS authorizes the evidential basis of answers — the property institutions are actually accountable for when their name is on the answer.
- The institution as the authorizing party. The greenlist places governance authority with the deployed-to organization. This is a locus-of-control distinction from both model-level approaches (developer authority) and guardrail toolkits (application-developer discretion).
- Deterministic, fail-closed enforcement. The execution gate is pipeline structure, not prompt content or classifier judgment. Where authorized evidence is absent, governed refusal with human routing is the designed outcome — in deployment testing, the system explicitly declares “not in provided sources” rather than importing outside material.
- Per-claim evidence preservation. Every institutional claim cites its authorized source, producing answers that are auditable and correctable at the level of individual statements rather than whole responses.
- Deployment and evaluation history. The architecture is operational across multiple institutional pilots (university deployments; a federal-service pilot) and has been subjected to blind comparative evaluations against leading generation-first systems using a published criteria matrix, including a hallucination/scope-intrusion penalty metric, and to repeated-run tests demonstrating that its zero-unauthorized-source behaviour is invariant across runs where generation-first systems vary.

The evaluation evidence deserves one framing note. In comparative testing, generation-first systems sometimes performed well on individual runs — including runs with zero unauthorized sources. This is not a counterexample to the architectural claim; it is the claim. Good behaviour that nothing enforces is a disposition, observable per run; behaviour an execution gate enforces is a property, certifiable across runs, model updates, prompts, and users. In generation-first systems, governance is a prompt. In

governance-first systems, it is an execution constraint. Institutions can audit constraints; they cannot audit dispositions.

7. Limitations and Open Questions

Honest positioning requires stating what this paper does not establish. First, scope: COMPAiSS as deployed governs public institutional information; it does not address clinical or personal-data reasoning, agentic action authorization (CaMeL's territory), or training-time alignment, and it presupposes that the institution's authorized sources are themselves accurate and maintained — the architecture guarantees provenance, not the correctness of the provenance's content. Second, evaluation scale: the comparative evaluations to date are structured but small (tens of questions, small numbers of repeated runs), conducted with a published matrix and blind scoring but not yet independently replicated or peer reviewed. Third, the field itself: nearly all of the closest literature — including CaMeL, Llama Guard, Constitutional AI, and the convergent terminology reports — is preprint, vendor documentation, or self-published; the peer-reviewed core of this area is still forming, and positions taken here should be revisited as it does. Fourth, boundary trade-offs: confining inference to authorized evidence produces thinner answers where institutional coverage is thin; we treat this as the correct trade-off for accountable institutions, but it is a trade-off, and measuring its cost to answer completeness is ongoing work.

8. Conclusion

The direction of the field is no longer in serious dispute: control is moving upstream, from filtering what models say toward governing what they may do — and, in the position developed here, what they may know an answer from. Within that movement, COMPAiSS occupies a specific and testable position: it is one implementation of an authorization-first architectural approach, distinguished by institution-authorized evidence before inference, deterministic execution gating with fail-closed governed refusal, per-claim evidence preservation, and an accumulating base of deployment evaluations. The distinction that matters to accountable institutions is not which underlying model is most capable, but which system property they can certify. An execution constraint on evidence is such a property. That is the position COMPAiSS stakes within the emerging literature, and the standard against which it invites evaluation.

References

Peer-review status is indicated for each entry. All records were verified as live and existing in July 2026; one additional item circulating in earlier drafts (an arXiv-hosted "AI governance control stack" PDF) is excluded here because stable bibliographic metadata could not be verified.

Akarkach, M. (2026a). Pre-inference governance – Canonical definition (Version 1.1) [Technical report, not peer reviewed]. Zenodo. https://doi.org/10.5281/zenodo.18401330

Akarkach, M. (2026b). Pre-inference governance beyond medicine: Authorization architectures for high-risk autonomous systems (Version 1.0) [Technical report, not peer reviewed]. Zenodo. https://doi.org/10.5281/zenodo.18261708

Akarkach, M. (2026c). Preventing epistemic exposure in frontier artificial intelligence systems: Pre-inference authorization as an architectural solution to inference extraction and discovery risk [Technical report, not peer reviewed]. Zenodo. https://doi.org/10.5281/zenodo.18795429

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., … Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback [Preprint, not peer reviewed]. arXiv. https://doi.org/10.48550/arXiv.2212.08073

Debenedetti, E., Shumailov, I., Fan, T., Hayes, J., Carlini, N., Fabian, D., Kern, C., Shi, C., Terzis, A., & Tramèr, F. (2025). Defeating prompt injections by design [Preprint, not peer reviewed]. arXiv. https://doi.org/10.48550/arXiv.2503.18813

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., … Khabsa, M. (2023). Llama Guard: LLM-based input–output safeguard for human–AI conversations [Preprint, not peer reviewed]. arXiv. https://doi.org/10.48550/arXiv.2312.06674

NVIDIA Corporation. (2023–2026). NeMo Guardrails: Programmable guardrails for LLM applications; NeMo Guardrails plugin for NeMo Relay: Runtime governance for LLM and tool execution [Technical documentation, not peer reviewed]. https://docs.nvidia.com/nemo/guardrails/ ; https://github.com/NVIDIA-NeMo/Guardrails

COMPAiSS. (2026). Comparative evaluation results: Service Canada Q&A, blind protocol, criteria matrix v3; and repeated-run governance-compliance tests [Internal evaluation reports]. Available on request; summaries at https://compaiss.ai/demos