Cost Savings by Design

Why COMPAiSS Costs Less Without Sacrificing Accuracy

EXECUTIVE SUMMARY

For institutions managing approximately 100,000 AI-assisted queries annually, conventional enterprise AI systems typically incur operating costs between $90,000 and $200,000. A COMPAiSS deployment addressing the same workload operates at approximately $15,000 to $30,000 per year.

This 6× to 12× cost differential is not achieved through discounted pricing, reduced capability, or lower-quality models. It results from a structural architectural inversion. Conventional systems allow AI inference to execute on every query — incurring cost even for irrelevant, unsupported, or refused requests. COMPAiSS introduces an execution gate that prevents model inference unless an authorized institutional answer exists.

As a result, COMPAiSS allocates budget exclusively to authoritative, value-adding interactions, while structurally eliminating cost, risk, and governance overhead associated with unauthorized AI reasoning.

> This analysis addresses operating cost and governance implications of an execution-gated architecture at institutional scale. It does not cover capital expenditures for non-AI infrastructure, change-management costs, or downstream productivity effects. All figures are operating-cost estimates under the assumptions stated in Section 5.

INSTITUTIONAL CHALLENGE: THE SCALING GAP IN AI

Universities, hospitals, public-sector organizations, and other highly regulated institutions manage large, fragmented bodies of authoritative information — policies, procedures, eligibility rules, and regulatory obligations distributed across hundreds or thousands of documents.

Demand for this information continues to rise:

- Students, staff, patients, and citizens expect immediate answers
- Human support teams cannot scale proportionally
- Traditional service models strain budgets and staffing capacity

AI systems are increasingly deployed to absorb this demand. However, how AI is architected determines whether it reduces institutional cost and risk — or amplifies both.

THE CONVENTIONAL PARADIGM: HIDDEN COSTS OF GENERATION-FIRST AI

Most enterprise AI systems follow a generation-first sequence:

1	Inference begins immediately when a question is submitted
2	Documents are retrieved (often via RAG pipelines)
3	A response is generated

4	Safety, moderation, and governance controls intervene post-generation
5	Full interaction logs are stored for compliance

This pattern creates three structural cost drivers.

1. INFERENCE WASTE	2. PERSISTENT INFRASTRUCTURE OVERHEAD	3. COMPENSATORY GOVERNANCE
Inference runs even for irrelevant, exploratory, or disallowed questions. Refusals still consume tokens and infrastructure because the model has already executed.	RAG architectures require continuous document ingestion, embedding computation, and vector database operations regardless of actual usage volume.	Because the model is permitted to speculate, institutions must pay for moderation, safety tooling, and review processes to mitigate errors after generation.

> **Cost is incurred before the system knows whether a question should be answered at all.**

THE COMPAISS DISTINCTION: EXECUTION-GATED ARCHITECTURE

COMPAiSS reverses this sequence:

- Inference does not execute unless authorization succeeds
- Unsupported or out-of-scope queries trigger no model call
- No tokens, retrieval pipelines, or moderation systems run for unauthorized queries

> **Cost avoidance is structural, not behavioral.**

The authorization gate operates through deterministic, non-AI algorithms that match queries against predefined institutional structures and approved sources. The gate itself runs at negligible compute cost relative to large language model inference.

Process Flow Comparison

CONVENTIONAL ENTERPRISE AI	COMPAiSS
User question → Inference begins → Retrieval → Answer generation → Safety filters → Moderation → Output	User question → Authorization check → If unauthorized: terminate with guidance If authorized: AI inference → Generate answer from greenlisted sources → Output

> **This is not optimization of an existing model — it is elimination of entire cost drivers.**

Safe Failure and Bounded Error

FALSE NEGATIVES	FALSE POSITIVES

Legitimate questions blocked: Users receive redirection to authoritative sources or guidance to refine the request. Impact is temporary and low-risk.	Edge cases allowed: Responses remain grounded exclusively in approved institutional sources, preventing fabrication.

COMPARATIVE FINANCIAL ANALYSIS: ANNUAL PROJECTIONS

Baseline: ~100,000 Annual Queries | ~60% Authorization Rate

- Token pricing reflects current enterprise market norms at time of writing
- Authorization rate is a planning assumption; actual rates vary by scope maturity
- All figures are annual operating costs, excluding one-time onboarding

Token estimates reflect total model context per query, not only the user's question. In conventional RAG systems, retrieved documents, citations, and system instructions are inserted into the prompt prior to generation, often constituting the majority of tokens consumed per interaction.

PARAMETER	CONVENTIONAL RAG	COMPAiSS
Annual query volume	~100,000	~100,000
Authorization rate	100% (inference runs on all queries)	~60% (planning assumption)
Token composition per authorized query	User query + retrieved documents + instrulee.	User query + bounded generation (no document insertion)
Average tokens per authorized query	~15,000–30,000	~3,000–8,000
Model class	Enterprise LLM (GPT / Claude / Gemini)	Enterprise LLM (Claude / Gemini)
Operating environment	Regulated, audit-required	Regulated, audit-required
Annual inference (token) cost only	**$10k–$30k**	**$7k–$18k**

> In retrieval-augmented generation (RAG) systems, token usage is dominated by the insertion of retrieved document excerpts into the model prompt prior to generation. COMPAiSS does not insert documents into the model context; documents are used exclusively for authorization and scope control. As a result, authorized COMPAiSS queries consume materially fewer tokens per interaction.

While COMPAiSS inference is denied for approximately 40% of queries to avoid unnecessary cost, those users still receive a high-value safe failure response (e.g., direct links to authoritative policies), providing institutional service at zero marginal compute cost.

Side-by-Side Cost Comparison

COST COMPONENT	CONVENTIONAL RAG	COMPAiSS
Inference (tokens)	**$10k–$30k**	**$7k–$18k**
RAG infrastructure	**$30k–$60k**	**$0**

Safety & moderation	**$30k–$75k**	**$0**
Monitoring & compliance	**$10k–$20k**	**$5k–$15k**
TOTAL ANNUAL COST	**$90k–$200k**	**$15k–$30k**

> Note: These figures represent operational AI inference, infrastructure, and governance costs. Commercial licensing terms, contractual fees, or institution-specific service agreements are addressed separately and vary by deployment size and scope.

STRUCTURAL INTEGRITY: WHY ACCURACY IS A BYPRODUCT OF DESIGN

Lower cost does not imply lower accuracy.

Hallucinations require active inference. By preventing inference for unauthorized queries, COMPAiSS removes the structural conditions under which fabrication occurs.

Unlike systems that attempt to filter or insure against hallucinations after generation — leaving residual tail risk — COMPAiSS eliminates the possibility of speculative reasoning by preventing unauthorized inference altogether.

> **Accuracy is enforced through bounded scope, not post-hoc correction.**

CONCLUSION: STRATEGIC INSTITUTIONAL ALIGNMENT

At institutional scale, COMPAiSS delivers both cost reduction and risk mitigation through the same mechanism: preventing unauthorized AI reasoning before it occurs.

This architectural distinction enables institutions to meet rising demand for automated support while maintaining accuracy, auditability, and fiduciary responsibility.

> **COMPAiSS allocates budget exclusively to authoritative, value-adding interactions — while structurally eliminating the cost, risk, and governance overhead of unauthorized AI reasoning.**

COMPAiSS Inc. | Cost Savings by Design