Architectural Comparison: Standard AI/RAG vs COMPAiSS

Standard Generation-First AI / RAG Architecture

STAGE	WHAT HAPPENS AT THIS STAGE	*WHAT THIS MEANS FOR ACCURACY & RISK*
Model Inference (Always Open)	The AI model is active and ready to generate responses by default. Inference begins as soon as a question is received.	*The system assumes it will provide an answer before knowing if one is appropriate.*
User Question Intake	The user's question is received in any language and passed into the system for processing.	*No institutional scope or authorization check occurs at this point.*
Language Translation / Normalization	If the question is not in English, it may be translated as part of the retrieval and reasoning process.	*Translation can influence which documents are retrieved and how meaning is interpreted.*
Embedding Generation	The question is converted into a mathematical representation to enable similarity search against stored documents.	*The system searches broadly across all available data without institutional filtering.*
Vector / Document Retrieval (RAG)	The system retrieves documents or passages that seem relevant based on similarity matching. This may include external, outdated, or non-authoritative sources.	*Retrieved content may not be institutionally verified or authorized.*
Prompt Assembly	The AI combines the user's question with retrieved documents to construct a prompt for generating an answer.	*The AI may fill gaps, infer context, or draw from training data beyond what was retrieved.*
Post-Generation Controls	After the answer is generated, it may pass through filters, safety checks, moderation layers, or human review to catch errors.	*Controls reduce but do not eliminate hallucinations or unsupported claims.*
Final Answer Delivered	The answer is translated back (if needed) and presented to the user, often with disclaimers about potential inaccuracy.	*Users receive answers that may include unauthorized links, unverified information, or mixed sources.*

In this model, AI is open from the start and answers are controlled through stages, but hallucinations are not eliminated.

COMPAiSS Execution-Gated Architecture

STAGE	WHAT HAPPENS AT THIS STAGE	*WHAT THIS MEANS FOR ACCURACY & RISK*
Model Inference (Always Open)	The AI model is not active by default. Inference is gated and will only occur if authorization succeeds.	*The system does not assume it will answer before determining if it should.*
User Question Intake	The user's question is received in any language and immediately evaluated for institutional scope and authorization.	*Questions outside institutional authority are blocked before any AI reasoning begins.*
Language Translation / Normalization	If the question is not in English, it is translated solely to normalize meaning. This is a linguistic step only, not a retrieval or reasoning step.	*Translation does not influence which sources are considered or how answers are determined.*
Embedding Generation	The normalized question is converted into a mathematical representation, but only within a pre-bounded institutional truth space.	*Search is restricted to greenlisted, institution-approved sources before any AI reasoning occurs.*
Vector / Document Retrieval (RAG)	The system retrieves only from authorized institutional sources. If no authoritative material exists, the process stops here.	*Only verified, institution-approved content can inform an answer.*
Prompt Assembly	If authorization succeeds, the AI is instantiated and combines the question with authorized materials. The AI cannot infer beyond what is explicitly provided.	*The AI is confined to institutional sources and cannot fill gaps from general training data.*
Post-Generation Controls	Because the answer is based solely on authorized sources, post-generation filtering is unnecessary for accuracy or authority.	*Hallucinations based on unauthorized sources are structurally prevented, not retroactively caught.*
Final Answer Delivered	The answer is translated back (if needed) and presented to the user, grounded exclusively in institutional materials. If no answer was authorized, the user is informed.	*Users receive only precise, authorized answers or a clear statement that no answer is available.*

In this model, AI reasoning is gated before it begins, preventing unsupported answers from being generated at all.