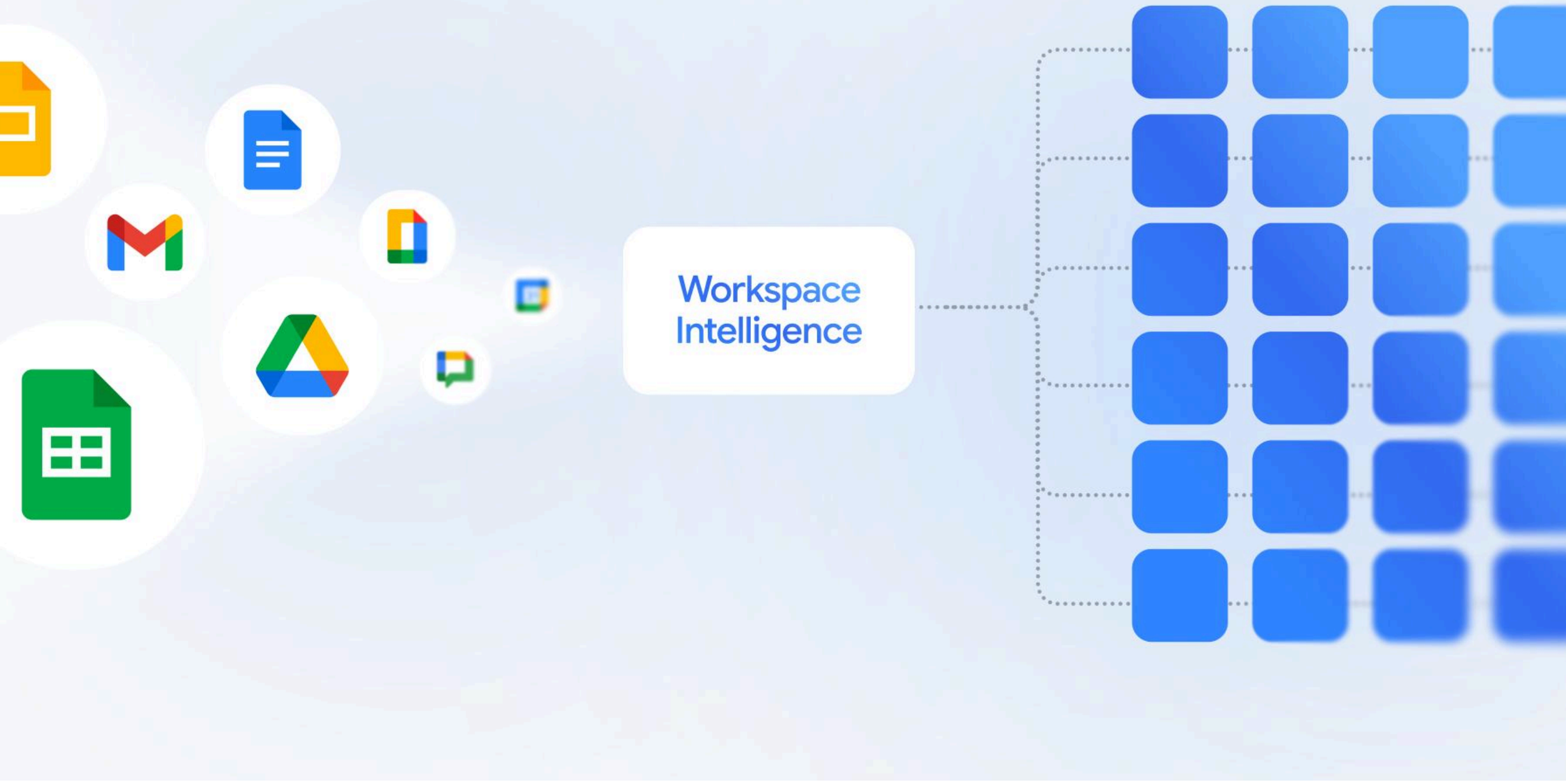
Workspace Intelligence

Contextual AI for the enterprise



The value of an AI assistant isn’t measured solely by its reasoning capabilities, but also by its ability to understand your context and complete complex tasks on your behalf

And in an enterprise environment, this must be accomplished in a secure, governed, and compliant manner.

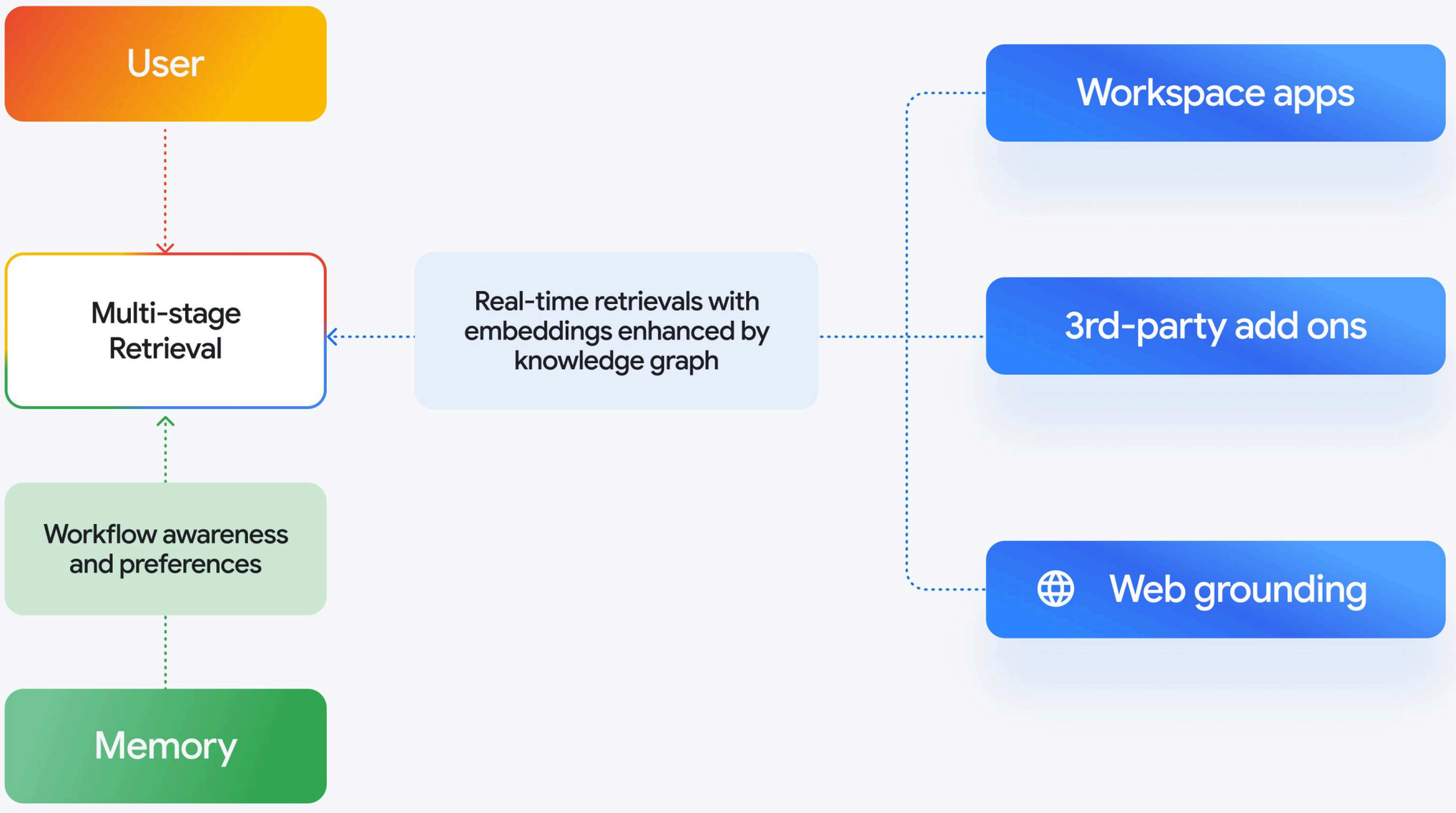
Workspace Intelligence is the foundational architecture that makes this possible at planet scale. It is a secure, dynamic system that inherently understands complex semantic relationships within your specific work ecosystem – your Workspace content (like in Gmail or Google Drive), your active projects, your collaborators, and your organization's domain knowledge – without the need to provide this context for each of your prompts or tasks. By strictly adhering to your existing enterprise data boundaries and controls, Workspace Intelligence ensures that every generative AI interaction is grounded in the reality of your day-to-day work, delivering the right information, in the right format, at exactly the right time.

Table of contents

The need for contextual AI in the enterprise

Frontier large language models (LLMs), like Gemini, possess incredible multimodal reasoning and generative capabilities out-of-the-box. Yet, they inherently lack specific knowledge and context of a user's preferences and immediate projects, organizational structure, and institutional knowledge. Consequently, users often face the "blank canvas" problem: they spend excessive time manually gathering documents from various different surfaces, copying and pasting background context, and writing lengthy, repetitive prompts just to ground the AI sufficiently to get a useful response.

Workspace Intelligence solves this friction. We define Workspace Intelligence as the secure, intelligent system of retrieving and reasoning over an organization's semantic data and ambient Workspace signals to produce a more accurate, personalized, and helpful response. It acts as the critical universal semantic context layer between the computational power of Gemini and the specialized, ever-evolving knowledge of the enterprise.



How context is assembled

When a user submits a prompt, Workspace Intelligence utilizes a sophisticated, multi-stage retrieval architecture to assemble context. Rather than executing simple keyword searches, a query understanding layer parses the natural language intent to dynamically plan a coordinated, multi-source data retrieval strategy across Workspace apps, third-party add-on sources integrated with Workspace, and the web. If the initial retrieval requires more depth, the system can autonomously initiate supplementary query loops to gather targeted, necessary information. Finally, a relevance engine digests this raw data by intelligently sanitizing, scoring, and compressing the gathered context. This multi-layered process ensures the underlying Gemini model receives dense, high-signal information optimized for its context window length, stripping away noise to deliver accurate and grounded responses.

The three pillars of Workspace context

To provide precise grounding for every interaction, Workspace Intelligence synthesizes information across three foundational pillars. It leverages real-time retrieval to surface relevant enterprise data and knowledge graph connections, maintains continuous workflow awareness to track recent cross-application interactions, and securely integrates web grounding to provide external factual context. This multi-layered approach ensures that Gemini understands not just the explicit request, but the user’s organizational environment, current workstream, and the necessary external information.

01 Real-time Retrieval

When a user asks Gemini in Workspace to do a task, Workspace Intelligence generates a real-time search across the user's enterprise corpus, including emails, documents, chats, and calendar events.

- **Enterprise knowledge graph**
  The knowledge graph of your Workspace data acts as a sophisticated, continuously updated map of relationships within the user’s organization. Rather than viewing files and emails in isolation, the graph understands the connections between entities, surfacing the most important data to accomplish your task.
- **Beyond keyword search**
  Traditional token-based retrieval relies on exact keyword matching, which often fails to capture human intent. Workspace Intelligence utilizes embedding-based retrieval. Embeddings capture the deep semantic meaning of documents. For example, a search for “Q3 revenue projections” will accurately surface a document titled “Fall Financial Estimates,” even if the exact keywords are absent.

02 Workflow Awareness and Memory

Context isn’t just about historical emails and files; it’s about ambient awareness of the user's current task. Workspace Intelligence securely aggregates recent, cross-application interactions to build a short-term “memory” or scratchpad for Gemini.

- **Ambient awareness as a ranking signal**
  By understanding short-term work context, Gemini can make more intelligent tie-breaking decisions during retrieval. For example, if a user asks for “the latest project matrix,” the system might find several relevant files. However, if the Workflow Awareness layer notes that the user was just reading a specific matrix document and discussing it in a chat five minutes ago, Workspace Intelligence applies a significant relevance boost to that specific file.
- **Reference resolution**
  This persistent context enables seamless, multi-turn conversations across applications. A user can exit a Google Meet call and simply ask, “Draft an email based on the meeting I just had,” and the system will instantly identify the relevant calendar event, attendees, and accompanying meeting notes to execute the task.

03 Web Grounding

When an enterprise user's prompt requires external factual grounding, the system accesses live Google Search. This capability allows Workspace Intelligence to securely blend internal enterprise knowledge with current world events, industry news, or external market data – all within a single, cohesive response.

For all of the processing described, including web grounding, Google acts as the data processor and only the relevant sub-portion of the prompt is submitted to the search index. You retain the ownership of your data end-to-end and are protected by the enterprise-grade privacy commitments outlined in the "Security, Privacy, and Enterprise Trust" section below.

Engineering for quality: Evaluation and performance constraints

Delivering comprehensive, multi-loop context in real-time is a complex engineering challenge. We optimize Workspace Intelligence to balance two competing priorities: system recall (finding all the right information) and user-facing latency (delivering it quickly).

- **Consolidated ranking architecture**
  Historically, retrieving data from disparate enterprise data sources (email vs. documents vs. chat) required fragmented, highly-latent architectures. Workspace Intelligence relies on a modernized, unified infrastructure that allows for cross-corpus ranking in a single pass.
- **Balancing recall and latency**
  Because foundational generative models have a latency floor dictated by prefill and decoding speeds, fetching 100% of available enterprise data is not feasible; it would overwhelm the context window and introduce unacceptable delays. Instead, indexing traffic is heavily optimized for high-throughput, generating embeddings for massive volumes of data in the background. Conversely, the serving traffic is heavily optimized for latency, using specialized, smaller routing models to ensure Workspace Intelligence can retrieve the most critical "golden artifacts" in its first planning loop, targeting seamless, low single-digit second response times.
- **Evaluation methodologies**
  To continuously ensure quality at scale without compromising data confidentiality and privacy, Google utilizes advanced, automated AI-driven evaluations. These methodologies rely on synthetic data and secure, localized models to safely assess retrieval accuracy. This ensures the routing and retrieval architecture improves continuously while meeting our enterprise-grade security commitments.

Security, privacy, and enterprise trust

Workspace Intelligence is built on the fundamental, non-negotiable principle that your data remains yours:

- **Privacy by design**
  Your content is not reviewed by humans or used for Gemini model training outside your domain without permission. Your content is not used for ads targeting.
- **Security by default**
  Our underlying Gemini models are built on top of Google's zero trust and secure-by-default infrastructure. We extend our secure-by-design principles to the operation of models, with robust protections against malicious AI use, including prompt injection risks. With AI-powered phishing and malware protection in Workspace, we block the vast majority of threats from ever reaching users.
- **Granular data governance**
  Gemini only interacts with data that your users already have access to. As an admin, you can further restrict how Gemini accesses sensitive data with advanced data governance and sovereignty controls. Admins will also be able to choose which sources are consulted by Workspace Intelligence using a new control in the Admin Console.
  Workspace Intelligence respects data processing restrictions established by administrators within their domain.
- **Out-of-the-box compliance**
  Gemini in Workspace has one of the most comprehensive sets of safety, privacy, and security certifications internationally recognized by regulatory and compliance bodies, including being the first generative AI assistant for productivity and collaboration suites to have achieved ISO 42001 – the world's first international standard for Artificial Intelligence Management Systems – and FedRAMP High authorization.

We achieve this within Workspace Intelligence via the following controls:

- **Strict access controls (ACLs)**
  The entire retrieval and orchestration architecture rigorously adheres to existing enterprise permissions in Workspace. The system only retrieves, ranks, and generates context from files, emails, and data that the user already has explicit permission to access. If a user cannot search for a file organically, the AI cannot see it, retrieve it, or reason over it on their behalf.
- **Security by design and tenant isolation**
  All intelligence – including the connections mapped in the enterprise knowledge graph and any dynamically retrieved artifacts – is localized strictly within the customer's tenant boundaries. This data is rigorously isolated. It is a core Google Workspace commitment that your customer data is **not** used to train AI models outside your organization without permission.

Conclusion

AI is only as helpful as the context it is given. Workspace Intelligence acts as the secure, intelligent bridge between Gemini’s reasoning capabilities and the practical, fast-paced realities of personal and enterprise productivity. By mapping complex relationships, utilizing state-of-the-art precomputed embeddings for dynamic retrieval, and maintaining an ambient, real-time awareness of the user's workflow, Workspace Intelligence ensures that enterprise AI transcends generic assistance to become an indispensable, context-aware partner.

Experience a more helpful AI today

Learn more