Gemini for Google Workspace

Privacy, Security, Data Governance, & Compliance White Paper

April 2025 edition

While artificial intelligence (AI) has enhanced digital product experiences for many years, the advent of generative artificial intelligence (Gen AI) is rapidly transforming how we work.

As Gen AI introduces new modalities of communication and collaboration, security practitioners are tasked with keeping up with this rapid pace of innovation, while avoiding unnecessary risks to their organizations.

At Google Workspace, we believe that every company can take full advantage of Gen AI without compromising on privacy, security, and compliance. Based on our experience supporting thousands of enterprises during their AI enablement journeys, we want to ensure that security and IT professionals are confident in their decision to use Gemini for Google Workspace. Throughout this paper we'll explore how we've incorporated Google's AI Principles into a set of privacy, security, data governance, and compliance controls built into Gemini for Workspace.

Disclaimer: The following paper covers using Gemini for Google Workspace with a qualifying edition of Google Workspace, and does not cover the privacy, security, and compliance considerations for NotebookLM and the Gemini app. You can find more information about NotebookLM and the Gemini app in our Generative AI in Google Workspace Privacy Hub.

Table of contents

The market for Gen AI solutions is diverse, encompassing native features within core business systems, specialized niche providers catering to specific use cases, and accessible models that facilitate custom development. Regardless of the chosen path, the first step organizations should take when evaluating solutions is to understand how privacy, security, data governance, and compliance controls are implemented. Top of mind concerns when evaluating solutions include:

- ☐ **Privacy:** What safeguards are implemented to ensure that my user's data privacy is protected?
- ☐ **Security:** What protections are in place to protect my organization and users from malicious use?
- ☐ **Data governance:** How can I ensure that my data governance policies are respected and enforced?
- ☐ **Compliance:** Can I leverage AI without violating my compliance requirements?

Gemini for Workspace is a suite of Gen AI tools designed to help your teams work smarter and faster across Workspace apps, built on-top of Google's Gemini family of large language models (LLMs). These models can be multimodal—meaning they are capable of using text, images, and audio for both input and output—and are designed to integrate with agents. Our approach in developing Gemini for Workspace was based on the same privacy- and security-first principles that Workspace was developed on.

With Gemini for Workspace, you get:

Privacy by design: Your content is not reviewed by humans or used for Gemini model training outside your domain without permission. As a reminder, it is also not sold or 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 protections.

Out-of-the-box compliance: Gemini for 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.

Summary of Gemini for Workspace privacy and security controls

	Gemini for Workspace DOES NOT...	Gemini for Workspace DOES...
Data access	Access Workspace content that your users don't have permission to access	Access relevant Workspace content that your users have permission to access based on their prompts
Data use	Use your users' content, prompts, or generated responses to train Gen AI models outside your domain without permission	Use your users‘ prompts and relevant Workspace content to generate a response
Data protection	Share your users' prompts or generated responses with other users or organizations	Automatically apply your existing data protection controls when inserting generated responses into emails or documents
Abuse prevention	Use potentially malicious sources to generate a response	Automatically block and exclude content that is identified as malicious (such as spam, phishing, or malware) from responses

☑ **Privacy by design:** What safeguards are implemented to ensure that my user's privacy is protected?

When it comes to understanding how an AI model interacts with your data in Workspace, there are two key concepts:

Model training: The process of teaching a model to extract the correct patterns and inferences from data by adjusting the probability of a given outcome. Google's foundational language models are trained primarily on publicly available, crawlable data from the internet.

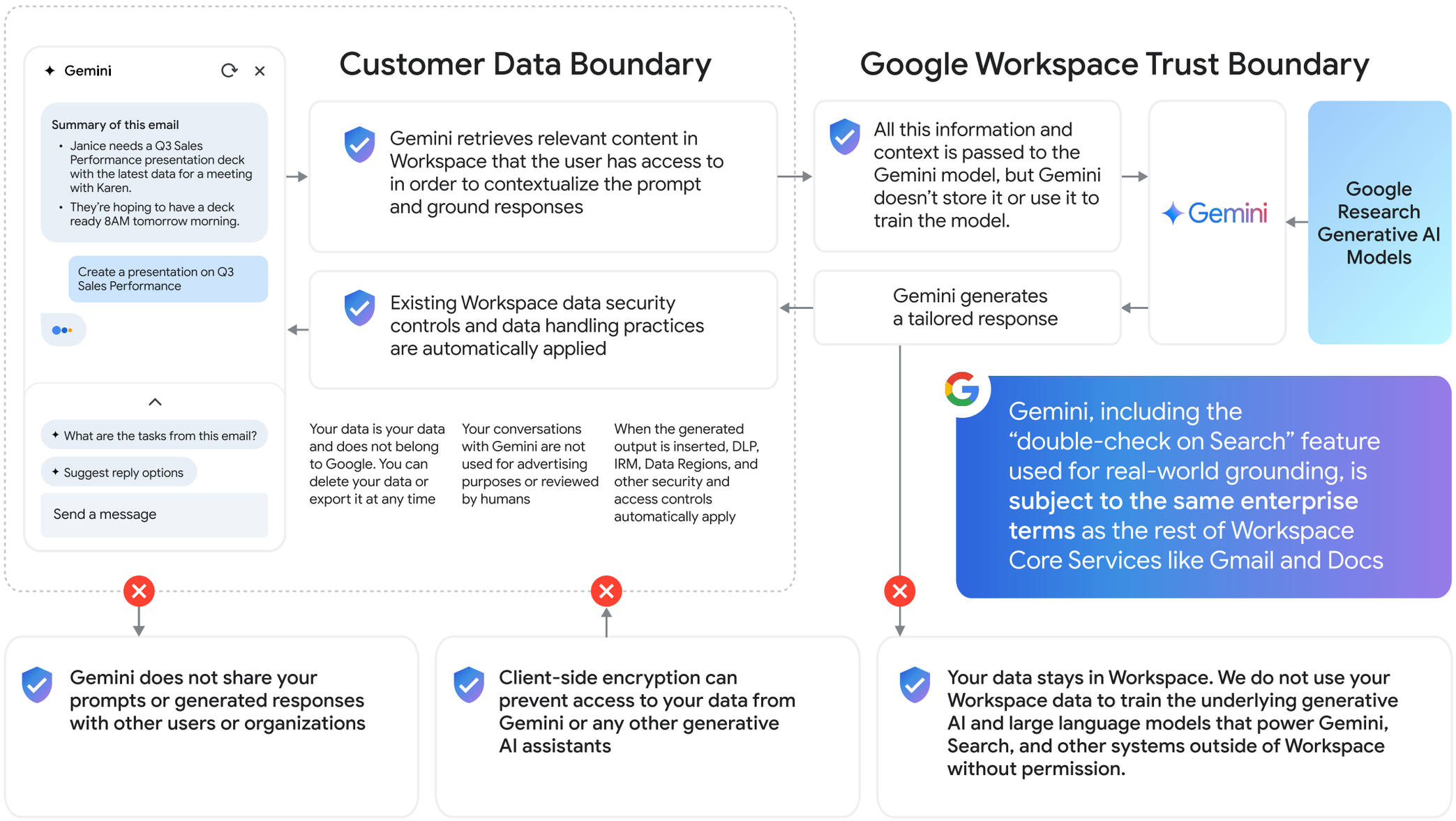
Corpus grounding: The act of a model evaluating a specific input (such as a Google Drive file) to respond to a prompt.

With Gemini for Workspace, your data access permissions and privacy are fully respected throughout the prompt lifecycle:

1. Gemini for Workspace only retrieves relevant content your users have access to in Workspace (referred to as **corpus grounding**). This helps clarify the prompt's meaning, and makes sure the response is helpful based on real-world facts.

2. This information and context is passed to the Gemini model and Gemini for Workspace creates a tailored response from within your **trust boundary**. This is the virtual barrier that controls which information your users can access and share, both within and beyond your organization. Your organization's interactions with Gemini for Workspace all happen within this boundary and all your organization's security, privacy, and access controls are automatically applied. This means your organization can control where your data is stored and processed, and ensure that only authorized parties can access that data.

Gemini for Workspace – Life of a Prompt



Gemini respects the existing data access and security **controls that you already have in Workspace.**

For example, if you don't have the permission to access a document in Drive, Gemini will not retrieve it.

If there is an IT policy that prevents you from downloading, printing, or copying a document, Gemini will also not retrieve it.

☑ **Security by default:** What protections are in place to protect my organization and users from malicious use?

As a pioneer in the practice of Secure by Design, Google applied these same principles to the development of our foundational AI models which power Gemini for Workspace.

Secure AI model development

Google's approach to securing the development of our foundational models has three pillars:

1. **Secure environment:** Our model development environment leverages Google's zero trust and secure by default infrastructure. This includes robust access management, asset inventorying, and risk tagging for all ML assets.

2. **Supply chain integrity:** To ensure model integrity, we follow our Responsible Generative AI toolkit guidance, including maintaining observability for training data, including evaluations for source tampering, data poisoning, and training data quality.

3. **Rigorous testing:** Our foundational Gemini models undergo rigorous adversarial training and red team testing by teams of world class safety experts.

This end-to-end approach enables advanced AI experiences that put safety first, but we don't stop there:

- Our safety and security teams provide 24 / 7 / 365 monitoring of all Google products, services, and infrastructure.

- Through our Vulnerability Reward Program, we collaborate with and incentivize the security research community across 68 countries to identify and address vulnerabilities in our generative AI products. In 2024 alone, we awarded $11.8 million to more than 600 researchers who contributed to the safety and security of our products.

- Our Google Threat Intelligence Group (GTIG) closely studies adversarial misuses of AI, including prompt attacks or other AI-specific threats. With the learnings from GTIG, we continuously improve our AI models to make them less susceptible to misuse, and we apply our intelligence to improve Google's defenses and protect users from cyber threat activity.

Secure model operation

Simply building secure models is not enough, their operation needs to be vigorously defended against adversarial use. Google has built and continues to enhance controls in Gemini to defend against the adversarial misuse of AI, including advanced abuse and prompt injection defenses. Workspace customers can experience three times fewer email security incidents than with traditional solutions. This is due, in part, to our multiple layers of advanced threat defenses that block more than 99.9% of spam, phishing attempts, and malware. When content is identified to be malicious, such as spam, we automatically block Gemini from accessing it. Workspace utilizes these protections to develop advanced protections specifically for the emerging threat vector of direct and indirect prompt injection attacks.

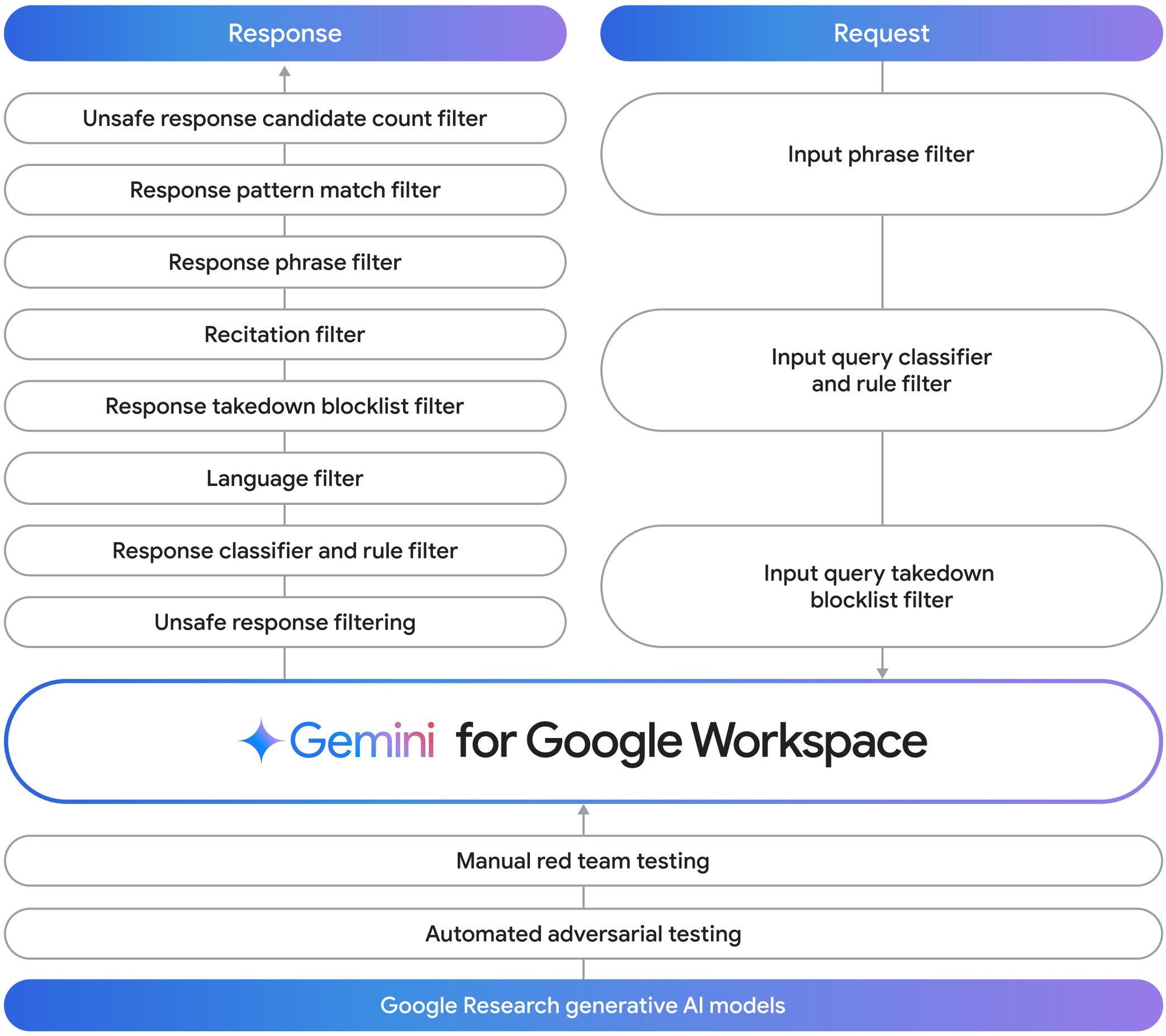
Prompt injection attacks are cybersecurity threats specifically targeting AI tools that use prompts for model instructions. Much like malicious software code that is executed by a computer's operating system, prompt injections manipulate the LLM's interpretation and output to bypass built-in security mechanisms and lead to unintended and potentially harmful outcomes. Prompt injections also have similarities to phishing as they often rely on deceptive communications to trick the LLM into behaving in unintended ways.

Similar to our progress against other forms of malicious attacks, Workspace is making significant investments in this area by building a layered approach to protect against prompt injection attacks. This approach includes advanced monitoring that leverages supervised fine tuning and the sanitization of inputs, URLs, and responses at each layer to prevent malicious instructions from being performed.

Expanding on this concept further, as a pioneer in AI research, innovation, and products, Google has developed proprietary approaches to detect and block many forms of prompt injection attacks. For example, below are some of the specific detection mechanisms that prompts in Workspace are evaluated for:

- **Indirect malicious instructions:** Identify and mitigate malicious instructions found in files, emails, and attachments used to generate content and output responses.
- **Malicious safety abuse attempts:** Detect abusive instructions to prevent model safety bypasses.
- **Malicious URLs:** Help ensure that web links and images presented to Gemini for Workspace users are safe from data-exfiltration, phishing, and malware.

Our early analysis has shown that this layered defense approach successfully detects and mitigates the majority of prompt injection attacks.



Automated adversarial testing

To help ensure our defenses are ready to combat the threats our models will face, we have developed a red-team framework. Our red-team testing consists of several optimization-based attacks that generate prompt injections, such as inserting malicious prompt instruction into a Gmail message or a document in Google Docs. These optimization-based attacks are designed to be as strong as possible; weak attacks do little to inform us of the susceptibility of an AI system to indirect prompt injections.

Our ongoing investment in security protections to safeguard current and future AI systems against new adversarial threats, such as prompt injection attacks, is informed by insights gleaned from our red-team framework, internal research from groups like the Google Threat Intelligence Group, and industry research. In the Google Security blog, you can read more about our specific investments in red-teaming to defend against prompt injection attacks.

☑ **Granular data governance:** How can I ensure that my data governance policies are respected and enforced?

Secure Gen AI deployments start with strong data governance. Gemini respects existing Workspace data access and security controls, and retrieves relevant data that the user has the permission to access. For organizations concerned about broad oversharing and overly permissive sharing behavior, Workspace offers robust data governance controls:

1. **Classification labels:** Classification labels, a data-classification solution for Google Drive and Gmail, provide an adaptable framework for identifying and categorizing information to enable granularity in data lifecycle management, data protection, auditing, and discoverability-and-search scenarios. With classification labels, organizations can specify up to 150 unique labels, each with flexible metadata structures and configurable permissions, through an administrator-defined taxonomy. Workspace also offers AI classification, an AI-powered solution for automating data classification.

2. **Data loss prevention rules:** Workspace data loss prevention (DLP) capabilities allow administrators to control the sharing of sensitive information with configurable content rules that identify sensitive content and apply policy enforcements. DLP rules can use admin-defined or pretrained content detectors and classification labels as rule conditions to trigger policy enforcements such as blocking external sharing of Drive files, Gmail messages, and Chat messages; or disabling the option to download, copy, print, and email files in Drive. Learn more about data loss prevention.

3. **Trust rules:** With trust rules, administrators can create granular policies to control who can access their organization's Drive files. Policies can apply to individual users, groups, organizational units, and domains to specify:
    - Which users' files can be shared with internal or external users
    - Which users can receive files from internal or external users
    - Which internal or external users can be invited and add items to shared drives

Because trust rules provide flexibility in establishing collaboration boundaries, they can help your organization secure sensitive information and maintain compliance with industry standards and regulations. Learn more about creating and managing trust rules.

4. **Context-Aware Access:** Using Context-Aware Access, you can create granular access control security policies for apps based on attributes such as user identity, location, device security status, and IP address. For apps that are core services, such as Gemini for Workspace, policy evaluation is continuous. For example, if a user signs in to a core service at the office and walks over to a coffee shop, a Context-Aware Access policy for that service is rechecked when the user changes location.

Limiting Gemini's access to sensitive information

Gemini for Workspace is designed to be an extension of the user, only accessing and interacting with data that the user already has access to. However, some organizations have specific use cases where they want to further restrict which data Gemini can access on behalf of a user. To do so, customers can leverage AI classification and data loss prevention (DLP) capabilities together to help identify sensitive data, automatically apply classification labels, and enforce Information Rights Management (IRM) controls based on the classification labels. This helps to restrict Gemini for Workspace from retrieving targeted data for the users under the IRM restriction. For example, if a user isn't allowed to download, print, or copy files based on the IRM policy, Gemini will not retrieve those files or their content on the user's behalf. In addition, generated output inserted into emails in Gmail or documents in Drive are automatically evaluated against in-scope DLP policies set by domain administrators.

For scenarios where data is of a highly-confidential or sensitive nature, Workspace client-side encryption (CSE) controls can be applied to prevent Gemini for Workspace from being able to access the data. CSE, in general, provides an additional layer of data protection and helps prevent unauthorized access from any third-party, including Google and third-party entities itself. Administrators can also turn on CSE by default for select organizational units (for example, finance or HR).

Workspace also offers granular audit logs in Drive for activity triggered by Gemini for Workspace. For example, if Gemini for Workspace accesses data from a set of files in response to a user query, an "item content accessed" event is generated for each of the accessed files in the Drive log events.

☑ **Compliance:** Can I leverage AI without violating my compliance requirements?

Industry recognized certifications

Gemini for Workspace can help your organization stay compliant with your regulatory requirements. Gemini for Workspace has one of the most comprehensive sets of safety, privacy, and security certifications, including SOC 1/2/3, ISO 9001, ISO/IEC 27001, 27701, 27017, 27018, and 42001 certifications. In particular, the achievement of ISO/IEC 42001, the world's first international standard for Artificial Intelligence Management Systems (AIMS), certifies that Gemini for Workspace has been developed, deployed, and maintained responsibly with appropriate ethical considerations, data governance, and transparency.

	Gemini for Workspace	Gemini app gemini.google.com	Microsoft 365 Copilot[1]	ChatGPT Enterprise[2]
SOC 1	✓	✓		
SOC 2/3	✓	✓		✓
ISO 27001	✓	✓		
ISO 27017	✓	✓		
ISO 27018	✓	✓	✓	
ISO 27701	✓	✓		
ISO 9001	✓	✓		
ISO 42001	✓	✓	✓	
CSA STAR	✓	✓		✓
GDPR	✓	✓	✓	✓
CCPA	✓	✓	✓	✓
HIPAA	✓	✓	✓	
FedRAMP High	✓	✓		
BSI C5	✓	✓		

1. Based on analysis of publicly available Microsoft Compliance Offerings documentation as of 4/2/2025
2. Based on analysis of publicly available OpenAI Security & Privacy documentation as of 4/2/2025

Gemini for Workspace is the first generative AI assistant for productivity and collaboration suites to have achieved FedRAMP High authorization and meet BSI C5 criteria. It can also help your organization meet HIPAA compliance.

Ready to get started?

If you're ready to get started, check out our step by step AI adoption guidelines and whitepaper, find additional information in the Generative AI in Google Workspace Privacy Hub, and get started with a no-cost trial.