Google

September 2026

Google Private AI Compute: Enabled with Secure Server-Side Memory

From Google Platforms and Devices, Google DeepMind, Google Core, and Google Cloud.

We believe privacy is the foundation of personalized AI assistance

The most effective AI doesn't just answer a user's query; it anticipates needs and understands the rich personal context behind them, such as their current task, their ambient environment, or their upcoming commitments. This requires processing complex and often sensitive data, such as audio, end-to-end encrypted messages, or device screen content.

Historically, the architecture for securely processing this type of data was on-device processing. The advancements to more proactive and personal AI experiences require a level of deep reasoning and advanced capabilities that demands the computational power of larger models, like the latest Gemini models, hosted in the cloud.

Private AI Compute in the cloud is our privacy platform, built to deliver the speed and power from the cloud while extending the **user security and privacy assurances of on-device processing.** When a user interacts with Google AI experiences running on Private AI Compute, the system is engineered to ensure that their personal information, unique insights, and how they use them are **private only to the user.**

In our latest update, we have extended the capabilities of Google Private AI Compute with a secure, persistent, and isolated storage layer. This advancement enables the next generation of agentic and AI experiences that require long-term context, cross-device or cross-platform context, and user preferences. More details about this trusted memory architecture can be found in the section below "Introducing secure server-side memory".

Private AI Compute is built on years of Google's existing investments in security and privacy and is engineered to deliver the following requirements.

User data processed by Private AI Compute is not available to anyone other than the user, including Google

Information is processed within a system that is designed to protect sensitive user data, which is referred to in this document as the protected execution environment, based on a hardware root of trust. No user data leaves the protected execution environment without explicit user intent or action.

Hyper-scalable private inference

To bring everyone the best of our AI, Google's cloud infrastructure is highly optimized to produce exceptional serving efficiency, capacity, and computational power. This design enables Google's most advanced AI models to be served at rates of millions of user queries per second. Private AI Compute uses this end-to-end globally distributed and optimized model serving infrastructure to deliver private inference on the latest Gemini models, powered by Google's own custom Tensor Processing Units (TPUs).

External verifiability

For both the initial release and the update enabling secure server-side memory, external auditors have validated that the system design meets strict privacy and security guidelines. A summary of their reports is available: 2025 report here, and 2026 report here. We continue to improve the transparency and attestation mechanisms that will let users and outside experts verify the system directly over time.

Whether executing a real-time prompt or retrieving historical user context, every workload operating within the Private AI Compute Platform shares a unified privacy and security baseline. The following sections detail how we deliver on these requirements for the stateless system and explain our memory-enabled architecture.

Table of contents

01. Our design approach for the platform

Private AI Compute builds on our deep investment in delivering AI responsibly and years of Google's innovations in privacy infrastructure and advanced security, such as Private Compute Core to protect your most sensitive mobile device data; Federated Learning and Differential Privacy, which enable us to improve our products without revealing sensitive information; and end-to-end encrypted backup. On these foundations, combined with our existing Google Cloud security architecture, we designed Private AI Compute with the following additional multi-layered protections to ensure that user data is not available to anyone other than the user.

Create a protected execution environment across both CPU and TPU workloads ("trusted nodes")

For CPU workloads we use an AMD-based hardware Trusted Execution Environment (TEE) to encrypt and isolate memory and processing from the host.

For TPU workloads such as serving LLM workloads, we use a hardened TPU platform that delivers privacy and security properties comparable to a typical TEE, by using our TPU hardware (Google's Titanium Intelligence Enclave).

Establish encrypted communication channels between trusted nodes

As the Private AI Compute serving infrastructure involves a multi-node distributed system, the system performs peer-to-peer attestation and encryption between the trusted nodes to ensure user data is decrypted and processed solely within a protected environment.

Attest trusted nodes

Attesting trusted nodes ensures their integrity as part of establishing encrypted communication channels such as Noise and Application Layer Transport Security (ALTS), to ensure that user data is shielded from broader Google infrastructure. The initiation of these encrypted connections uses bi-directional attestation: each workload requests and cryptographically validates the workload credentials of the other, ensuring mutual trust within the protected execution environment. Workload credentials are provisioned only upon successful validation of the node's attestation against internal reference values. Failure of validation prevents connection establishment, thus safeguarding user data from untrusted components.

In the Private AI Compute architecture, the client establishes a Noise encryption connection with a frontend server and establishes mutual attestation of auditable binaries over that connection. Subsequently, the frontend server establishes an ALTS encryption channel with other services in the scalable inference pipeline and with model servers running on the hardened TPU platform. These services and servers are provisioned only upon successful attestation validation. The overall trust in the Private AI Compute system is established by transitive trust between these servers.

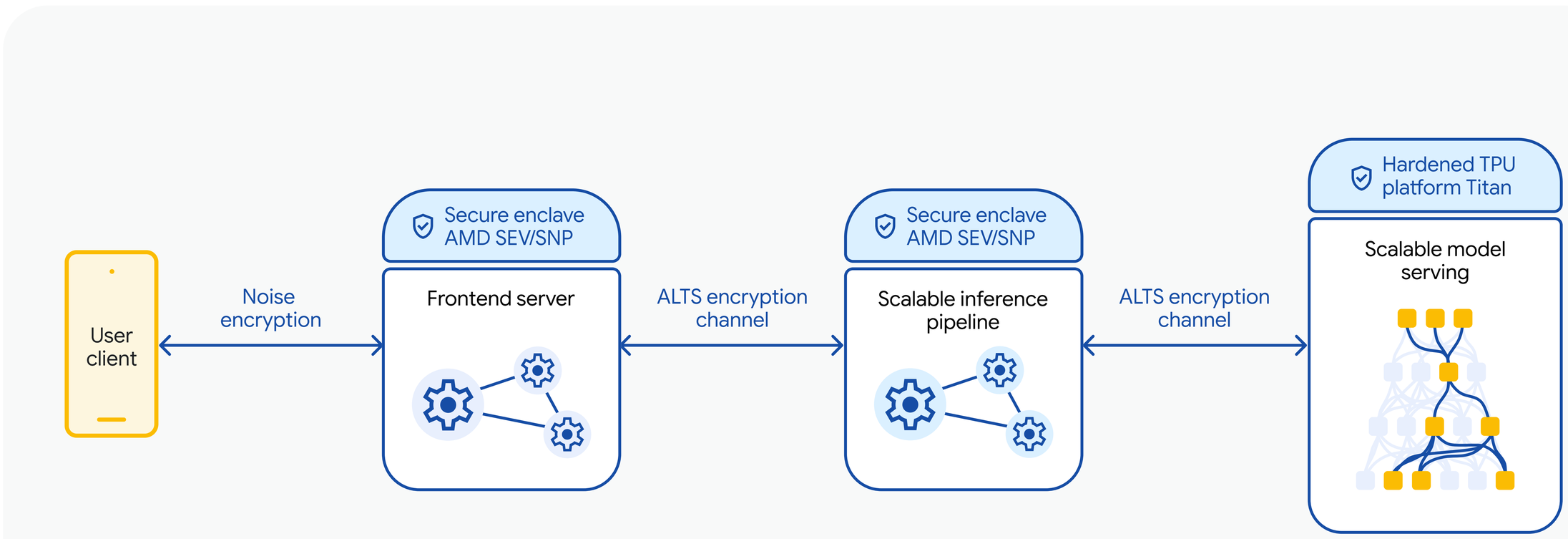

Figure 1: Simplified diagram to illustrate the Private AI Compute chain of trust.

Reduce the trusted computing base of AI services

Private AI Compute minimizes the number of components and entities that must be trusted for data confidentiality, to only the essential components required to maintain system integrity and protect sensitive user data.

Ensure private telemetry and analytics

Deployments of Private AI Compute that require analytics and aggregate insights make use of privacy-enhancing technologies (PETs), such as confidential federated analytics, to ensure that only anonymous statistics (e.g. differentially private aggregates) are visible to Google. With confidential federated analytics, unaggregated data is processed by open source software protected by hardware TEEs, currently Oak on AMD SEV-SNP with a hardware root of trust. The privacy properties of confidential federated analytics are transparent and can be verified by external parties.

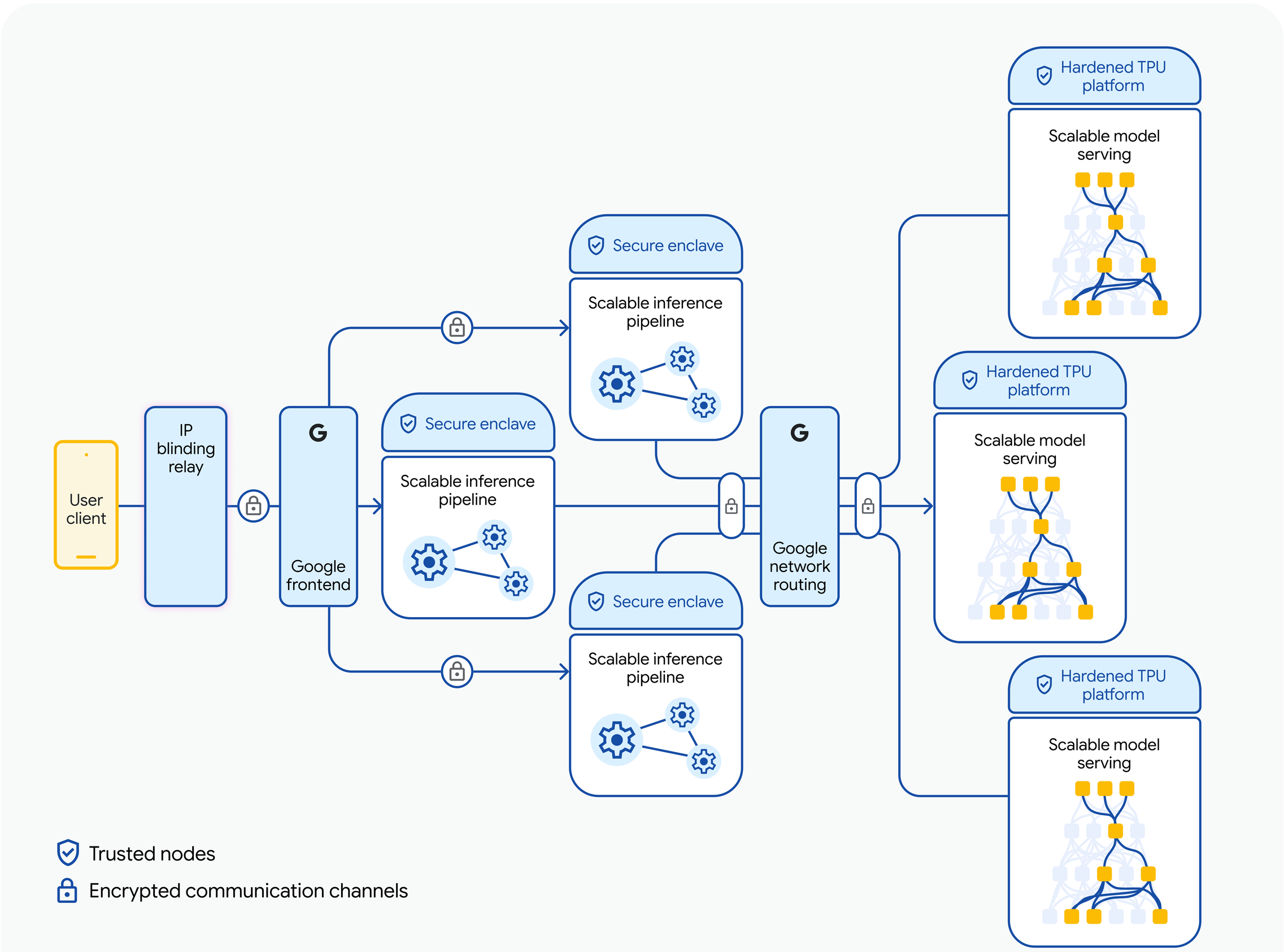

Figure 2: The system architecture of Private AI Compute.

02.

Protections against privacy and security threats to user data access

In addition to the above privacy engineering measures, we know that strong privacy requires strong security to maintain the integrity guarantees of the system, including malicious modification of its privacy properties.

Private AI Compute is built on the foundation of our existing leading security infrastructure, including encryption for client-server communications; and binary authorization ensuring only signed, authorized code and validated configurations are running across our software supply chain. Binary Authorization for Borg, for example, provides similar security guarantees to the industry standard SLSA framework. Robust audit trail logging provides the foundation for all these measures, ensuring clear identification and comprehensive review of any security-relevant events or deviations.

As the sections below detail, Private AI Compute is implemented on top of this robust foundation and goes further – to defend against security risks such as insider access, service exploitation, and data exposure. Private AI Compute isolates user data in Virtual Machines (VMs) to contain compromise in case of a bug, hardens our systems against physical exfiltration with memory encryption and input/output memory management unit (IOMMU) protections, and uses hardware and software attestation to ensure system integrity.

Protections for privileged access misuse

Insider threats can happen when privileged access is abused to gain unauthorized access to data. For example, an insider could gain unauthorized control over critical system functions, make unauthorized changes to exfiltrate user data; or the insider can identify and target specific users or data flows and try to redirect and manipulate them, using knowledge such as client IP addresses or user credentials.

Mitigations

Private AI Compute is designed to implement the following protections against any single or coordinated insider access.

Ephemeral by design for stateless inference

(see Section "Introducing secure server-side memory" for the stateful implementation). The system is designed so that inputs, model inferences, and computations are only kept as long as needed to fulfill the user's query. Attackers cannot access past data. User data is processed in a protected execution environment at the time of inference request and discarded when the user session is completed.

No privileged access to user data

Private AI Compute is designed so that administrative access to user data is not possible. This achieves the highest level of controls for production services, preventing human access to the data even in "break glass" emergency scenarios. We do this with:

- **TEE on CPU platform:** The CPU workloads are hosted in confidential virtual machines, a hardware-isolated, memory-encrypted environment shielded from the underlying infrastructure. This ensures user data protection during processing because the workload in a guest virtual machine is protected from the host.
- **Zero shell access on hardened TPU platform:** Arbitrary execution (e.g. shell access) is not possible on nodes running in the Private AI Compute system, including our hardened TPU platform.
- **Physical security:** Google's best-in-class, multi-layered physical security ensures malicious attackers cannot access data centers globally.

Run key services on a state-of-the-art confidential computing platform based on AMD's hardware Trusted Execution Environment

Frontend services run in a confidential virtual machine: a hardware-isolated, memory-encrypted environment shielded from the underlying infrastructure. This ensures user data protection during processing because the workload in a guest virtual machine is protected from the host and the code is verified via attestation.

Verify server authenticity to prevent eavesdropping and tampering

Before any data exchange occurs, clients establish trust with a Private AI Compute endpoint through validating the endpoint server's identity. This relies on remote attestation and cryptographic key validation, along with a secure session protocol, to confirm that the endpoint server is genuine, unmodified, and actively protected by the referenced security measures.

Non-targetability with IP protection for stateless inference

Even if an attacker were to defeat all other defenses in place, they would still need to distinguish a specific user's data from the immense volume of total network traffic to intercept it. We make this computationally infeasible in our stateless system.

- **IP-blinding relay:** We use IP blinding relays which are operated by third-parties to tunnel all traffic bound for the Private AI Compute system. This removes the ability for an attacker to link your IP address–or any other network identifying information–to your specific query, and thereby removes the ability to distinguish one user's traffic from another.
- **Isolation of authentication and authorization:** We isolate the system's authentication and authorization from inference using Anonymous Tokens. Device authentication and rate limiting are handled by a separate server, which exchanges device credentials for an anonymous, cacheable token to manage usage for the entire system. This token is presented to the Private AI Compute server to authenticate the user without revealing their identity. A user's identity never travels with their data on the inference path.

Ensure detection of cybersecurity incidents

To protect against emerging threats, detection analytics leverage relevant security signals, enriched by frontline threat intelligence and expertise from Google's Threat Analysis Group and Mandiant teams. This ensures a deep understanding of evolving adversary techniques. By integrating Gemini's advanced reasoning, the analysis is accelerated to proactively detect and mitigate incidents based on real-world attack patterns.

Protections against data exposure from unintended errors or vulnerability exploits

Unintended errors in computing systems can lead to data exposure, and these errors can include software bugs, hardware malfunctions, misconfigurations, or even unexpected interactions within complex systems. This can lead to unintended logging, sharing user data with monitoring tools, and inadvertently granting human access.

Additionally, potential attackers can exploit flaws in the hardware components, as well as vulnerabilities in the operating system, hypervisor, and application code to bypass security mechanisms, and grant unauthorized access to system memory and thus user data.

Mitigations

In order to deliver robust protection against both unintended errors and vulnerability exploits, Private AI Compute is designed with the following measures.

Carefully control the input and output from Private AI Compute system jobs

Granular egress policies have been put in place to prevent sensitive data from leaving the system across diverse channels such as monitoring, data logs, or core dumps. These policies cover each data point according to its sensitivity, with a default of no egress to ensure user data cannot be accessed.

Safeguard service integrity through VM isolation and runtime protection

Private AI Compute services run on confidential computing and hardened TPU platforms with hardware roots of trust. Confidential computing uses the virtual machine layer to isolate services and dependencies to ensure a single instance compromise can be contained and promptly remediated, preventing adverse effects on other services.

Robust to continuous availability and updates

Private AI Compute leverages Google's best-in-class secure infrastructure, which pushes out frequent updates to address security and reliability issues, such as Distributed Denial of Service (DDoS) prevention and compliance. Private AI Compute is engineered to preserve privacy commitments actively across updates that address security vulnerabilities, optimize performance, and deliver new or improved features.

03.

Introducing secure server-side memory

The stateful extension of Private AI Compute allows Google AI experiences to have a persistent, per-user memory that lives in the Cloud, while not being readable by Google.

Why persistent memory

Ephemeral inference cannot support experiences that depend on continuity across invocations: an assistant that remembers a user's long-running project, the decisions it already made on their behalf, or the preferences it inferred in the past across sessions. On-device storage alternatives face practical constraints: physical capacity limits, friction syncing across multi-device surfaces, and the prohibitive bandwidth and latency overhead of repeatedly transmitting massive historical contexts over the network to cloud-hosted frontier models. Secure server-side memory extends personal context beyond device storage into a Cloud environment engineered so that only attested private workloads running within the protected execution environment acting on the user's behalf can decrypt and access it.

How this memory capability is built

Private AI Compute secure server-side memory operates entirely within the platform's Protected Execution Environment.

Per-user, hardware-isolated database.

At its core is the memory Oak Server, a stateful, per-user database running inside a hardware TEE. Records are encrypted with per-user keys that are never visible outside the Trusted Computing Base or to Google infrastructure. Decryption occurs strictly within the TEE using keys bound to the user, allowing the surrounding platform to store, replicate, and back up the encrypted database without ever having access to the underlying plaintext.

Orchestrator that keeps data inside the boundary.

An orchestrator mediates between the AI model and the memory Oak Server, ensuring that data routing, model calls, and storage operations remain strictly within the protected boundary. User data never leaves the enclave in plaintext.

Attested, encrypted channels

The orchestrator, the memory server, and the inference pipeline authenticate one another using the same bi-directional attestation and encrypted channels described in stateless system. A component that cannot produce valid attestation evidence cannot join the data path.

Written in memory-safe language

The memory TEE application is written in Rust and runs in the Oak Containers runtime, whose core components are also written in Rust, substantially reducing the risk of memory-safety vulnerabilities in this new code.

Open source

The memory Oak Server and Oak Containers runtime are open source, allowing third-party researchers to inspect implementation code and verify that binary digests deployed in production correspond directly to auditable source releases. Because persistent state carries a fundamentally higher risk profile than ephemeral computation, Google has made the memory Oak Server open source for independent public review. Additionally, for establishing what runs in production: the published source is tied to the deployed enclave through reproducible builds, public endorsement of the resulting binary digest in an append-only ledger, and attestation of that same digest by the enclave before any key is released — see the section External verifiability for the platform.

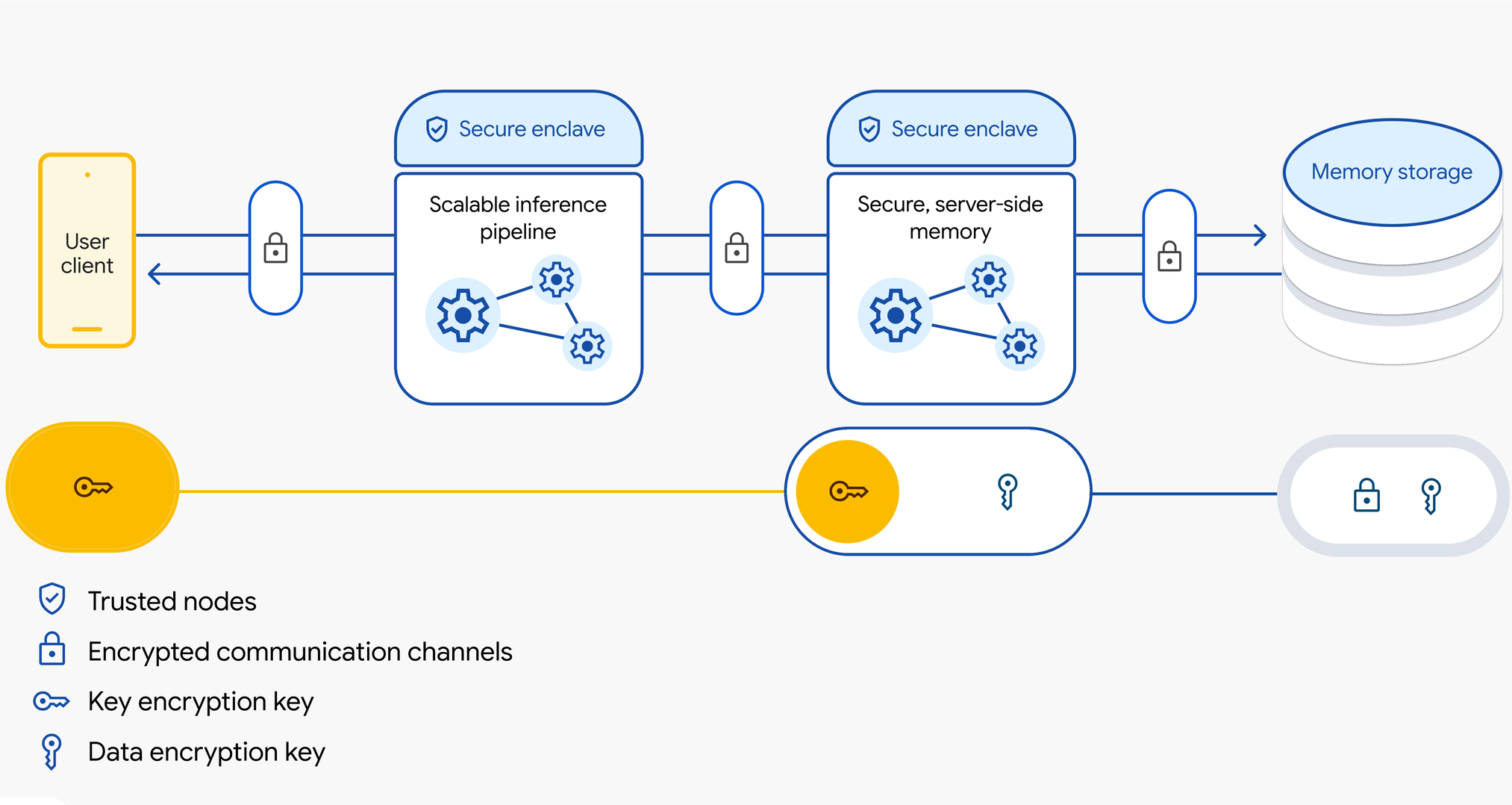

Figure 3: A simplified diagram of Private AI Compute with secure server-side memory.

By combining hardware-enforced secure enclaves, encrypted channels, and per-user databases shielded by device-derived encryption keys, this architecture ensures your data stay fully private and under your control.

How memory and inference interact in a request lifecycle

To fulfill an agentic query requiring historical context, memory and inference coordinate across secure enclave boundaries:

1. **Attested ingress and routing:** The client establishes an encrypted session via the Noise Protocol. The request reaches the orchestration enclave inside an AMD SEV-SNP confidential VM.
2. **Attested context retrieval:** The orchestrator establishes a mutually attested ALTS channel to the memory Oak Server. Upon hardware verification of the enclave's measurement, the user's decryption keys are unsealed to the database engine. The relevant records are decrypted strictly in volatile enclave RAM and returned to the orchestrator.
3. **Augmented inference:** The orchestrator merges the retrieved memory context with the active prompt and securely transmits the bundle over an attested internal channel to the hardened TPU platform. The model evaluates the prompt entirely within the TPU boundary; no plaintext prompt or activation crosses that boundary, and the platform's protections against observation of that computation are described in the Threat model for persistent data.
4. **State commit and volatile purge:** If the session produces new memories, facts, or updated preferences, the orchestrator commits them back to the memory Oak Server, where they are re-encrypted under the user's key and written to persistent storage. Immediately upon response delivery, the orchestrator and TPU platform wipe all volatile prompt context, tokens, and intermediate activations, ensuring zero plaintext residual state.

Key management

Memory records are encrypted under keys that are bound to the user, not to Google's infrastructure, and the key material is only ever available in plaintext inside an attested TEE. Key release to the memory server is gated on attestation: an enclave that does not present valid attestation evidence matching an endorsed memory binary cannot obtain the keys, and therefore cannot read a user's memory — including if the enclave is a modified or unauthorized build.

Threat model for persistent data

The security goals for a persistent store are the following:

Isolated by design, rather than ephemeral

When an experience utilizes the Private AI Compute Platform’s memory capabilities, memory data is retained by definition. Protection comes from hardware isolation and per-user encryption; the data exists in a form that no one outside the attested enclave can interpret.

No privileged access

There is no administrative path to plaintext user data, even in break-glass scenarios.

Compromise containment

The memory server runs in a confidential VM, so a compromise of one instance is contained and does not expose other users' state.

Egress control

Default-deny egress policies apply: memory contents cannot leave through monitoring, logging, or core dumps – unless with user intent.

Non-targetability

The inference path is capable of non-targetability because it is impossible to link a query to a user: nothing about the request is tied to a user identity. A persistent memory store is different by construction — a request must reach the user's store, so the system must resolve a stable per-user identifier. Consequently, the stateful system does not claim network-level non-targetability. Instead, targeting a specific user's memory yields only opaque ciphertext. Without the user's keys, which are only available inside an attested enclave, the data remains unreadable.

04. External verifiability for the platform

We believe it is critical that users can trust in the claims we make about our software.

To this end, Google has made the underlying binaries for platforms and features like Android, Private Compute Core, confidential federated analytics, and our ML suite available for public inspection and review. This means ensuring these claims can be verified for the software that is used to process their data, and be verified as the same software that third parties can inspect and audit.

Building on this, releases of Private AI Compute ensure that the software processing user data can be both externally audited and cryptographically attested. The sections below describe the assurances available today through third-party expert audits, our system-level attestation, and a tamper-evident public ledger of production binary digests, and the direction we are taking to put more of that evidence directly in the hands of users and outside experts.

Third-party privacy and security audits

In order to ensure that the privacy of the system is verifiable by others outside of Google, we enable third-party privacy and security audits of our system binaries and source code. Expert review provided a detailed analysis of and recommendations for our system. We are committed to continuously improving our security posture.

A summary of the 2025 audit finding is available here and the 2026 audit finding is available here.

Attestation: establishing a verifiable binary identity for our compliant servers

Google's production infrastructure ensures that all software and hardware on the machine is authorized for the Private AI Compute system. Private AI Compute application binaries are only scheduled on machines that are running intended privacy-preserving software. Its critical components are available for third party audit and digests are published to a ledger.

Private AI Compute continues to anchor its trust in hardware, combining system-level attestation, measured from the first mutable code to the OS via the Titan security chip with hardware based TEEs. Upon successful hardware and software attestation, server instances that adhere to our strict privacy assurances are provisioned with a cryptographic key. Clients verify this key before transmitting any data, establishing a cryptographic anchor for the server's binary identity. This prevents connections to unauthorized or modified services and ensures an authenticated, encrypted channel into the protected execution environment. For user-side transparency, Private AI Compute network requests can be inspected directly in the Network Logs on supported Pixel and Android devices.

05. What's next

We believe transparency and verifiability are key pillars of trust.

To this end, we have made the underlying binaries for platforms and features like Android, Private Compute Core, confidential federated analytics, and our ML suite available for public inspection and review. Transparency and external verifiability are also central to the Private AI Compute Platform. With the introduction of secure server-side memory, we have taken another foundational step forward by anchoring persistent state in open-source, buildable enclaves that outside experts can inspect.

Building on this foundation, future releases of Private AI Compute will continue to move verification evidence outward, enabling users and independent experts to verify system integrity directly. This includes:

Client-side attestation verification

Moving from Google-asserted compliance toward client-side cryptographic verification, enabling user devices to independently validate server attestation evidence and endorsed binary identities before transmitting sensitive data.

Independently witnessed transparency

Evolving our published binary ledger toward a continuous, append-only transparency log observed and co-signed by independent third parties, ensuring the public record of deployed software is tamper-evident and consistent.

Continuous scrutiny and engagement

Broadening reproducible build coverage across additional system components, maintaining recurring third-party audits, and actively engaging independent security researchers and privacy experts for continuous feedback on our platform architecture.

We look forward to continuing to share more information about Private AI Compute with you, including additional capabilities that will run on this privacy-preserving infrastructure.