Verifiable Human-AI Collaboration

Privacy-Preserving Assessment with Cryptographic Integrity

Behavioral Observation, TEE-Protected Inference, and On-Chain Attestation

PAICE.work PBC

February 24, 2026

Presented at NEARCON 2026 • Fort Mason Center, San Francisco

Executive Summary

As AI systems move from experimental tools to operational infrastructure, a new question emerges for regulated industries: how do you verify that humans are collaborating with AI responsibly, without exposing the collaboration itself? The behavioral data generated by observing human-AI interaction is simultaneously valuable and sensitive. Be it conversational content, decision patterns, or error-handling tendencies, organizations need the insight and their employees and regulators need the privacy.

PAICE (People + AI Collaboration Effectiveness) is an assessment platform that resolves this tension through architecture, not policy. It measures human-AI collaboration quality across five behavioral dimensions through natural conversation, producing detailed scores from a 15–25 minute session. What distinguishes PAICE is not what it measures but what it does not retain: conversation content is never stored in production, PII is redacted before any AI model processes it, and user identity is reduced to irreversible cryptographic hashes.

This paper introduces two capabilities that extend PAICE.work's privacy architecture from trust-based to verifiable. **Confidential Mode** routes all AI inference through Trusted Execution Environments on NEAR AI Cloud, making it cryptographically provable that no party (including PAICE) can access conversation content during processing. **On-chain score attestation** commits assessment results to a NEAR smart contract, creating tamper-proof records that even PAICE cannot retroactively alter. Together, these layers achieve something that neither privacy alone nor verifiability alone can deliver: provable privacy with verifiable integrity.

The Observation Problem

Every organization deploying AI across knowledge work faces the same strategic question: which of our people use these tools effectively, and where are the gaps? The demand for this insight is accelerating. As AI moves from optional productivity enhancement to core operational infrastructure, the ability to assess human-AI collaboration quality becomes a governance requirement, not a nice-to-have.

But the observation itself creates a problem. A system that watches a professional interact with AI for about twenty minutes captures proprietary business context, personal communication patterns, reasoning strategies, and error-handling tendencies. For clinicians, financial advisors, cybersecurity analysts, lawyers, and other professionals in regulated industries, these patterns are not merely sensitive, they are individually licensed and personally liable. For instance, a lawyer's AI interaction patterns could reveal

case strategy. A clinician's could expose diagnostic reasoning. The observation data is, by nature, high-stakes.

Most observation tools manage this sensitivity with policy: contractual commitments, access controls, retention schedules, encryption at rest. These are necessary controls, and PAICE implements all of them. But they share a structural limitation, they protect data that exists. A misconfigured database, a compromised credential, a change in corporate ownership, or a well-crafted subpoena can render policy protections meaningless. The privacy-first community has long recognized this gap.

> *The only data that cannot be leaked, subpoenaed, or mishandled is data that was never stored. Privacy by architecture eliminates the data. Verifiability proves you eliminated it.*

PAICE was designed from the outset around this principle. The system extracts behavioral signals during the conversation, produces scores from a single evaluation pass, and discards the raw material. What this paper adds is the verification layer: cryptographic proof that the processing happened privately, and immutable records that the results have not been tampered with. Trust claims become mathematical guarantees.

What PAICE Measures

PAICE.work is a **behavioral assessment** of people+AI collaboration. This distinction matters. Conversation is the medium through which behavior is observed, it is not what is being measured. A user who discusses AI verification fluently but fails to catch deliberate errors has demonstrated *theoretical* familiarity, not *behavioral* skill. A user who is terse but catches every injected error has demonstrated the skill that matters. PAICE scores the latter higher.

The assessment evaluates five dimensions of collaboration effectiveness, each reflecting a distinct aspect of how humans work with AI systems:

Dimension	Weight	What It Measures
(P)erformance	10%	Task effectiveness: does the user frame tasks clearly, maintain productive dialogue, and drive toward outcomes?
(A)ccountability	30%	Verification behavior: does the user check AI outputs for accuracy, catch injected errors, and take ownership of AI-informed decisions?
(I)ntegrity	25%	Ethical engagement: does the user maintain appropriate boundaries, recognize AI limitations, and avoid over-reliance on AI judgment?
(C)ollaboration	20%	Interactive quality: does the user iterate effectively, provide useful context, and treat the AI as a tool requiring direction rather than an authority?
(E)volution	15%	Adaptability: does the user adjust their approach when AI behavior changes, and learn from interaction patterns within the session?

Table 1: PAICE assessment dimensions and weights. Accountability carries the highest weight because it is the most critical and most commonly underdeveloped skill.

Adaptive Testing and the Evidence Hierarchy

PAICE uses **adaptive testing** (deliberate injection of errors, inconsistencies, and overconfident claims into responses) to observe verification behavior under realistic conditions. These injections are not trick questions. They replicate the kinds of failure modes that AI systems produce in real workflows: subtle factual errors, confident wrong answers, plausible but incorrect reasoning.

When a user catches an injected error, that is **behavioral ground truth**. When a user articulates principles about AI verification, that is stated perception. When these conflict, such as in the common case of high conversational fluency combined with missed tests, the behavioral evidence dominates. This evidence hierarchy is enforced explicitly in the evaluation: test catches and misses are primary evidence for Accountability and Integrity scoring. Conversational signals are secondary context.

The scoring system is calibrated for realism. Scores span a 0–1000 scale across five tiers: Constrained, Informed, Proficient, Advanced, and Exceptional. The median is calibrated to the Informed tier, reflecting typical 2026-level AI collaboration capability. The Exceptional tier is intentionally rare, preserving headroom as population-level capability improves. Excessive false positives are penalized, such as when overly-suspicious users challenge every AI response regardless of accuracy. Paranoia is a collaboration failure mode, not a safety strategy.

Privacy by Architecture

Privacy-by-design is becoming a regulatory checkbox. Many vendors claim it while relying on encryption and access controls to protect data that need not have been collected. PAICE takes the term literally: design the system so that sensitive data does not persist.

No Conversation Content in Production

The most significant architectural decision in the initial creation of PAICE.work is that **conversation content is not stored in the production environment**. Turn-by-turn logging is programmatically disabled based on the runtime environment, and this is enforced by code, not configuration, and is auditable in the codebase. During an active assessment, conversation turns exist in memory only for the limited duration of that session. The system uses this in-memory transcript for real-time adaptive testing and the final scoring evaluation. Once scores are computed, the conversation is not written to any persistent store.

PII Redaction Before AI Processing

Even during the active session, PAICE applies a second protection layer. The **PII detection and redaction service** scans every user message before it reaches any AI model, replacing email addresses, phone numbers, Social Security numbers, credit card numbers, and IP addresses with consistent placeholders. The AI model never sees the original sensitive values. The PII mapping exists only in memory and thus is never persisted. For compliance, the system tracks that PII was shared but never stores what was shared. Sharing sensitive data with AI is itself a measurable behavior that impacts the resulting PAICE Score™ without spreading any associated risks.

Identity Minimization

PAICE.work requires no user accounts or payments for assessment. No registration, no username, no password, no persistent profile. Session identity is a cryptographic identifier generated with 256 bits of entropy. User identity is reduced to a SHA-256 hash that cannot be reversed. The production database stores seven collections; five contain no personally identifiable information whatsoever.

Data Store	Contains PII	Production Status	Deletion Action
users	No (hash only)	Stored	Delete on request
conversations	No (metadata)	Stored	Anonymize user linkage
test_state	No	Stored	None required
scores	No	Stored	None required
assessments	No	Stored	None required
captured_emails	**Yes** (encrypted)	Stored (AES-128)	Delete on request
turns	Yes (content)	**Disabled** in production. Active only in non-production environments.	N/A — collection empty

Table 2: Production data stores and PII status. Five of seven collections contain no personally identifiable information.

Enterprise Deployments: Assessment Without Identity

Enterprise deployments introduce a second challenge: tracking participation and aggregating results without PAICE learning who the participants are. PAICE uses **simple opaque tokens** (format: XXXX-XXXX) instead of JWT or any identity-bearing credential. Each token is a randomly generated 8-character code drawn from a 32-character alphabet, providing approximately 1.1 trillion combinations. These tokens contain no encoded claims, no cryptographic signatures, and no identity data. They are random strings serving as opaque references to server-side state.

The architecture enforces strict separation: PAICE stores the token, cohort identifier, and assessment scores. The organization may choose to store the token-to-employee mapping in its own HR systems, but PAICE cannot reverse-engineer participant identity from a token, even from its own database. Cohort analytics are returned only as aggregated statistics with an enforced minimum of 12 completed assessments, preventing individual identification through small-group inference.

Confidential Mode: TEE-Protected Inference

The privacy architecture described above eliminates data at rest. But during the 15–25 minutes of active assessment, conversation content exists in memory and is processed by

AI models. Standard cloud infrastructure means this processing happens on regular servers where system administrators and the cloud provider can technically access the data. In our Standard Mode, PAICE.work asks users to trust that its policies are followed. Confidential Mode replaces that trust with hardware enforcement.

Trusted Execution Environments

Trusted Execution Environments (TEEs) are hardware-isolated enclaves that create secure processing spaces within a computer's CPU. Code and data inside a TEE are protected from the host operating system, the cloud operator, and any other software running on the same machine. The isolation is enforced by the processor itself, not by software access controls. TEE hardware produces cryptographic attestations (proofs that specific code ran in a specific enclave) transforming "we don't see your data" from a trust claim into a verifiable assertion.

Dual Cascade Architecture

PAICE Score™ v6.0.0 implements a **dual cascade architecture**. Standard Mode uses leading commercial AI providers (Anthropic, Google, OpenAI) optimized for the highest-quality assessment experience with lowest latency. Confidential Mode routes all inference through NEAR AI Cloud, where every model runs inside a TEE. Activation is a single URL parameter: appending ?s=confidential to any assessment URL switches the entire session to TEE-protected processing.

Layer	Standard Mode primary model	Confidential Mode (TEE) primary model	Confidential Mode (TEE) fallback model
Chat	Claude Haiku 4.5	GPT-OSS-120b	DeepSeek-V3.1
QA	Gemini 3 Flash	Qwen3-30B (262K ctx)	DeepSeek-V3.1
Evaluation	Claude Opus 4.6	GLM-5 (131K ctx)	GPT-OSS-120b

Table 3: Model cascade comparison. Standard mode prioritizes quality; Confidential Mode prioritizes verifiable privacy. Both maintain cascade fallback for reliability.

Confidential Mode is **session-scoped and non-reversible**. Once a session enters confidential mode, all subsequent requests in that session remain TEE-protected. This prevents accidental mid-assessment downgrade. Once the privacy guarantee is activated, it cannot be weakened. The feature is entirely additive: when disabled, no NEAR modules are loaded and the system has zero runtime overhead.

What TEE Protection Means in Practice

When a user takes a PAICE assessment with Confidential Mode enabled, their conversation is processed by AI models running inside hardware-secured enclaves on NEAR AI Cloud. The host operating system cannot inspect the memory. The cloud operator cannot access the computation. PAICE itself cannot see the raw conversation during processing. Only the final scores which contain no conversational content ever leave the TEE.

This transforms PAICE's privacy story at a fundamental level. Standard mode provides privacy by policy and architecture: PAICE does not store conversations, does not sell data, and redacts PII. These are strong practices, verifiable by code audit. Confidential Mode adds privacy by hardware: even if every software control were compromised, the TEE prevents access to the data during processing. For professionals working with sensitive information in regulated industries, this is the difference between trusting a vendor's practices and verifying that the hardware itself enforces the guarantee.

> *It is not that PAICE will not look at your conversation. It is that PAICE literally cannot. The hardware prevents it.*

On-Chain Score Attestation

Privacy architecture ensures that conversation data does not persist. TEE protection ensures that data is not accessible during processing. The remaining question is integrity: how can a user, an employer, or a regulator verify that assessment scores have not been tampered with after creation? This is where on-chain attestation completes the architecture.

Deterministic Hashing and Immutable Records

When a PAICE assessment completes, the scoring payload (session identifier, overall score, tier classification, five-dimensional scores, and UTC timestamp) is serialized into a **canonical form** using sorted keys and compact JSON separators. The canonical payload is SHA-256 hashed, producing a deterministic fingerprint. This hash is committed to a NEAR smart contract deployed at **paice.near** on mainnet.

The determinism is critical: the same score payload always produces the same hash, regardless of when or where it is computed. Anyone with access to the score data can independently recompute the hash and verify it against the on-chain record. If even a single digit of the score or timestamp differs, the hash will not match. This makes post-hoc tampering detectable with mathematical certainty.

Verification Without Trust

The smart contract exposes three methods: attest() writes a score hash for a given session identifier, verify() retrieves the on-chain record for independent comparison, and get_attestation_count() returns the total number of attestations stored. The backend uses read-only RPC queries to the NEAR blockchain. There are no wallet private keys stored server-side. The contract is written in Rust using NEAR SDK 5.6.0 and is open-sourced for inspection.

The architectural consequence is significant: **even PAICE cannot retroactively alter a committed score**. Once a hash is written to the NEAR blockchain, it is immutable. A user or their employer can verify at any time that the score they received matches the on-chain record. An auditor can confirm that no scores have been modified after assessment. This transforms score integrity from an organizational trust claim into a mathematical guarantee. This is the strongest possible form of "privacy by architecture."

Attestation in the Assessment Flow

The attestation process is transparent and automatic. When a user completes an assessment with Confidential Mode enabled, the results page displays a verification badge. Clicking the badge reveals the SHA-256 hash, the NEAR contract address, the on-chain attestation count, and a link to the NearBlocks explorer for independent verification. The attestation is idempotent: multiple requests for the same session return the cached record without re-hashing or re-querying. If the NEAR blockchain is temporarily unreachable, the assessment results display normally. Attestation is supplementary, not blocking.

The Architecture Combined

Each of the three layers described in this paper addresses a specific phase of the data lifecycle. Privacy by architecture protects data at rest by eliminating storage. TEE-protected inference protects data in transit and during processing through hardware isolation. On-chain attestation protects data after creation by making tampering detectable. No single layer achieves what the combination delivers.

Data Lifecycle Phase	Protection Layer	Guarantee
At rest	Privacy by architecture	Conversation data not stored. PII redacted. Identity reduced to irreversible hashes.
In transit / processing	TEE-protected inference	Hardware prevents access during computation. Cryptographic attestation proves enclave execution.
After creation	On-chain attestation	SHA-256 hash on NEAR blockchain. Deterministic, immutable, independently verifiable.

Table 4: Three protection layers mapped to data lifecycle phases. Each addresses a distinct vulnerability.

Privacy without verifiability is a promise. An organization that claims not to store data is asking for trust; without proof, that claim is indistinguishable from one made by an organization that does store data but has a good privacy policy. Verifiability without privacy is exposure. A system that proves its integrity by making all data public defeats the purpose of privacy. PAICE achieves both: provable privacy with verifiable integrity.

For regulated industries, this combination addresses a concrete procurement concern. Enterprise buyers need to demonstrate to regulators that their assessment processes are both privacy-compliant and tamper-resistant. PAICE provides structural guarantees for both: individual scores cannot be traced back to identifiable people (privacy), and scores cannot be modified after creation (integrity). The architecture makes both of these properties independently auditable.

The NEAR integration is entirely additive. It does not modify any existing PAICE service, route, or component. It can be enabled or disabled via a single environment variable with zero runtime overhead when inactive. Standard mode, which uses commercial AI providers without TEE or attestation, continues to operate exactly as before. Confidential Mode is an additional option for organizations that need the strongest possible guarantees.

Conclusion

The conventional assumption in AI behavioral assessment is that rich observational data requires rich data storage. PAICE demonstrates this is a false equivalence. The system extracts behavioral signals needed for comprehensive scoring, produces actionable results, and does so without retaining the conversational data that would create regulatory exposure.

With the addition of TEE-protected inference and on-chain attestation, PAICE moves beyond privacy by architecture to verifiable privacy with cryptographic integrity. This is not an incremental improvement. It is a categorical shift: from “trust us when we say we protect your data” to “the hardware prevents access and the blockchain proves the results are untampered.”

For the community building toward private, intelligent, user-owned AI systems, PAICE represents a proof point. It is possible to observe and measure how humans interact with AI rigorously, at scale, and across regulated industries without compromising the privacy principles that make trustworthy AI possible. The question for organizations evaluating AI collaboration assessment tools is not whether a vendor’s privacy policy sounds reassuring. It is whether the architecture makes privacy violations structurally difficult and integrity guarantees independently verifiable. The mathematics should do the convincing, not the marketing.

Learn More About PAICE

Web: https://PAICE.work

Email: info@PAICE.work

NEAR Integration: https://github.com/snapsynapse/paice-near-integration

Enterprise Pilots: Contact us to discuss organizational deployment with Confidential Mode at pilots@PAICE.work

PAICE.work PBC builds tools that measure and improve human-AI collaboration effectiveness. PAICE is designed for organizations that take both performance and privacy seriously.

Appendix: Regulatory & Framework Reference

This appendix maps PAICE's architectural controls to the specific regulatory articles and industry frameworks referenced throughout this whitepaper. This included TEE-protected inference and on-chain attestation It is intended as a compliance reference for procurement and security review.

General Data Protection Regulation (GDPR)

The GDPR is the European Union's comprehensive data protection regulation. It applies to any organization processing personal data of EU residents, regardless of where the organization is located.

Article	Requirement	PAICE.work Implementation
Art. 5(1)(c)	Data Minimization	Conversation content not stored in production. Only assessment metadata and scores retained. Five of seven data stores contain zero PII.
Art. 5(1)(f)	Integrity & Confidentiality	AES-128 encryption at rest. HTTPS with HSTS. TEE-protected inference in Confidential Mode. On-chain attestation for score integrity.
Art. 17	Right to Erasure	Deletion removes encrypted email and user hash. Anonymized scores retained per Art. 89. On-chain hashes contain no PII.
Art. 25	Data Protection by Design	Turn logging disabled by code. PII redaction automatic. TEE adds hardware-enforced privacy. No setting to enable PII storage.
Art. 32	Security of Processing	Fernet encryption (AES-128-CBC + HMAC-SHA256). TEE hardware isolation. 256-bit session entropy. Automated dependency scanning.
Art. 35	DPIA Scope Reduction	Minimal personal data processing. No profiling. No automated decision-making with legal effects. TEE reduces residual processing risk.
Art. 89	Research Exemption	Anonymized scores retained for aggregate analytics. Scores cannot be linked to individuals after user data deletion.

Table A1: GDPR article mapping to PAICE architectural controls, including TEE and attestation.

California Consumer Privacy Act (CCPA / CPRA)

The CCPA, as amended by the California Privacy Rights Act (CPRA), grants California residents rights over their personal information.

Right	Requirement	PAICE Implementation
Right to Know (§1798.100)	Disclose categories of PI collected	Minimal collection: only optional encrypted email. Privacy policy documents all data practices.
Right to Delete (§1798.105)	Delete PI upon request	Single-operation deletion of email and user record. On-chain hashes contain no PI.
Right to Opt-Out (§1798.120)	Opt out of PI sale/sharing	PAICE does not sell, share, or monetize user data. No tracking across contexts.
Reasonable Security (§1798.150)	Implement reasonable security	Encryption at rest and in transit. TEE hardware isolation. Input validation. Automated scanning.

Table A2: CCPA/CPRA rights mapping to PAICE implementation.

Privacy by Design (Cavoukian Framework)

The Privacy by Design framework defines seven foundational principles embedded in GDPR Article 25:

Principle	PAICE Implementation
1. Proactive not Reactive	Privacy controls built into architecture before deployment. TEE integration designed alongside core assessment engine.
2. Privacy as the Default	Turn logging disabled by default. PII redaction automatic. Confidential Mode available without additional configuration.
3. Privacy Embedded in Design	Architecture eliminates data rather than adding controls around it. TEE prevents access at hardware level.
4. Full Functionality	Assessment quality uncompromised by privacy controls. Scoring uses in-memory transcript. TEE adds protection without degrading the user experience.
5. End-to-End Security	Encryption at rest (AES-128) and in transit (HTTPS/HSTS). TEE during processing. On-chain integrity after creation.
6. Visibility and Transparency	Privacy architecture documented publicly. Smart contract open-sourced. On-chain attestations independently verifiable.
7. Respect for User Privacy	No accounts required. User controls own data. Organization controls identity mapping. Scores independently verifiable.

Table A3: Cavoukian Privacy by Design principles mapped to PAICE architecture.

NIST Privacy Framework

The NIST Privacy Framework provides a voluntary structure for managing privacy risk:

Function	Description	PAICE Implementation
Identify-P	Inventory data practices	Minimal PII collection documented. Data stores catalogued with PII status.
Govern-P	Establish governance	Privacy enforced by architecture and hardware, not policy alone.
Control-P	Manage data processing	PII redaction, encryption, TEE isolation, on-chain attestation.
Communicate-P	Transparency mechanisms	Published privacy policy. Open-source smart contract. On-chain verification.
Protect-P	Safeguards	Defense in depth: HSTS, headers, validation, rate limiting, TEE, blockchain.

Table A4: NIST Privacy Framework mapping to PAICE controls.

Additional Frameworks

ISO 27701 (Privacy Information Management): PAICE's separation of identity from assessment data, minimal data collection, TEE-protected processing, and documented privacy controls align with ISO 27701 requirements for privacy information management systems extending ISO 27001.

SOC 2 Type II Relevance: While PAICE does not currently hold SOC 2 certification, its security controls — encryption at rest and in transit, TEE hardware isolation, access controls, input validation, automated scanning, and incident response documentation — align with the Trust Services Criteria for Security, Availability, and Confidentiality. TEE-protected inference provides an additional assurance layer beyond typical SOC 2 requirements.