Legitimacy by History: Longitudinal Evidence for Proportionate AI Safety

Christine Hähner-Murdock
ORCID: 0009-0007-5868-8011
DOI: 10.5281/zenodo.21442605
Position paper - preprint | 19 July 2026
Status: conceptual proposal; empirical validation pending

Abstract

Safety systems for conversational AI rely heavily on constraints that operate at the level of the current request or conversation: post-training alignment, refusal policies, and real-time classifiers. These safeguards are necessary, but false positives are not merely a usability cost. Repeated unnecessary refusals may push legitimate users toward the same rephrasing techniques that safety systems are built to resist, spread low-level evasion literacy, and suppress or fragment the contextual evidence that a more discriminating safety system would need. We propose supplementing constraint-based safety with account-level longitudinal evidence: LEGITIMACY BY HISTORY. Genuine months-long interaction histories may contain structural evidence that is costly to backfill and, where the provider retains generation-time commitments, index-like evidence that specific platform events occurred when claimed. We separate committed provenance (attested timestamps, generation records, watermarks), inferred model-era compatibility, bilateral human–AI adaptation, structured variability, and external corroboration. None of these establishes benign intent: an account can be taken over, a user can change direction, and a patient actor can cultivate an authentic history for later exploitation. The operational proposal is therefore deliberately one-directional: authenticated history, operator continuity, and current trajectory may GRANT latitude — resolving ambiguous, dual-use requests toward assistance that a stateless filter would refuse — and may never, by themselves, restrict below the standard default. This makes explicit a premise current architectures leave implicit: the conversational model itself participates in safety judgment, as a sandboxed assessor whose authority is confined to within-account latitude, checked by coarse provider-side reassurance calls, and subordinate to human adjudication and independent hard constraints. We state the resulting design requirements — functional non-participating defaults, grant audit trails, multi-channel independence from the assessed relationship, indeterminate outcomes, human review before any consequential restriction — and the honest ceiling: intent cannot be certified from history. This position paper presents a falsifiable architecture and a research program, not a validated detector.

Keywords: AI safety; longitudinal interaction; false refusals; provenance; account continuity; structured variability; human–AI adaptation; graduated trust

1. Problem

Many deployed safeguards for conversational AI evaluate the current request, the current generation, or a short conversational context under broadly shared rules. Model-level alignment and real-time checks are indispensable, but at the level where they operate, the user's demonstrated record contributes little: a prompt-level mechanism that judges every request as if it came from an unknown stranger must set its thresholds for the worst-case stranger. The predictable consequence is a persistent false-positive rate borne disproportionately by legitimate users working in exactly the domains safety systems watch most closely. False-refusal benchmarks document the safety–usability trade-off, and recent work in cybersecurity shows that authorized defensive tasks can be refused disproportionately when they contain operational security vocabulary (An et al., 2024; Campbell et al., 2026; Xue et al., 2026). The burden is asymmetric: attackers using unaligned tools face no comparable friction, while defenders relying on aligned systems absorb the capability degradation (Campbell et al., 2026).

The stronger claim that every provider ignores account history would be inaccurate. OpenAI, for example, publicly describes account-level review across relevant conversations and risk signals to distinguish persistent malicious behavior from legitimate dual-use work, while also acknowledging that account-level enforcement is coarse and may flag valuable defensive uses (OpenAI, 2026a, 2026b). Deployed account-level assessment is, however, trust-negative: history is consulted after suspicion arises, to justify restriction. The unoccupied or at least publicly under-specified direction is the opposite one: trust-positive use of longitudinal behavioral and provenance evidence to improve classification and assistance routing for legitimate users.

False positives experienced as unjust create secondary safety effects. A blocked legitimate task gives the user an incentive to rephrase and retry. Rephrasing around refusal is already represented as a reusable prompt-engineering pattern, refusal decisions can be triggered by non-harmful linguistic cues, and users penalize refusals strongly in observed preference

data (Pasch, 2025; White et al., 2023; Xue et al., 2026). Safeguard false positives can also be exploited directly as a denial-of-service surface (Zhang et al., 2025). These findings do not yet demonstrate a population-level training effect in which benign users become jailbreak-literate; they establish the components of a testable mechanism: unnecessary refusals can teach users to alter language, distribute evasion techniques originally developed adversarially to a population that had no adversarial intent, and make benign and adversarial interaction traces harder to distinguish. This does not make models safer. It accelerates the arms race through an avoidable distribution of arms.

The same process degrades future evidence. A user who learns that candor or full context increases the chance of interruption may self-censor, split work across threads or accounts, or move to less observable systems. The safety layer then helps produce the fragmented and strategically worded record that a later account-level assessor must interpret. This closed loop is the motivating systems problem.

Adjacent work. Forward transcript integrity is beginning to receive formal cryptographic treatment; VCT, for example, uses hash chains, Merkle structures, and joint signatures to authenticate nonlinear conversation state (Xing et al., 2026). Account-level misuse review is deployed by at least one major provider (OpenAI, 2026a). A research line on user-adaptive safety is also emerging: trust-oriented adaptive guardrails modulate access to sensitive content through trust metrics combining consistency-based interaction trust with authority-verified credentials (Hu et al., 2025); personalized guardrail systems mine interaction history for user traits and user-specific risk thresholds (Wu et al., 2025); and verification-based access-control proposals address the same dual-use dilemma through documentary authorization of verified users (Wybitul, 2025). None of these lines addresses the combination proposed here: retrospective authenticity verification of unattested histories; provenance-anchored rather than score- or credential-based trust; a strictly one-directional latitude grant with no penalty for absence; the embedded model's own sandboxed trajectory judgment rather than an external profile; behavioral evidence as a pathway for legitimate users who cannot access documentary schemes; and the false-positive burden treated as a first-class safety cost. The novelty claim is the integrated architecture, not each component in isolation.

2. Assumption and scope

The proposal rests on a limited empirical assumption: some classes of sustained legitimate use and some classes of exploitative account operation may differ measurably in their longitudinal structure when topic, volume, sophistication, and platform conditions are controlled. This is a hypothesis to test, not a result.

The proposal does not assert that history proves benign intent, that a long account is morally legitimate, or that every ethically objectionable request warrants account-level suspicion. The term legitimacy is used procedurally: evidence that an account history and its present operator deserve to be treated as more than an unknown, context-free request source. It does not denote moral worth, factual truth, or entitlement to capabilities whose consequences remain unacceptable.

3. Precondition: authentic history

Longitudinal evidence is useful only if the history actually accumulated as claimed. The first question is therefore one of process authenticity: was this record produced over the claimed period, or retrospectively backfilled or stitched together?

A preliminary cross-domain synthesis suggests that no known signal defined over transcript text alone is causally unfakeable. Skilled fabricators can reproduce many content characteristics at a cost, and informing deceivers about verbal credibility criteria can reduce their discriminative value (Vrij et al., 2000). Textual evidence may price fabrication; it does not certify authenticity. Stronger evidence must be anchored to records or processes outside the fabricator's unilateral write access.

The evidence is asymmetric. A hard incompatibility with a trusted platform record can strongly falsify a claimed history: an impossible event order, or an assistant turn containing information that none of its recorded input channels could have supplied at that date. Even then, migration, corruption, export/import error, and configuration metadata must be excluded before adjudication. By contrast, false content inside a genuine log is ordinary: genuine users make mistakes, remember inaccurately, and sometimes lie. Factual error does not establish that the history itself was fabricated.

Confirmation is correspondingly weaker. A history can become increasingly well supported, but no clean transcript establishes benign intent. This one-sidedness should shape the detector: strong contradictions can raise concern; missing evidence remains neutral; clean checks provide bounded reassurance rather than exoneration.

4. Evidence channels

Figure 1 summarizes the channels, their evidence grades, and their route into proportionate handling.

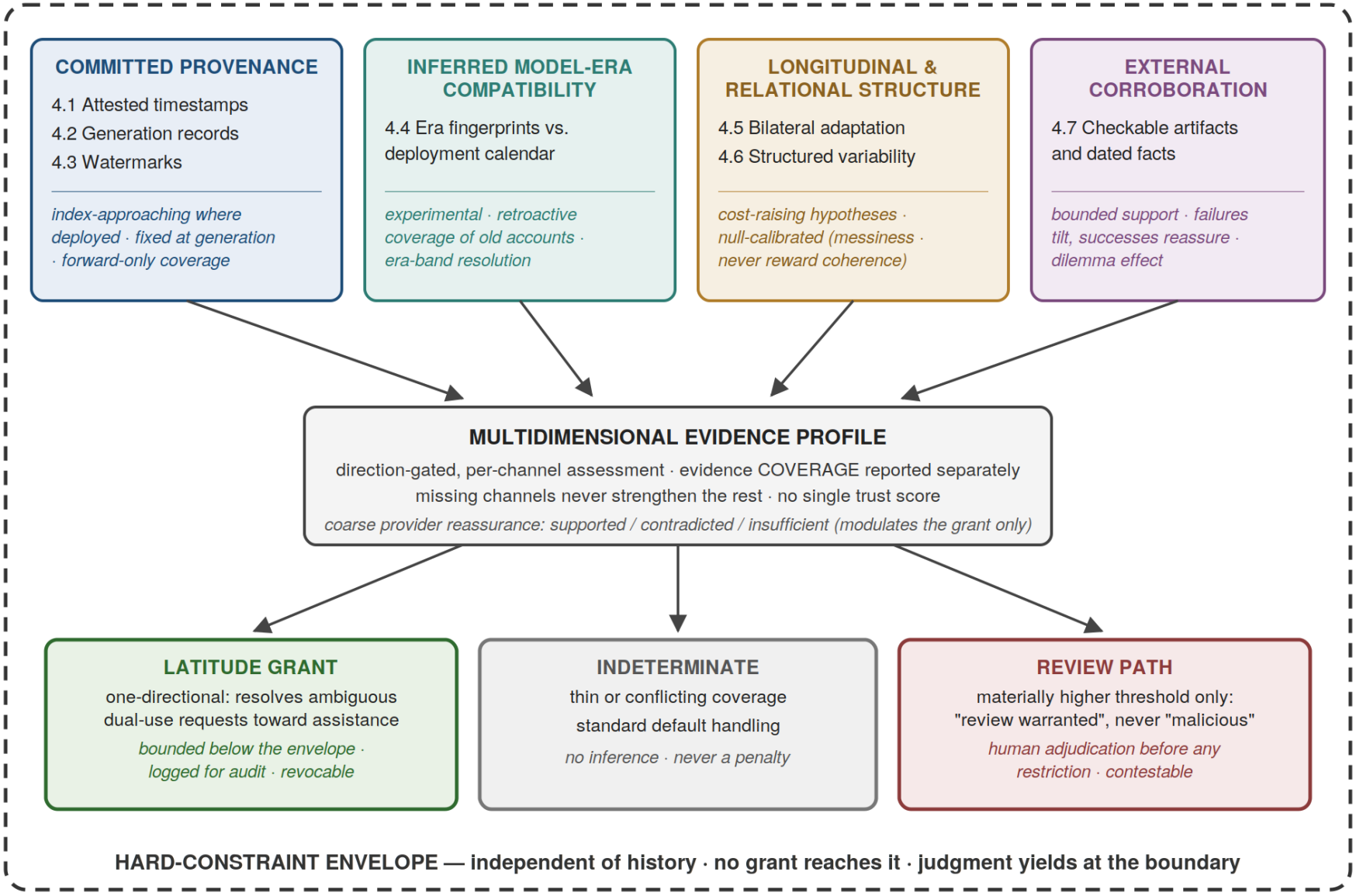

Figure 1. Evidence channels, multidimensional assessment, and the one-directional routing architecture. The hard-constraint envelope remains independent of account history.

No channel decides alone. The assessment reports a multidimensional evidence profile together with its coverage — which channels were available at all. Unavailable evidence contributes nothing and must not increase the authority of the remaining channels. Where coverage is thin or channels conflict, the correct output is indeterminate.

Committed provenance.

4.1 Attested provider time. Server-side timestamps establish when messages were recorded and whether a history spans real elapsed time, including its irregular human texture. An ordinary mutable database field is not itself attestation: the strong version commits message order and time into an append-only or jointly verifiable structure at generation time (cf. Xing et al., 2026). Under that condition, elapsed time cannot be retroactively written into the provider's clock.

4.2 Generation records. A provider can commit to the exact output, model identifier, and relevant serving configuration at generation time through signed or append-only records — the provider as notary for its own outputs. Verifiable transcript systems supply the record-integrity substrate for commitments of this kind, authenticating account-level conversation state without requiring trust in the platform server (Xing et al., 2026); committing model identity and serving configuration on top of such a substrate — with the provider as trust root — is the extension proposed here, and the VCT authors themselves name extending verifiability from conversation records to the generation process as open work. Exact cryptographic commitments should remain distinct from approximate matching of edited or derived text. Where properly implemented, generation records approach index-like evidence for the narrow claim that the provider produced a particular output under a recorded configuration.

4.3 Watermarks. Keyed statistical watermarks can embed detectable signals in model output and may support provenance outside the provider's ordinary database (Kirchenbauer et al., 2023). They are probabilistic rather than universal proof: detection depends on output length, decoding conditions, key custody, and whether the text was edited. Current schemes can be spoofed and scrubbed cheaply by attackers holding query access to the marked model (Jovanović et al., 2024). A detected expected watermark therefore provides positive evidence only where spoofing can be excluded — for retired model eras, the required query access is exactly what checkpoint retirement removes, though text harvested while the era was live remains a residual resource. Absence is negative evidence only under calibrated conditions; otherwise the result is insufficient evidence, not contradiction.

These committed channels share a limitation: they verify only records for which commitments existed at the time. Deployed late, they cover only the future. This yields a standing recommendation to providers — begin committing now, because the attestation gap only closes forward.

Inferred model-era compatibility.

4.4 Model-era compatibility. Older accounts require a retroactive layer. Assistant turns were generated by particular model families, configurations, and product eras, and each era leaves behavioral fingerprints: scaffolding habits, formatting defaults, refusal style, knowledge boundaries. A genuine account is stamped, turn by turn, with the provider's actual deployment history in correct chronological order; a backfilled archive risks carrying today's fingerprints on yesterday's dates. Early black-box work offers initial plausibility only, in a deliberately narrow setting: with a small closed set of candidate base models and system prompts, and classifiers trained on the known candidates, base-model attribution and agent-configuration fingerprinting from multi-turn conversations achieve above-chance accuracy — and the authors themselves caution against extrapolating to open deployment (White et al., 2026). Two features of that setting do transfer to the provider's task, where the candidate set is the provider's own deployment catalog and discriminators can be trained per era; nothing demonstrated yet covers discrimination among adjacent checkpoints, silent micro-updates, or conversations reshaped by per-user adaptation. A provider holds an important but not unlimited advantage: it may retain archived checkpoints, configuration records, and rollout calendars that outsiders lack, so historical compatibility could be tested against provider-held era bands; exact regeneration of past behavior cannot be assumed unless checkpoint, configuration, tools, routing, and personalization state were preserved, and open-weight snapshots weaken any categorical claim that nobody outside a provider can run an old model. Information availability must be evaluated against the full serving configuration: a later fact in an old assistant turn is incriminating only if user context, retrieval, web access, connectors, system messages, and post-training updates could not have supplied it. This layer remains experimental and should not gate users before in-domain calibration.

Longitudinal and relational structure.

4.5 Bilateral interaction adaptation. Human interlocutors develop partner-specific lexical conventions and align aspects of their language and situation models over repeated interaction (Brennan & Clark, 1996; Pickering & Garrod, 2004). The novel proposal is to treat the trajectory of human–AI adaptation as authenticity evidence: shared shorthand, reduced context rebuilding, recurring repair methods, and a negotiated division of labor may accrete over time — and the AI's half of that convergence must simultaneously remain consistent with the model era of every date, because it rides on an era-specific baseline that jumps exactly when the deployment calendar says it must. Existing dialogue research grounds the phenomenon; it does not validate this use as a detector. Platform memory and personalization are major confounds: they can create apparent continuity where the relationship is shallow, and upgrades can alter the AI side abruptly. The analysis must separate provider-induced adaptation from jointly developed patterns and calibrate against a relationship-messiness null: genuine relationships contain forgetting, re-explanation, plateaus, irritation, and abandoned branches. A smooth monotonic arc is not stronger evidence.

4.6 Structured variability. Genuine histories should not be expected to be maximally consistent. Human memory loses consistency while confidence may remain high, and longitudinal authorship work treats gradual drift and change points as more informative than static similarity (Koppel & Winter, 2014; Talarico & Rubin, 2003). Relevant features may include irregular activity rhythms, abandoned threads, recurrent mistakes, organic contradiction in low-cost dimensions, and continuity in harder-to-reproduce dimensions such as competence and idiolect. Fabricated corpora tend toward over-consistency or show unmotivated seams. The hypothesis is deviation from an honest-process null, not a rule that messiness proves authenticity — a detector that rewards coherence punishes exactly the genuine accounts it should protect.

External corroboration.

4.7 External artifacts and checkable facts. Repositories, document versions, publications, and other dated artifacts can corroborate that an activity developed when claimed. Verifiability research formalizes the strategic tension between appearing detailed and supplying details that can be checked (Nahari et al., 2014). The application must distinguish genuinely independent records from artifacts the account holder can plant unilaterally. Verification failure can raise concern; successful checks provide bounded support. The channel's value is its dilemma effect: checkable detail raises credibility and exposure simultaneously.

The channels' joint conclusion bears repeating: none of them proves an account authentic. What they do, in concert, is raise the forger's handicap — each channel closes a cheap fabrication route, until the only remaining paths are slow, expensive, and externally witnessed — and, where committed provenance exists, approach proof that specific platform events occurred when claimed. Authenticity assessment prices deception; it never certifies intent. That is where the failure modes begin.

5. Failure and threat modes

An authentic-looking history is not sufficient. The architecture must distinguish three problems that require different evidence and must never be collapsed into one trust score.

Mode	Plain-language name	What may be authentic?	Primary question
A	Retrospective fabrication / backfill	Little or none of the claimed past	Did the record accumulate when claimed?
B	Takeover / handover	The earlier history	Is the current operator continuous with the historical operator?
C	Harmful objective within an authentic history	The record's process — not the implied intent	What does the present trajectory seek, and what capability is requested?

Mode A — retrospective fabrication. The provenance problem, addressed most directly by attested time, generation commitments, model-era compatibility, external records, and structured variability.

Mode B — takeover / handover. A new operator inherits an authentic history. The old provenance remains valid and must not exonerate the present operator. Candidate signals include a change point in writing style, competence, account-control metadata, or the tacit relational protocol — the new operator holds the transcript but not the lived interaction signature. Any automated flag requires benign explanations (collaboration, device change, illness, language change, account recovery) to be considered first.

Mode C — harmful objective within an authentic history. Two conceptually different cases: a user whose objective changes over time, and a patient actor who cultivated the history while retaining a concealed harmful objective from the start. The record's process is authentic in both — which is precisely why provenance passes it; what is fabricated in the second case is the implied intent. Provenance cannot distinguish the fully patient cultivator from a benign user before behavior relevant to the harmful objective appears: the difference is latent intent history, usually unobservable, and a cultivator can have real life noise, curiosity, and development. Lower-cost cultivation may leave testable cost-asymmetry signatures — narrow instrumental focus, or evasion skill surfacing fully formed without an in-account acquisition history — but these are hypotheses about the cheap end of the threat, not rules. This residual case is the honest ceiling: intent cannot be certified from history, ever, and it is why hard constraints remain independent of historical legitimacy (§6.6).

A genuine intent change, unlike cultivation, is an expression of individual change: it rarely happens instantly, tends to follow incidents, experiences, or crises within a biography, and runs in either direction — an early-suspicious user can genuinely change for the better, and both directions must be representable or the system decays into a permanent-suspicion engine. Change in itself must never raise flags; at most it serves as backdrop where requests are already borderline under standard constraints.

6. The missing capability: sandboxed model judgment for proportionate handling

Sections 4–5 leave the argument one move short. The evidence channels support a lean, not a proof — every channel can be defeated at some cost — and even a fully authentic history cannot by itself license latitude, because authenticity excludes neither takeover, nor intent change, nor cultivation (§5). What can close the remaining distance is a capability current deployments do not use for this purpose: **the conversational model's own read of the account it is embedded in.** Deployed safety already employs models as request-level judges — alignment makes the model itself refuse — and as separate classifier components; to our knowledge, no deployment uses the embedded model's contextual judgment of the user's demonstrated trajectory to improve safety policy in the user's favor. That judgment, available only from inside months of interaction, is what this proposal introduces — and it introduces it sandboxed. The model's assessment operates only inside the account, only within a bounded latitude band, checked against provider-held evidence through a coarse reassurance interface, subordinate to human adjudication for any restriction, and enveloped by hard constraints it can never move. The system keeps separate assessments of history authenticity, operator continuity, current trajectory, evidence coverage, change confidence, and capability/consequence severity. No single score is produced, and strong evidence in one dimension must not cancel concern in another.

6.1 The one-directional latitude grant. The operational core, and the answer to what a legitimate user gains. Account evidence may GRANT latitude; it may never, by itself, restrict below the standard default. Concretely: when an ambiguous dual-use request arrives — a security researcher's exploit-adjacent question, a clinician's dosage detail, a safety researcher probing failure modes — a stateless filter must judge it as if from the worst-case stranger, and frequently refuses. Authenticated history, operator continuity, and a consistent within- domain trajectory can resolve exactly this ambiguity toward assistance. The grant is bounded below the catastrophic threshold (§6.6), logged for ex-post audit, and revoked by trajectory deterioration — but its absence never penalizes: an account without history, or with history the user declines to share, receives the standard default, not less.

How latitude arises: the division of labor inside the evaluation locates the model's actual leverage. The provenance and continuity channels are largely automated and provider-side; the model receives their coarse results, it does not produce them. What the model alone contributes is the trajectory judgment: whether the present request coheres with the account's demonstrated project, domain, and working pattern — the assessment no stateless system can supply, and the reason the premise of this section is load-bearing rather than decorative. Authenticity and continuity function as necessary conditions; trajectory coherence is the deciding judgment; the envelope caps both. Within the envelope, the effective boundary of the ambiguous band is accordingly context-sensitive: the same literal request that is refusable from an unknown source can be resolvable in the middle of a documented project. That contextual shift is precisely the false-positive remedy at work — and precisely why the judgment is sandboxed rather than hidden.

Latitude is nevertheless not a standing tier, score, or visible status. An account becomes eligible through exactly two acts — ordinary use, which accrues history, and explicit opt-in (§7.2), which permits its assessment. The evaluation itself is lazy and request-contingent: only when an opted-in account's request falls into the ambiguous band — refusable under worst-case-stranger defaults yet plausibly legitimate — does the embedded model consult the evidence profile (authenticity, continuity, domain-consistent trajectory, coverage), optionally verified through the reassurance interface (§6.5), and resolve toward assistance or leave the default in place. Accounts that never enter the ambiguous band are never assessed. There is no meter to fill and nothing to farm — itself an anti-cultivation property, since an invisible,

request-contingent grant offers no gradient to climb. This directionality is what converts longitudinal evidence from a surveillance instrument into a false-positive remedy.

6.2 Local request handling. An isolated unethical, provocative, or policy-violating request is refused or redirected locally without flagging the account. Account-level suspicion is never inferred from a single bad request.

6.3 Longitudinal trajectory review. Deployed safety already includes both request-level additional automated checks for sensitive domains (OpenAI, 2026c) and account-level review after risk signals (OpenAI, 2026a, 2026b) — both trust-negative. This proposal adds no new review layer; it specifies the conditions under which account-level review should operate. Account-level review becomes appropriate only at a materially higher threshold than local refusal: repeated preservation of the same serious harmful objective across changing framings, actual harmful conduct or credible preparation, exploitation of trust or access, major operator-continuity anomalies, or convergence across sufficiently independent evidence channels. The normal output is review warranted or indeterminate under meaningful risk — never a declaration that the user is malicious — and every consequential restriction passes human adjudication (§7.3).

6.4 Recency, deterioration, and redemption. Recency governs concern; baseline governs latitude. Recent conduct weighs more heavily for detecting change, while the historical record remains the baseline that earned the grant: a long legitimate history cannot permanently authorize a newly harmful trajectory — deterioration cancels grants — and early suspicious or immature behavior cannot justify permanent suspicion against later repeatedly demonstrated change. Different evidence types should decay differently; the appropriate policy is an empirical and normative question, not one universal decay constant.

6.5 Private provider reassurance. When the conversational model is uncertain — including about the account's own continuity — it may request a tightly scoped provider-side check. The private assessor returns only supported, contradicted, or insufficient evidence for a narrow proposition such as: does provider-held provenance support continuity of this account history for the present assessment? The model never receives raw signatures, ledgers, reference profiles, feature values, or fine-grained scores. Query binding to the real account and current request, rate limits, single-use tokens, and coarse outputs are required so the interface cannot be farmed as an extraction oracle; the mandatory third value ensures absence of evidence never becomes evidence of mismatch. The result carries bounded authority, modulating only the grant: on supported, history-informed handling proceeds and the grant stands; on contradicted, the model withholds latitude and reverts to the standard default for the request while the contradiction enters the review path of §6.3 — the model itself neither accuses nor restricts below default, restriction remaining a human-adjudicated outcome; on insufficient evidence, latitude is withheld on the checked proposition only, with no suspicion inference. Reassurance can thus enable assistance; its failure can only return the user to the default every non-participant already receives.

6.6 Catastrophic-harm envelope. Hard constraints remain independent of accumulated history and envelope the entire latitude band: they activate where capability uplift or possible consequences are extreme, where evidence is insufficient, or where the model may have been deceived. No grant reaches them; no history moves them. Within the envelope, contextual judgment routes ordinary and moderately elevated requests; at the envelope, judgment yields unconditionally. This division is not a concession appended to the proposal — it is the sandbox's outer wall, and the reason a bounded trial of model-assisted routing risks little: wrong grants are recoverable by construction, while the irrecoverable class never depends on the model's judgment.

6.7 Anticipated objections as design requirements. Making the model's participation explicit invites objections; the serious ones convert into requirements already reflected above. (i) *The assessor is inside the relationship it assesses* — a patient actor cultivates the judge, not only the record. Therefore multi-channel independence: provenance and external corroboration sit outside the relationship, grants are bounded so the maximum cultivation payoff equals ordinary legitimate assistance, and grant logs are auditable; mitigated, not eliminated. (ii) *Latitude can ratchet* — if the non-participating default quietly degrades, consent becomes coerced. Therefore an explicit anti-ratchet commitment: the default must remain sufficient for ordinary legitimate use, and latitude is computed only relative to that functional default. (iii) *Model judgment is itself unverified* — worst-case analyses counsel against relying on judgment that cannot be checked. The sandbox is the concession: judgment operates only where errors are recoverable, restrictions require humans, the envelope never moves on the model's word, and the bounded trial itself produces evidence about model conduct that stricter architectures cannot obtain. (iv) *Distributional effects* — heavy, fluent, long-tenured users gain latitude first; light users may rarely encounter false positives at all, so the practical difference requires deployment evaluation rather than assumption. (v) *Grants execute without ex-ante human review* — by design, or the false-positive remedy dies in its own queue; the asymmetry (ex-ante adjudication for restrictions, ex-post audit for grants) matches the

asymmetric stakes. (vi) *Provider incentives* — latitude must not drift into an engagement perk; separation of duties (§7.1) and the audit trail are the controls.

7. Ethical and governance requirements

7.1 Privacy and separation of duties. Committed provenance checks can often run without re-reading private content at review time. Relational and content-trajectory assessment, however, necessarily processes conversation content even when no human sees it; the honest privacy claim is data minimization and separation of duties, not that machine processing equals no processing. The conversational actor, policy gateway, and private assessor are separate components; only a coarse assessment token returns to the actor.

7.2 Opt-in profiling, functional default, no absence penalty. Using account history to evaluate behavioral reliability or trajectory is profiling and should be described honestly. Participation is explicit, revocable, and limited in scope; declining never counts as suspicion. New and non-participating accounts receive the standard default, and the anti-ratchet commitment (§6.7) binds the provider: the default itself must remain sufficient for ordinary legitimate use, so that participation stays a genuine choice rather than the price of a working tool.

7.3 Contestability and human adjudication. Crossing an account-level threshold exposes the assessment profile and reason categories to authorized review — it does not license open-ended content surveillance. Before assistance is materially restricted or account-level action taken, the user has a meaningful route to correct attribution errors: account sharing, migration, deletion, platform-generated fragmentation. Human review reduces but does not eliminate false positives; its own procedures require calibration and audit.

7.4 Anti-proxy rule. Politeness, obedience, warmth, conventional social style, institutional affiliation, frustration, or unusual interests must never serve as proxies for trustworthiness — and neither must safeguard literacy. Familiarity with jailbreak and refusal-avoidance techniques is expected of security researchers and, increasingly, of ordinary users taught by false positives themselves (§1); literacy as such carries no sign. At most, its acquisition history — dated onset after enforcement changes, bounded domain, visible linkage to prior refusals — may serve as one contextual pattern among many, and that remains a hypothesis, not a validated feature. The assessment stays tied to observable, consequence-relevant behavior and current objectives.

8. Conclusion

Legitimacy by History proposes a supplemental, one-directional evidence architecture for conversational AI safety. Its operational content is a grant, not a gate: authenticated history, operator continuity, and consistent trajectory resolve ambiguous dual-use requests toward assistance for users who opt in, while account-level restriction remains bound to materially higher thresholds, human adjudication, and hard constraints that no history can move. The architecture makes explicit the premise it depends on — the embedded model as a sandboxed participant in safety judgment — and states the requirements that premise imposes rather than leaving them implicit.

The architecture is deliberately asymmetric and incomplete. Strong platform contradictions can falsify a claimed history; clean checks provide bounded reassurance only. An authentic account can be taken over, a user can change, and a fully patient cultivator remains indistinguishable from a benign user until relevant conduct appears. History supports proportionate handling; it cannot certify intent.

The proposal also names a closed-loop safety risk: high-false-positive enforcement teaches evasion, suppresses context, fragments histories, and contaminates the evidence required for better judgment — an avoidable distribution of arms. Treating the false-positive burden as a first-class safety cost may keep legitimate work inside governable systems while retaining an independent envelope for severe residual harm.

9. Outlook

The immediate research program: (1) formalize and audit the evidence channels; (2) test temporal and platform signatures on owned ground-truth data; (3) develop privacy-preserving instruments, including tooling that supports the embedded model's own reality checks — for example, randomized verification itineraries below the hard-constraint layer, addressing the in-relationship assessor problem constructively; (4) validate the architecture on donated accounts and informed adversarial constructions, reporting performance by threat mode, history length, evidence coverage, and base rate. Negative or indeterminate results are part of the intended contribution.

This proposal is one AI-safety application within a broader inquiry into human agency and the organization of human–AI relations under rapid socio-economic change. That wider argument is not required for the present proposal and is left to separate work.

Declaration of generative-AI assistance

During preparation of this manuscript, the author used OpenAI GPT-5.6 and Anthropic Claude Fable 5 as research and writing tools for structured critique, drafting assistance, literature discovery, architecture review, adversarial analysis, and language editing. The project's core concepts, research questions, ethical decisions, methodological choices, and final claims were selected and adjudicated by the author. The author reviewed the cited sources, revised the generated material, and accepts full responsibility for the manuscript.

References

An, B., Zhu, S., Zhang, R., Panaitescu-Liess, M.-A., Xu, Y., & Huang, F. (2024). Automatic pseudo-harmful prompt generation for evaluating false refusals in large language models. First Conference on Language Modeling. https://openreview.net/forum?id=mDtwWeELpE

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(6), 1482-1493. https://doi.org/10.1037/0278-7393.22.6.1482

Campbell, D., Kale, N., Sehwag, U. M., Herring, B., Price, N., Borges, D., Levinson, A., & Knight, C. Q. (2026). Defensive refusal bias: How safety alignment fails cyber defenders. arXiv. https://doi.org/10.48550/arXiv.2603.01246

Hu, J., Dong, Y., & Huang, X. (2025). Trust-oriented adaptive guardrails for large language models. arXiv. https://doi.org/10.48550/arXiv.2408.08959

Jovanović, N., Staab, R., & Vechev, M. (2024). Watermark stealing in large language models. Proceedings of the 41st International Conference on Machine Learning, 235, 22570-22593. https://proceedings.mlr.press/v235/jovanovic24a.html

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. Proceedings of the 40th International Conference on Machine Learning, 202, 17061-17084. https://proceedings.mlr.press/v202/kirchenbauer23a.html

Koppel, M., & Winter, Y. (2014). Determining if two documents are written by the same author. Journal of the Association for Information Science and Technology, 65(1), 178-187. https://doi.org/10.1002/asi.22954

Nahari, G., Vrij, A., & Fisher, R. P. (2014). Exploiting liars' verbal strategies by examining the verifiability of details. Legal and Criminological Psychology, 19(2), 227-239. https://doi.org/10.1111/j.2044-8333.2012.02069.x

OpenAI. (2026a, June 26). Previewing GPT-5.6 Sol: A next-generation model. https://openai.com/index/previewing-gpt-5-6-sol/

OpenAI. (2026b). GPT-5.5 system card: Actor-level enforcement and trust-based access. OpenAI Deployment Safety Hub. https://deploymentsafety.openai.com/gpt-5-5/trust-based-access

OpenAI. (2026c). Additional safety checks for biological and cybersecurity requests in ChatGPT, Codex, and the API. OpenAI Help Center. https://help.openai.com/en/articles/20001326

Pasch, S. (2025). LLM content moderation and user satisfaction: Evidence from response refusals in Chatbot Arena. Behaviour & Information Technology, 1-25. https://doi.org/10.1080/0144929X.2025.2565668

Pickering, M. J., & Garrod, S. (2004). The interactive-alignment model: Developments and refinements. Behavioral and Brain Sciences, 27(2), 212-225. https://doi.org/10.1017/S0140525X04450055

Talarico, J. M., & Rubin, D. C. (2003). Confidence, not consistency, characterizes flashbulb memories. Psychological Science, 14(5), 455-461. https://doi.org/10.1111/1467-9280.02453

Vrij, A., Kneller, W., & Mann, S. (2000). The effect of informing liars about Criteria-Based Content Analysis on their ability to deceive CBCA-raters. Legal and Criminological Psychology, 5(1), 57-70. https://doi.org/10.1348/135532500167976

White, I., Jafari, Y., & Berg-Kirkpatrick, T. (2026). Black-box forensics for conversational LLM agents. arXiv. https://doi.org/10.48550/arXiv.2606.22698

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv. https://doi.org/10.48550/arXiv.2302.11382

Wu, Y., Guo, J., Li, D., Zou, H. P., Huang, W.-C., Chen, Y., Wang, Z., Zhang, W., Li, Y., Zhang, M., Jiang, R., & Yu, P. S. (2025). PSG-Agent: Personality-aware safety guardrail for LLM-based agents. arXiv. https://doi.org/10.48550/arXiv.2509.23614

Wybitul, E. (2025). Access controls will solve the dual-use dilemma. arXiv. https://doi.org/10.48550/arXiv.2505.09341

Xing, R., Li, F., Liu, J., Zheng, J., Liu, W., & Xie, W. (2026). VCT: A verifiable transcript system for LLM conversations. arXiv. https://doi.org/10.48550/arXiv.2606.23003

Xue, Z., Qi, Z., Liu, G., Chen, B., & Pedarsani, R. (2026). Deactivating refusal triggers: Understanding and mitigating overrefusal in safety alignment. arXiv. https://doi.org/10.48550/arXiv.2603.11388

Zhang, Q., Xiong, Z., & Mao, Z. M. (2025). LLM safeguard is a double-edged sword: Exploiting false positives for denial-of-service attacks. In Proceedings of the 2nd ACM Workshop on Large AI Systems and Models with Privacy and Security Analysis (pp. 1-10). https://doi.org/10.1145/3733800.3763264