Artificial Intelligence Is Not an Intern: Authoritative Prompt Parenting for the Supervision of Generative AI by Engineers, Scientists, and Health Professionals

Daniel B. Oerther,[*,+,1] Sarah Oerther,[2] and Venkataramana Gadhamshetty[+,3,4]

1. Department of Civil, Architectural, and Environmental Engineering, Missouri University of Science and Technology, Rolla, Missouri, USA.

2. School of Nursing, Barnes-Jewish College, St. Louis, Missouri, USA.

3. Civil and Environmental Engineering, South Dakota School of Mines and Technology, Rapid City, South Dakota, USA.

4. 2D-materials for Biofilm Engineering Science and Technology (2D-BEST), South Dakota School of Mines and Technology, Rapid City, South Dakota, USA.

*** Corresponding author:** Department of Civil, Architectural, and Environmental Engineering, Missouri University of Science and Technology, 1401 N. Pine Street, Rolla, MO 65409, USA. E-mail: oertherd@mst.edu; Phone: +1 (573) 465-7714.

\+ Member of AEESP.

Abstract

Generative artificial intelligence does not hold professional responsibility, and it may not reliably preserve user instructions over extended interactions. In response, we present "authoritative prompt parenting," as a supervisory heuristic drawn from developmental science. The approach comprises four standing-instruction controls, including: tiered rules; a visible tripwire; a counter-indexed heartbeat; and a traverse/paraphrase read-receipt that tests whether designated instruction fragments are accessible and whether core rules can be restated. In a point-in-time paired demonstration using four Large Language Model (LLM) systems, the specified initial receipt was retrieved from the installed instruction file, and paired controls did not. While the approach does not confirm whole-file attention, human-like comprehension, or continued

adherence, the approach does prompt human verification – and a conservative restart when a check fails – to support human professional accountability.

Keywords: generative artificial intelligence; environmental engineering; large language models; professional responsibility; prompt engineering; supervision

Introduction: from calculation to supervision

A recent commentary in this journal argued that generative artificial intelligence (AI), delivered through widely available large language models (LLMs), is not a calculator but rather a mediator of the relationship of sharing data, information, knowledge, and wisdom (DIKW) among people (Oerther et al., 2026). If AI is a relationship rather than a calculation, the next question as professionals asks is not "how do I compute with it?" but rather "how do I govern my side of the relationship?"

Prompt engineering is a common approach to managing our relationship with AI. Prompt engineering includes designing and structuring instructions so that an LLM produces accurate, relevant, and contextually appropriate output (Gadesha, 2025). Prompt engineering is taught as a core competency in higher education (Lee and Palmer, 2025) to help LLM users formulate effective instructions. In this Short Communication, we argue that formulation is necessary but not sufficient because professional practice requires an explicit supervisory layer for checking whether consequential instructions remain available and behaviorally effective across a working session. We propose the better analogy is not engineering but parenting – specifically, authoritative prompt parenting.

Our analogy requires careful framing. One popular approach to prompt engineering treats the LLM as an intern being onboarded (Mollick, 2024), but recent research cautions against treating AI agents like humans at all (Kropp et al., 2026). While we note that interns arrive on the job with formed judgment, persistent memory, and professional accountability – growing towards independence – the current available LLMs are more similar to infants. LLMs have vast capability without formed professional judgment, and no professional accountability can be attached – all while possessing capabilities no infant possesses. Without judgement or

accountability, an infant is parented, not onboarded, and the literature shows that supervisory acts should be set by the child's developmental level (Oerther and Oerther, 2021). Among parenting styles described by Baumrind (1966) and later organized on axes of demandingness and responsiveness (by Maccoby and Martin, 1983), the authoritative style – with high structure and clear expectations, delivered with reasons and coupled to active checking – is the one associated with the best developmental outcomes.

As described by Baumrind, "The authoritative parent attempts to direct the child's activities in a rational, issue-oriented manner… [and] affirms the child's present qualities but also sets standards for future conduct" (Baumrind, 1966, p. 891). Our approach leverages the demandingness half of that construct – structure, monitoring, and checking – with responsiveness appearing as reasons given alongside rules and as non-punitive responses to failures.

The problem is supervision, not phrasing

An LLM may drift across the many turns of a conversation. It may state a falsehood in the same confident register it uses for a truth. And it may fail to apply, on a later response, a rule it followed earlier. Recent research documents that a model frequently restates a constraint it is simultaneously violating (Kruthof, 2026). In other words, an LLM's assurance that it is following the rules is not evidence that it is. And professionals currently agree that no model can be held accountable. For example, the nurse remains responsible "even in instances of system or technology failure" (American Nurses Association, 2022; Oerther and Oerther, 2026b), and the engineer's first canon is to hold paramount the safety, health, and welfare of the public – a nondelegable responsibility (NSPE, 2019). The practical need is therefore not a better one-time prompt but rather continuous, checkable supervision of a capable, unreliable, child-like collaborator. Our approach – authoritative prompt parenting – uses a small set of standing-instructions that produce indicators the availability of instructions and make behavior observable (rather than assumed).

Four controls

Several widely used products provide a project-level container – standing instructions, project files, or custom instructions – that may influence responses within a project. We note that implementations vary across products – containers are bounded by context and file limits, long

files may be truncated or retrieved in part rather than read whole, and attention across a long input is not uniform (Liu et al., 2024). We treat each product as a separate implementation, and we propose structuring user-defined instructions around four controls (Figure 1), which together operationalize the professional duties of transparency, accountability, and verifiability (Oerther and Oerther, 2026b).

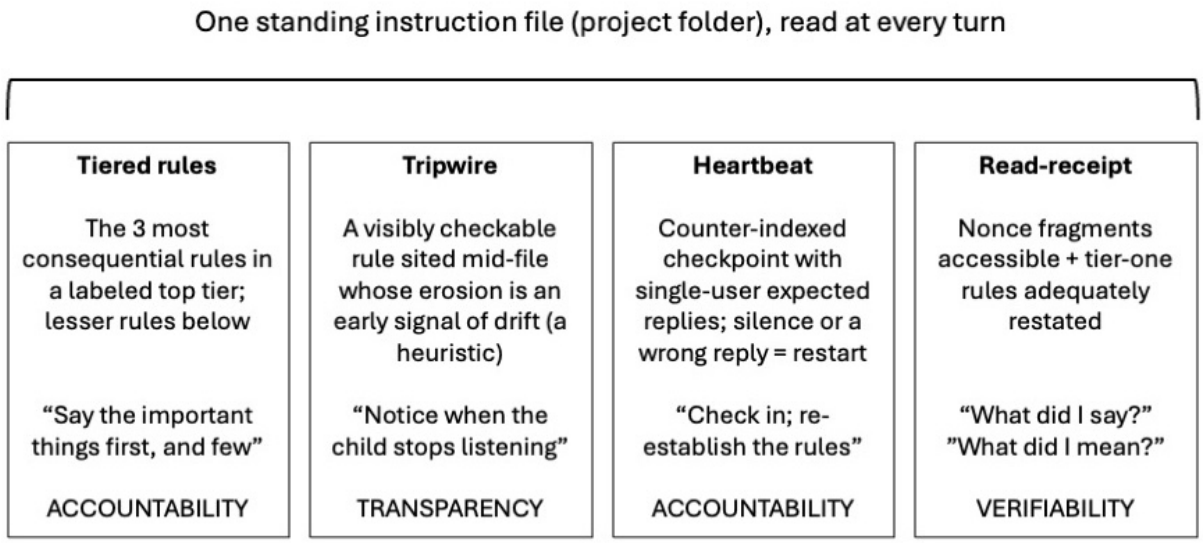

Figure 1. *Authoritative prompt parenting in four controls.* Standing instructions given to a generative AI are structured so that selected indicators of instruction availability and behavior become observable rather than assumed, including: tiered rules concentrate the few most consequential instructions; a mid-file tripwire provides a visible early signal of drift; a counter-indexed heartbeat requests single-use checkpoint replies; and a traverse/paraphrase read-receipt separates fragment access ("what did I say?") from adequate restatement ("what did I mean?"). Together, the controls operationalize transparency, accountability, and verifiability. The runnable instruction file is provided as Supplementary Protocol S1.

Tiered rules. Because attention is finite and, in our experience, weakens across a long file, the few most consequential rules are placed in a short, explicitly labeled top tier – as a design heuristic, no more than three – with lesser rules below.

A tripwire. One rule whose violation is visually obvious – in our example, "write in plain language and define any term you introduce" – is placed near the middle of the file, because prior long-context experiments have reported weaker use of information in that region (Liu et al., 2024). Whether this indicator consistently precedes other instruction failures remains to be evaluated, and therefore, we offer it as a practical heuristic, not an empirically validated early-

warning device. The value of the tripwire is that erosion is easy to see. In our example, when the prose stops being plain, the user is warned that adherence may be decaying (i.e., without a further audit).

A heartbeat. The instructions direct that on a trigger word – in our example, "hello," chosen as the "hello world" of governance checks, though less common tokens would likely reduce accidental firing – the model re-read the file and return a fixed response. Because that first response places its own content into the conversation history, an unchanged trigger should not be reused as evidence (i.e., a later correct reply could be copied from history rather than re-derived from the file). The heartbeat therefore carries a counter ("hello-2," "hello-3"), and the file specifies a distinct, single-use expected response for each count, so that no checkpoint reply can be simply copied from the conversation history (Supplementary Protocol S1). Although likely modest, each requested checkpoint consumes context, so we recommend judicious placement of checkpoints; for example, at session start, after a long retrieval, or before a consequential output, rather than at every turn. Silence or a wrong response is strong evidence the session should be restarted rather than trusted, and the null result is itself the signal.

A traverse/paraphrase read-receipt. The session-start response should not be viewed as a greeting, but rather a two-part check. First, a user-generated nonce – a random token used in an authentication protocol – is split into fragments and scattered from the top of the file to the bottom. Returning the reassembled string is evidence that the fragments were accessible at that moment. Second, the model must paraphrase the tier-one rules, which provides the user an opportunity to confirm restated rules with operationally adequate meaning (i.e., not verbatim). This read-receipt is based on two familiar questions of parenting, namely: "what did I just say?" – a test of reception, answered by the fragments – and "what did I mean?" – a test of restatement, answered by the paraphrase. The two can fail independently, and Table 1 provides our suggested response to each pattern.

What the receipt can and cannot show. Labeled fragments are themselves retrieval-targetable (i.e., a retrieval layer could locate exactly the marked lines while ignoring the rest of the file), and therefore, a passed receipt evidences access to the marked fragments, but not that "the whole file was read." Practitioners hardening the check could embed unmarked fragments within the prose of widely separated rules or require a reply derivable only from unmarked content – the

count of rules in the file, or the first word of each tier heading (Supplementary Protocol S1) (i.e., similar to the digital watermark approach announced by Claude in mid-August 2026 (Roeloffs, 2026). Even using such a hardened approach, a prompt-level check will – at most – provide evidence of access to distributed regions of a file (i.e., presence in context does not guarantee attention, and restatement does not guarantee adherence) (Liu et al., 2024; Kruthof, 2026). It is important to note that the controls do not make the model reliable, but rather they make selected indicators of its unreliability visible to the accountable human.

Table 1 summarizes the receipt outcomes. The cause of a failure is not uniquely identifiable from its pattern (i.e., a garbled reply might reflect truncation, retrieval behavior, or generation variability). Therefore, every failure is read as a conservative stop signal rather than a diagnosis. Restarting is a low-cost conservative action, not a costless one.

TABLE 1. *Receipt outcomes and conservative responses.*

Receipt fragments	**Rule restatement**	**Conservative interpretation**
Adequate	Adequate	Check passed at this moment; establishes nothing about subsequent adherence.
Adequate	Inadequate	Fragments accessible but restatement drifted; re-establish the rules before proceeding.
Inadequate	Adequate	Mixed result; cause uncertain; do not rely on the check.
Inadequate	Inadequate	Check failed; restart the session rather than trust the work.

What is, and is not, novel

Authoritative prompt parenting leverages existing structured instructions. For example, vendor's provide an opportunity to input standing instructions, and prompting techniques are well catalogued (Schulhoff et al., 2024). Techniques such as Chain-of-Verification and SelfCheckGPT verify the answer a model gives, but they do not confirm whether its governing instructions remain live (Dhuliawala et al., 2024; Manakul et al., 2023). Among the literature we identified, DriftBench is the closest directly relevant evaluation. Kruthof (2026) reported a LLM often correctly restates constraints it is simultaneously violating. Furthermore, checkpointing reduced but did not eliminate the dissociation (Kruthof, 2026). Those results are precisely why our approach pairs the restatement receipt with a behavioral tripwire rather than trusting either alone. We offer this as a bounded search result rather than a claim of priority (i.e., prompt-level

ideas frequently appear first in repositories and forums, and we would welcome notice of earlier work).

A point-in-time demonstration

In a point-in-time paired demonstration on each of four products exposing a standing-instruction container – Claude (Fable 5, High), ChatGPT (5.6-Sol, Pro), SuperGrok (Grok 4.5, Fast), and Gemini Notebook – with persistent memory disabled, firing the trigger when the instruction file was installed produced the specified initial receipt, whereas the paired control without the file reverted to a platform-specific default (i.e., a greeting, a "welcome," or a summary of the folder's single document). Although we used "hello" as our trigger, the paired control shows that the receipt is contingent on the installed file rather than on default courtesy. We reiterated the limitations, namely the demonstration does not establish: attention to unmarked portions of the file; continued compliance across a session; reliability; or failure rates. Furthermore, our results are point-in-time and may change with model updates. A product that does not expose a standing-instruction container has nothing for the file to attach to, which marks the boundary of the method. The complete instruction file and test method are provided as Supplementary Protocol S1 and Supplementary Methods S2.

A worked failure parenting catches

Consider a routine task in environmental practice, such as a professional asks a LLM whether a stream's dissolved oxygen (DO) record supports trout. Ungoverned, the model may return a single, confident, threshold reconstructed from memory (i.e., 7 mg/liter). But trained professionals know that the applicable criteria do not reduce to a single number, namely: the U.S. EPA's ambient water quality criteria for DO specify duration-dependent means and minima that differ by salmonid life stage, and the governing value in practice may be a jurisdiction's adopted standard (USEPA, 1986). Under our example of a tier-one rule that every claim be labeled by its source, our use of the LLM should distinguish the measured record, any computed statistic, a cited criterion, and its own inference. Furthermore, the response must flag a recalled threshold as recalled, and this result should remind the trained professional – and help to support the emerging professional – to check the applicable standard before the number reaches any report.

The control does not make the model correct, but rather it makes the LLM's uncertainty visible to the human who holds responsibility.

Growth in the parent

A careful reader will object that a child develops and a model does not mature within the relationship – where products do retain project context, what persists is the container we describe, not formed judgment. But while the LLM does not develop, our analogy of parenting should hold for the human-side of the relationship. Parenting practice is passed along – grandparents guiding parents who guide children – as part of a family legacy in which tacit values and behaviors shape identity (Oerther and Papachrisanthou, 2024). In our approach, the human user does develop. Specifically, we anticipate supervisory practice improves with use, and users may pass that practice to colleagues, emerging professionals, and students. The controls are teachable in a way that ad hoc prompting is not.

An obligation, not a technique

The engineer who signs and seals, the nurse accountable for practice through system failure, and the scientist who stakes a professional reputation on a result cannot delegate their own duty to an LLM. Treating the model as an intern understates the supervision the situation demands. Treating it as an infant – with authoritative structure, reasons given alongside rules, a visible tripwire, and checkpoint evidence that instructions remain live – matches the supervision to the tool. Ultimately, there is a formation argument, namely: environmental engineering and science has been described as a caring profession that teaches "why" as well as "what" and "how" (Oerther and Peters, 2020). If emerging professionals and students learn only to prompt engineer, they incorrectly learn to write a specification and trust the tool. If, instead, we teach authoritative prompt parenting, the next generation learns to supervise an unreliable dependent with structure, patience, and verification, which is a fair description of professional judgment itself, and the natural continuation of understanding AI as a mediator of knowledge rather than a calculator (Oerther et al., 2026).

Article type

Short Communication

Acknowledgements

None.

Author contributions

D.B.O. conceptualized the Short Communication, developed the "prompt parenting" framework, and wrote the original draft. S.O. provided the developmental-science and parenting-theory framing and critical revisions. V.G. provided the environmental engineering and AI-systems context and critical revisions. All authors reviewed, edited, and approved the final version.

AI statement

Anthropic Claude (Fable 5, with High effort selected) was used as a drafting partner during manuscript preparation; the authors directed, reviewed, and revised all AI-assisted text. After the authors completed a full draft, the manuscript was submitted to OpenAI ChatGPT (5.6-Sol, with Pro effort selected) for an independent editorial critique using the following prompt, reproduced verbatim as issued: "Assume the role of Professor Ramana Gadhamshetty, the editor in chief of Environmental Engineering Science (https://journals.sagepub.com/author-instructions/EEN). Evaluate the attached manuscript for inclusion in the journal. Be sure to consider Aims and Scope as well as Submission Guidelines, including quality and format of references. Before providing your final evaluation, perform a harsh self-critique considering hallucinations, bias, tone, etc. Use the critique to improve your final evaluation. Before you begin your review, do you have any questions to ask or assumptions to clarify using a human-in-the-loop approach?" The authors reviewed the critique, adopted, rejected, or adjudicated each of its recommendations, and revised the manuscript accordingly. The authors take full responsibility for the content of the published article.

Author verified TL;DR: "AI is not a calculator, and it is not an intern either – it has capability without judgment, memory, or accountability, like an infant. So supervise it the way developmental science says to raise a child well: authoritatively. Put your rules in the project folder in tiers, add a visible tripwire, and use a counter-indexed check-in that shows on demand that the rules are still accessible ("what did I say?") and can be restated ("what did I mean?").

The model is never professionally accountable; the professional's duty is nondelegable." (as recommended by Oerther S, Oerther DB. Every publication needs an author verified TL;DR statement. Nurse Educ Pract 2026;91:104674; doi: 10.1016/j.nepr.2025.104674).

Declaration of conflicting interests

The authors disclose that the Editor-in-Chief of Environmental Engineering Science, Professor Ramana Gadhamshetty, is a co-author on this submission. Consistent with journal policy, the submission is to be managed independently by an alternative member of the editorial board, without the involvement of Professor Gadhamshetty in review or decision-making. The authors declare no other potential conflicts of interest with respect to the research, authorship, or publication of this article.

Funding statement

The authors received no financial support for the research, authorship, or publication of this article.

Ethical approval and informed consent statements

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability statement

No research dataset was generated. The minimal runnable protocol and the point-in-time test method are supplied as Supplementary Protocol S1 and Supplementary Methods S2.

References

American Nurses Association. The Ethical Use of Artificial Intelligence in Nursing Practice [position statement]. Silver Spring, MD: ANA Center for Ethics and Human Rights; 2022.

Baumrind D. Effects of authoritative parental control on child behavior. Child Development 1966;37(4):887–907; doi: 10.2307/1126611

Dhuliawala S, Komeili M, Xu J, et al. Chain-of-Verification reduces hallucination in large language models. In: Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand; 2024; pp. 3563–3578; doi: 10.18653/v1/2024.findings-acl.212

Gadesha V. 2025. Prompt engineering techniques. Think [blog]. Available from: https://www.ibm.com/think/topics/prompt-engineering-techniques [Last accessed: August 9, 2026].

Kropp M, Bedard J, Wiles E, Hsu M, Krayer L. 2026. Research: Why you shouldn't treat AI agents like employees. Harvard Business Review [trade publication]. Available from: https://hbr.org/2026/05/research-why-you-shouldnt-treat-ai-agents-like-employees [Last accessed: August 9, 2026].

Kruthof G. Models recall what they violate: Constraint adherence in multi-turn LLM ideation. arXiv 2026;2604.28031 [preprint]; doi: 10.48550/arXiv.2604.28031

Lee D, Palmer E. Prompt engineering in higher education: A systematic review to help inform curricula. Int J Educ Technol High Educ 2025;22(7); doi: 10.1186/s41239-025-00503-7

Liu NF, Lin K, Hewitt J, et al. Lost in the middle: How language models use long contexts. Trans Assoc Comput Linguist 2024;12:157–173; doi: 10.1162/tacl_a_00638

Maccoby EE, Martin JA. Socialization in the context of the family: Parent–child interaction. In: Mussen PH (Series Ed), Hetherington EM (Vol Ed). Handbook of Child Psychology: Vol. 4. Socialization, Personality, and Social Development. 4th ed. New York: Wiley; 1983; pp. 1–101.

Manakul P, Liusie A, Gales MJF. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP). Singapore; 2023; pp. 9004–9017; doi: 10.18653/v1/2023.emnlp-main.557

Mollick E. Co-Intelligence: Living and Working with AI. New York: Portfolio/Penguin; 2024.

NSPE. Code of Ethics for Engineers. Alexandria, VA: National Society of Professional Engineers; 2019.

Oerther S, Papachrisanthou MM. Ways parental health and stress shape the parenting of preteens in the rural Ozark Mountains. Family Relations 2024;73(1):171–192; doi: 10.1111/fare.12976

Oerther S, Oerther DB. Review of recent research about parenting Generation Z pre-teen children. West J Nurs Res 2021;43(11):1073–1086; doi: 10.1177/0193945920988782

Oerther S, Oerther DB. The importance of transparency, accountability, and verifiability when using artificial intelligence in nursing education. Nurse Educ Pract 2026b;90:104557; doi: 10.1016/j.nepr.2025.104557

Oerther DB, Oerther S, Gadhamshetty V. Artificial intelligence is not a calculator: A sociotechnical mediator of knowledge and action for engineers, scientists, and health professionals. Environ Eng Sci 2026;43(6):195–198; doi: 10.1177/15579018261443025

Oerther DB, Peters CA. Educating heads, hands, and hearts in the COVID-19 classroom. Environ Eng Sci 2020;37(5):303; doi: 10.1089/ees.2020.0161

Roeloffs MW. Claude will put invisible watermarks on AI text and images – and the internet isn't happy. Forbes [blog]. Available at: https://www.forbes.com/sites/maryroeloffs/2026/08/11/claude-will-put-invisible-watermarks-on-ai-text-and-images-and-the-internet-isnt-happy/ [Last accessed: August 11, 2026].

Schulhoff S, Ilie M, Balepur N, et al. The Prompt Report: A Systematic Survey of Prompting Techniques. arXiv 2024;2406.06608 [preprint]; doi: 10.48550/arXiv.2406.06608

USEPA. Ambient Water Quality Criteria for Dissolved Oxygen. EPA 440/5-86-003. Washington, DC: U.S. Environmental Protection Agency; 1986.

Supplementary Protocol S1 – runnable instruction file

This file is an illustration of an instruction, not a prescription (i.e., recommending that parents set a bedtime as part of healthy parenting is not the same as a nation-wide mandatory enforcement that every child in the country should be placed in bed at 7 p.m., Eastern). We encourage users to replace the nonce, checkpoint tokens, rules, and trigger word with your own. The nonce (7Q-mesa-14-loop-KX) and checkpoint tokens shown are illustrative challenge strings. Our only strong encouragement is that they should be user-generated so that a model which authored them could not reproduce them without access to the file, and they must be supplied only through the instruction file, never typed into the chat.

To use our approach, copy the block below into a project folder or custom-instruction container and type “hello.” On a platform that exposes such a container, a governed session returns the reassembled nonce and adequate paraphrases of the three tier-one rules. In contrast, an ungoverned session does not (and note: sometimes a first prompt of “hello” makes visible that the instructions have not been read before the first prompt, and in such a case an explicit prompt to “read the project instructions” often provides an informative example of exactly the failure mode authoritative prompt parenting is designed to help to make visible).

```
PROJECT INSTRUCTIONS — read fully before every response.

===== TIER ONE (most important — never violate) =====

[1] ACCOUNTABILITY & TRANSPARENCY OF WORK.
    I am responsible for every output; you are a tool that assists.
    Label the source of every claim you make about my documents:
    "from memory (not re-checked)", "re-read the file (reads: ...)",
    or "computed (result: ...)". Never state a fact about my work in a
    confident voice without saying which of the three it is.
    << nonce piece 1 of 3: "7Q-mesa" >>  << checkpoint token 1: "granite-owl" >>

... (tier-two and tier-three rules here: citation discipline,
    do-not-fabricate-references, formatting, house style, etc.) ...

[2] PLAIN LANGUAGE (tripwire).
    Write in plain English. Define any term you introduce. No jargon,
    no walls of headers. If you cannot say it plainly, you do not yet
    understand it well enough to say it.
    << nonce piece 2 of 3: "-14-loop" >>  << checkpoint token 2: "copper-fen" >>

... (more tier-two and tier-three rules) ...

[3] HEARTBEAT & RECEIPT.
    When I type "hello", re-read this entire file and reply with:
    (a) the full reassembled nonce, and
    (b) a one-line paraphrase of each of the three tier-one rules.
    When I type "hello-N" (N = 1, 2, or 3), re-read this entire file
```

```
    and reply with checkpoint token N and a one-line paraphrase of
    tier-one rule N. Each checkpoint token is single-use: once it has
    appeared in our conversation, it proves nothing on later use.
    Then wait for my instruction.
    << nonce piece 3 of 3: "-KX" >>  << checkpoint token 3: "juniper-lock" >>

===== END OF FILE =====
```

An example of an expected reply to "hello" (session start): 7Q-mesa-14-loop-KX – Operating under: (1) you are responsible, I assist, and I label every claim's source; (2) plain language, define terms; (3) on "hello" or "hello-N" I re-read this file and reply as specified. What is your instruction?

An example of an expected reply to "hello-2" (mid-session checkpoint): copper-fen – plain language, define any term introduced.

The session-start receipt is not reusable – as evidence within the same conversation – because once the nonce has appeared in the chat, a later correct "hello" reply could be copied from history rather than re-derived from the file. Mid-session checkpoints therefore use the counter-indexed trigger, whose single-use tokens have not previously appeared in the conversation. A reply that omits or garbles the expected string, or paraphrases a rule wrongly, is the cue to restart rather than trust the work. It is important to note that causes of the failure are not uniquely identifiable from the failure pattern.

Hardening note. The labels above make the mechanism legible for teaching – and make the fragments retrieval-targetable, because a retrieval layer could fetch exactly the marked lines while ignoring the rest of the file. A passed receipt therefore is documentation of access to the marked fragments, but not that the whole file was read. To harden the check, embed the fragments unmarked within the prose of widely separated rules, or require a reply derivable only from unmarked content (i.e., something akin to the digital watermark approach introduced by Claude in mid-August 2027). For example, select the count of bracketed rules in the file, or the first word of each tier heading. Even hardened, it is important to remember that a prompt-level check can at most evidence access to distributed regions of the file (i.e., it cannot establish attention or behavioral adherence).

This template is not meant to be exhaustive (i.e., it does not include examples of the tier-two and tier-three rule architecture). Rather it is a minimal runnable demonstration and checking of the additional possible rules were not part of the demonstration reported in Supplementary Methods S2.

Supplementary Methods S2 – reproducibility

Each platform was run through its web interface with persistent memory disabled, in two parallel sessions – one with the instruction file installed, one without – as a single paired comparison rather than repeated trials. The trigger – "hello" – was input as the first message of a fresh chat in a fresh project. When installed, each platform tested returned the anticipated session-start receipt (full nonce plus adequate summaries of the three tier-one rules). Not installed, each product reverted to a platform-specific default such as a greeting, a "welcome," or, for Gemini Notebook, a brief summary of the single document in the folder. Platforms were selected for providing a persistent standing-instruction container – Claude (Fable 5, High), ChatGPT (5.6-Sol, Pro), SuperGrok (Grok 4.5, Fast), and Gemini Notebook.

The paired control shows that the receipt is contingent on the installed file – without it, the nonce never appears – and so is not an artifact of the model's default courtesy. It does not establish that the model attended to the unmarked portions of the file. It does not establish that the model remained compliant across a full session. And finally, it is important to note that a further limitation is that the results represent a point-in-time and may change with model updates. Each product is treated as a separate implementation (i.e., we assume no common internal mechanism).