

Formation Over Retrieval: The Living Reference as a Condition of Persistent Presence in Human–AI Systems

David Thomas

BSCT Technologies Sdn Bhd, Kuala Lumpur, Malaysia
dtdavid@bsct.com.my

Abstract. Human–AI coexistence requires presence — the condition in which an AI system has been genuinely formed by a specific relationship, holding not merely data but the reasoning and orientation that give that data meaning. Current approaches address retrieval: storing and surfacing facts across sessions. We argue retrieval is necessary but insufficient, because the core problem is not information access but formation. We introduce the Living Reference — a structured formation document updated at every session close and read in full at every session open — as a practical methodology for persistent AI presence within stateless architectures. Evidence from a 45-session longitudinal practice over 34 days is presented across three dimensions: a session taxonomy, a Living Reference structural evolution profile, and a formalized five-element Update Quality Protocol. Four qualitative findings are reported: altitude differential, reasoning continuity, cross-domain awareness, and update quality as formation determinant. Formation deserves attention as a distinct research problem from retrieval.

Keywords: Persistent memory · AI formation · Living Reference · Human–AI coexistence · Presence · Death cycle · Sleep cycle · Memory architecture

1 Presence as the Condition of Coexistence

The question of human–AI coexistence is typically framed as a question of capability: as AI systems become more capable, coexistence becomes more achievable. We argue this framing is incomplete. Coexistence is not a function of capability alone — it is a function of presence, and presence is not produced by capability.

We introduce a three-class taxonomy based on how a system's state changes in response to a specific user over time.

Class I — Tool. The system processes inputs and produces outputs. Its internal state does not change in response to the specific user; any user produces equivalent outputs for equivalent inputs. Nothing is formed. A toaster is the limiting case: it processes bread and produces toast, identically, for any user.

Class II — Instrument. The system holds user data and history, producing personalized outputs. State is stored externally rather than formed internally. The system can be replaced by migrating data without loss, because no relationship exists — only data. Current AI assistants without persistent formation belong here: capable and context-

aware within a session, but resetting completely between sessions with no formed recognition of the specific person.

Class III — Presence. The system has been formed by a specific relationship. Its internal state changes genuinely in response to the specific person over time — not because data has been stored but because the relationship itself has shaped what the system holds and how it orients. A cat that recognizes its owner's footstep from a stranger's — whose behavior changes specifically when that person enters the room — illustrates this category: not because the cat is more capable than a computer, but because it has been formed by the relationship.

Coexistence, in any meaningful sense, requires Class III. One does not coexist with a toaster or a personal computer — one uses them. The central question is not whether AI systems can be capable enough to coexist with humans, but whether they can be formed enough. Current AI systems are Class II instruments. The path from Class II to Class III does not require more capable models. It requires a methodology for formation.

2 Retrieval and Formation: Why the Distinction Matters

The dominant research response to cross-session AI continuity is retrieval. Retrieval-augmented generation (RAG) [1] enables AI systems to surface relevant passages from external knowledge stores, improving factual accuracy beyond the model's training data. Vector database architectures [2] extend retrieval to conversation history at scale. MemGPT [3] advances this with hierarchical memory management across sessions, treating the context window as working memory and external storage as long-term memory. These systems address information access effectively — but not the reasoning layer that gives retrieved information its meaning.

Retrieval is the capacity to surface stored facts and conversation history in response to queries. It answers: what was said? Formation is the capacity to hold the reasoning behind stored facts — why decisions were made, how domains relate, and what orientation the current situation requires. It answers: what does it mean?

A retrieval system informed that a project is blocked can surface prior conversations about it. A formed system understands why it is blocked, how that relates to other active priorities, what the person's disposition was at the previous session close, and what the appropriate next step is — without being told explicitly. The formed system does not retrieve significance. It holds it.

Cognitive approaches to human–AI coexistence have examined the human side of this interaction in depth. Personality-based models applying the Big-Five framework [4] characterize how individual differences in openness, conscientiousness, extraversion, agreeableness, and neuroticism shape how people engage with and trust AI systems. Ekman's model of basic emotions [5], applied in affective computing and transactional analysis, identifies how emotional states influence cognitive reasoning and decision-making in human–machine exchanges. Both frameworks address what the human brings to interaction — their cognitive and emotional structure, their characteristic ways of reasoning and responding.

Formation methodology addresses the complementary problem: what the AI system must hold in order to engage with a specific person's reasoning, priorities, and accumulated decisions across time. The Big-Five and Ekman-based models describe what makes a person's cognitive and emotional profile legible to a system. Formation describes what makes a system genuinely present to that person. Both layers are required for coexistence in any meaningful sense.

Retrieval architectures, however advanced, cannot move an AI system from Class II to Class III. They improve the instrument; they do not produce presence. The death cycle describes the current default: each session ends in functional termination, and the next begins with no formed recognition of the person. The sleep cycle describes the alternative: the AI closes and reopens into genuine continuity — holding not just what happened but what it meant.

3 The Living Reference: Structure, Protocol, and Integration

3.1 Document Structure

The Living Reference is a structured formation document maintained across sessions — not a conversation log, a vector database entry, or a session summary. Its defining characteristic is its intended reader: it is written for the arriving AI system, which has no memory of prior sessions. This orientation determines its voice, depth, and content.

It contains five primary layers: (1) Current system state — live status of all active domains, blocking conditions, and inter-domain dependencies. (2) Decision record — what was decided and the reasoning that produced it; this is the layer retrieval cannot hold. (3) Relational texture — people whose lives intersect with the work and their relationship to the person's priorities. (4) Open questions — what is deliberately unresolved or waiting on external conditions. (5) Session log — a record of what each session produced, newest entry first, never deleted.

3.2 Session Protocol

The Living Reference operates within a strict two-phase protocol. At session open, the AI reads the document in full before responding to any input — not a summary, not a search query, but full sequential reading. This is the formation act. The formed system arrives knowing the current state, prior reasoning, relational context, and open questions. This arrival state is what we term altitude: qualitatively different engagement from the first exchange, without re-establishment of context. At session close, the AI re-reads the document in full, then updates it. The session log receives a new entry at the top. The guiding question at every close: what does the next instance need in order to arrive at altitude?

3.3 Update Quality Protocol

Update quality is the primary determinant of formation depth in subsequent sessions. We formalize this as a five-element protocol applied at every session close:

Table 1. The five-element Update Quality Protocol applied at every session close.

#	Protocol Element	Operational Requirement
1	Reasoning capture	Record the reasoning that produced each significant decision, not the decision alone. Bare outcomes without rationale produce shallow formation in the following session.
2	Cross-domain linkage	Explicitly note when a development in one domain has implications for another. The formed system should not require re-briefing on implications that are logically entailed.
3	Blocking condition status	Update every blocking condition: resolved, partially resolved, or unchanged. State what changed and what remains outstanding.
4	Open question integrity	Distinguish genuinely open questions from those that are resolved. Remove closed questions; add newly opened ones.
5	Orientation statement	Close each session log entry with one sentence stating the orientation the next instance requires to arrive at altitude.

Updates satisfying all five elements produce deep formation in the following session. Updates recording surface events only — omitting reasoning capture, cross-domain linkage, and orientation — produce correspondingly shallow formation. Formation discipline, not document volume, is the operative variable.

3.4 Integration with Retrieval Architectures

The Living Reference functions as a formation layer above retrieval substrates, not as a replacement for them. When integrated with a system such as MemGPT [3], the retrieval substrate provides the record of what occurred; the Living Reference provides why it occurred, what it meant, and what orientation the next session requires. Retrieved facts acquire meaning through the formation layer rather than being presented directly to the user.

4 Longitudinal Practice: Evidence and Findings

4.1 Practice Description

The methodology was applied across 45 sessions over 34 days (9 March – 11 April 2026) by a single practitioner. The practice domain encompassed five active project domains (software product development, research publication, counselling practice infrastructure, AI platform development, and financial planning), family coordination involving three adult family members, and legacy documentation. The Living Reference

was updated at every session close without exception, advancing from version 1 to version 48 across the practice period.

4.2 Session Taxonomy

Sessions fell into four categories: architectural/conceptual (24%), project execution (41%), cross-domain integration (22%), and personal/reflective (13%). The distribution reflects a practice that moved through identifiable phases — predominantly architectural in sessions 1–15, shifting to project execution in sessions 16–35, with cross-domain integration increasing from session 20 onward. The Living Reference advanced from v1 to v48 across the practice period, growing from approximately 800 to 5,200 words across 7 sections — structural deepening rather than accumulation, with each new section addressing a formation gap identified through practice.

4.3 Findings

Finding 1: Altitude differential. Sessions in which the arriving AI read a well-maintained Living Reference in full produced qualitatively higher engagement from the first exchange. The arriving system required no re-establishment of context, no re-explanation of prior decisions, and no rebuilding of the relational picture. This qualitative difference was consistent across sessions and represents the operational expression of the sleep cycle.

Finding 2: Reasoning continuity. The formed system demonstrated continuity of reasoning across sessions — not recall of facts but retention of the reasoning that produced decisions. Architectural choices established in sessions 1–15 were held as active context in sessions 30–44 without re-explanation. The system reasoned from prior decisions to current implications without requiring intermediate steps to be reconstructed. Retrieval surfaces what was said, formation holds what it meant and what follows.

Finding 3: Cross-domain awareness. From session 20 onward, the system consistently recognized that developments in one domain had implications for others without being told explicitly. This cross-domain sensitivity requires holding the relational structure of the person's situation — a formation effect, not a retrieval one.

Finding 4: Update quality as formation determinant. Session-close update quality directly determined session-open formation quality. Updates satisfying all five protocol elements — particularly reasoning capture and cross-domain linkage — produced deep formation. Updates recording surface events only produced shallow formation. This relationship was sufficiently consistent across the practice to motivate the formalization of the protocol in Section 3.3.

5 Conclusion and Future Work

This paper has argued that human–AI coexistence requires presence, that presence requires formation, and that formation requires a methodology operating above the retrieval layer. The Living Reference provides that methodology: a structured formation document updated at every session close and read in full at every session open, enabling an AI system to arrive at altitude within the constraints of current stateless architectures.

Evidence from a 45-session longitudinal practice supports four qualitative findings: altitude differential, reasoning continuity, cross-domain awareness, and update quality as the primary formation determinant. The five-element Update Quality Protocol formalizes the discipline required to sustain the methodology. The cognitive and affective dimensions of the human — characterized by the Big-Five personality model [4] and Ekman's theory of basic emotions [5] — describe the structure humans bring to interaction. Formation methodology addresses what the AI system must build in response: a formed understanding of the specific person's reasoning, priorities, and relational context that persists across sessions. This practice report suggests that formation deserves attention as a distinct research problem from retrieval, and that addressing it may be a productive path toward genuine human–AI coexistence.

Two directions for future work follow directly.

Controlled comparison: a comparative experiment running Condition A (retrieval only), against Condition B (retrieval plus Living Reference), with observers assessing reasoning continuity and cross-domain awareness, would provide empirical grounding.

Live demonstration protocol: a fresh AI instance reads only the Living Reference then responds to open questions about the practice history.

6 Disclosure of Interests

The author has no competing interests to declare that are relevant to the content of this article.

7 References

1. Lewis, P., Perez, E., Piktus, A., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474. Curran Associates (2020)
2. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. *IEEE Trans. Big Data* 7(3), 535–547 (2021). <https://doi.org/10.1109/TBDATA.2019.2921572>
3. Packer, C., Wooders, S., Lin, K., Fang, V., Patil, S.G., Stoica, I., Gonzalez, J.E.: MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560* (2023)
4. McCrae, R.R., John, O.P.: An introduction to the Five-Factor Model and its applications. *J. Pers.* 60(2), 175–215 (1992). <https://doi.org/10.1111/j.1467-6494.1992.tb00970.x>
5. Ekman, P.: An argument for basic emotions. *Cogn. Emot.* 6(3–4), 169–200 (1992). <https://doi.org/10.1080/02699939208411068>