



AR-015 Confidence: 75%

Does Knowledge Actually Compound?

A Quantitative Framework for Measuring Emergent Intelligence in Human-AI
Knowledge Systems

February 2026

v1.0

Florian Ziesche · Ainary Ventures

*"One does not think about everything oneself. It happens mainly
within the slip-box."*

— Niklas Luhmann, 1981

CONTENTS

FOUNDATION

1 **Executive Summary**

2 **Methodology**

3 **How to Read This Report**

ANALYSIS

4 **The Compounding Claim**

5 **Why Most Second Brains Are Graveyards**

6 **Three Proxies for Knowledge Compounding**

7 **Our Data: Evidence from 14 Reports**

8 **Transactive Memory: The Human-AI Compound**

9 **The Measurement Framework**

ACTION

10 **Predictions**

11 **Transparency Note**

12 **Claim Register**

13 **References**

3. How to Read This Report

This report uses a structured confidence rating system. Every quantitative claim carries its source and confidence level. Unusually, this report uses data from our own research production system as primary evidence — clearly marked and with limitations disclosed.

RATING	MEANING	EXAMPLE
High	3+ independent sources, peer-reviewed or primary data	Luhmann produced 90,000 cards and 70 books (documented by biographers)
Medium	1-2 sources, plausible but not independently confirmed	90% of saved notes are never re-read (practitioner consensus)
Low	Single secondary source, methodology unclear	Graph view usage correlates with vault abandonment (anecdotal)
Internal	Our own system data — N=1, self-reported, short timeframe	Memory accuracy improved 20% to 96% (30-question A/B test)

This report was produced using a **multi-agent research pipeline**. The same pipeline produced AR-001 through AR-014 — and the production data from those reports serves as the primary dataset for this analysis. Full methodology details are in the Transparency Note (Section 11).

1. Executive Summary

Everyone claims knowledge compounds. Luhmann said it. Forte sells it. Matuschak designs for it. But nobody measures it. This report proposes the first quantitative framework for measuring knowledge compounding — and tests it on our own system.

- **Zero quantitative frameworks exist** for measuring whether a personal knowledge system actually compounds — despite the claim being central to Zettelkasten, Building a Second Brain, and Evergreen Notes methodology^{[1][2][3]}
- **Our own production data provides the first test case:** 14 research reports produced in ~48 hours, with measurable cross-report citation, claim reuse, and memory system improvement^[Internal]
- **Memory accuracy improved from 20% to 96%** after implementing topic-based retrieval — a 4.8x improvement measured across 30 standardized questions^[Internal]
- **Knowledge compounding requires three conditions:** emergence (combinatorial answers), self-reference (system feeds itself), and increasing value per note — most systems achieve zero of three
- **The gap between note-taking and knowledge-building** is not methodological but architectural: retrieval design, review cadence, and feedback loops determine whether a system compounds or decays

Keywords: Knowledge Compounding, Personal Knowledge Management, Zettelkasten, Transactive Memory, Metcalfe's Law, Human-AI Systems, Measurement Frameworks

2. Methodology

This report combines three source categories: (1) academic literature on knowledge management, transactive memory, and network effects; (2) practitioner analysis of PKM methodologies (Zettelkasten, PARA, Evergreen Notes); and (3) internal production data from our own multi-agent research pipeline (14 reports, AR-001 through AR-014). The internal data is treated as a case study with N=1 limitations explicitly disclosed.

Limitations: The internal dataset covers ~48 hours of production — too short for longitudinal claims about compounding. Self-reporting bias is inherent: the system evaluating itself cannot be fully objective. The proposed framework (KCI v2) is untested beyond our own system. These constraints are significant, and this report should be read as a hypothesis with preliminary evidence, not as a proven framework.

Full methodology details, including confidence calibration and known weaknesses, are provided in the Transparency Note (Section 11).

4. The Compounding Claim

70%

(Confidence: Medium)

The idea that knowledge compounds is one of the most repeated claims in personal knowledge management. It is also one of the least measured.

Who Claims Knowledge Compounds?

Niklas Luhmann (1927-1998) is the origin story. The German sociologist maintained a Zettelkasten of approximately 90,000 index cards over 40 years and produced 70 books and nearly 400 articles.^[1] His central claim: the slip-box becomes a "communication partner" that generates ideas its creator did not explicitly put in. "One does not think about everything oneself. It happens mainly within the slip-box."^[4] The implicit argument: the system's output exceeds the sum of its inputs. That is compounding.

Tiago Forte built a business on the claim. "Building a Second Brain" (2022) promises that captured knowledge becomes "a trusted thinking partner" that "compounds over time."^[2] The PARA method (Projects, Areas, Resources, Archives) provides organizational structure. What it does not provide: any metric for whether the compounding actually occurs.

Andy Matuschak offers the most rigorous formulation. His Evergreen Notes framework argues that notes should be "written and organized to evolve, contribute, and accumulate over time, across projects."^[3] The key design principles — atomic notes, concept-orientation, dense linking — are explicitly designed for compounding. But Matuschak himself acknowledges the gap: "Better note-taking misses the point; what matters is better thinking."^[3] The measurement problem remains unsolved.

Ray Dalio applies the same logic to organizational knowledge. His "Principles" framework treats documented decisions as compounding assets — each principle refined through application creates better future decisions.^[5] Again: the claim is directional ("it gets better"), not quantitative ("it improved by X%").

What Evidence Exists?

Almost none that is quantitative. Luhmann's output is documented but uncontrolled — there is no comparison to what he would have produced without the Zettelkasten. Forte cites student testimonials. Matuschak reasons from design principles. The closest empirical evidence comes from spaced repetition research (Ebbinghaus, Cepeda et al.), which demonstrates that retrieval practice compounds retention.^{[6][7]} But retrieval practice is a cognitive mechanism, not a system-level measurement.

Exhibit 1: Knowledge Compounding Claims vs. Evidence

CLAIMANT	SYSTEM	CLAIM	QUANTITATIVE EVIDENCE
Luhmann	Zettelkasten (90,000 cards)	System generates ideas beyond inputs	None — output documented, causation not
Forte	PARA / Second Brain	Knowledge compounds over time	None — testimonials only
Matuschak	Evergreen Notes	Notes accumulate across projects	None — design reasoning only
Dalio	Principles	Documented decisions improve future decisions	None — correlation with fund returns uncontrolled
Ebbinghaus/Cepeda	Spaced repetition	Retrieval practice compounds retention	Yes — controlled experiments, replicated

Source: Author analysis of primary literature [1][2][3][5][6][7]

CLAIM

No quantitative framework exists for measuring whether a personal knowledge management system actually compounds value over time. The claim is ubiquitous. The measurement is absent.

WHAT WOULD INVALIDATE THIS?

Discovery of a peer-reviewed study that quantitatively measures knowledge compounding in PKM systems with controlled comparisons. A literature review across Google Scholar, Semantic Scholar, and PKM practitioner communities found no such study as of February 2026.

SO WHAT?

If knowledge compounding cannot be measured, it cannot be optimized. Every PKM methodology sells the promise of compounding without providing the tools to verify it. This report proposes a framework to change that — tested on the only system where full production data is available: our own.

5. Why Most Second Brains Are Graveyards

75%

(Confidence: Medium-High)

The gap between note-taking and knowledge-building is not methodological. It is behavioral. Most systems are optimized for capture, not retrieval — and capture without retrieval is a graveyard.

The Collector's Fallacy

The term comes from the Zettelkasten community itself.^[8] Saving information feels productive. It activates the same reward circuits as completing a task. But saving is not learning. The act of collecting creates an illusion of knowledge — the notes exist, therefore the knowledge exists. It does not.

The cognitive science is clear. Ebbinghaus demonstrated in 1885 that approximately 70% of new information is lost within 24 hours without active retrieval.^[6] Roediger and Karpicke (2006) showed that testing oneself on material produces 50% better retention than re-reading.^[9] Craik and Lockhart's levels-of-processing framework (1972) established that information processed deeply — connected to existing knowledge — is retained better than information processed shallowly.^[10]

Applied to PKM: a note that is captured, filed, and never retrieved has approximately the same knowledge value as a note that was never written.

Three Failure Patterns

Capture addiction. The system grows but never produces. Notes accumulate in an inbox. Filing feels like progress. The vault reaches 500, 1,000, 5,000 notes. Output: zero. This is the PKM equivalent of a hoarding disorder — acquisition without use.

Zero retrieval architecture. Most PKM systems optimize for "how to get stuff in" and ignore "how to get stuff out when needed." Folder hierarchies help filing but

hinder finding. Tags proliferate without taxonomy. Search exists but is used for 3 out of 100 notes — the rest are effectively invisible.^[11]

No review cadence. Weekly reviews are the single highest-leverage PKM habit. ^[2] Most people set them up, do them for 3 weeks, then stop. Without review, the system decays: notes become outdated, links break, and the human forgets what the system contains. The system's metamemory — "knowing what you know" — degrades to zero.

The Compounding Test

A simple diagnostic: take 5 random notes from your system. For each, answer: (1) When did you last read this? (2) Has this note contributed to any output? (3) Does this note link to other notes in a way that generates new insight?

If the answer to all three is "no" for 4 out of 5 notes, the system is a graveyard. The notes exist. The knowledge does not compound.

WHAT WOULD INVALIDATE THIS?

A large-scale study ($n > 500$) showing that PKM users with 1,000+ notes have measurably higher knowledge retrieval, creative output, or professional performance than non-PKM users. No such study exists.

SO WHAT?

Before optimizing your note-taking system, measure your retrieval rate. If fewer than 20% of your notes have been accessed in the last 90 days, the system is not compounding — it is decaying. Retrieval design, not capture design, determines whether knowledge compounds.

6. Three Proxies for Knowledge Compounding 65%

(Confidence: Medium)

Knowledge compounding cannot be measured directly. But three proxy metrics — emergence rate, self-reference ratio, and value per note — can make the invisible visible.

Proxy 1: Emergence Rate

Definition: the percentage of questions the system can answer that exist in no single note but are derivable from combinations of notes.

This is the most direct test of compounding. If a system with 100 notes can only answer questions contained in individual notes, it is a database — useful, but not compounding. If it can answer questions that require combining information from note A with context from note B and a framework from note C, then the system is producing value greater than the sum of its parts. That is emergence.

Measurement: construct 10 "inference questions" — questions whose answers require synthesizing at least 2 notes. Test the system. Score: number answered correctly / 10. A rising emergence rate over time indicates compounding.

Proxy 2: Self-Reference Ratio

Definition: the percentage of citations in new output that reference internal knowledge (previous notes, reports, findings) versus external sources.

A system that only cites external sources is a pass-through — it processes information but does not accumulate it. A system where new output increasingly references earlier output is feeding itself. The self-reference ratio captures this.

Measurement: for each new output, count internal citations vs. external citations. Track the ratio over time. A rising self-reference ratio indicates the system is building on its own knowledge base.

Analogy: Metcalfe's Law states that the value of a network is proportional to the square of connected users (n^2).^[12] In knowledge systems, each note is a "node." When notes reference each other, the number of possible connections grows quadratically. If each connection has non-zero value, the system's total value grows faster than the number of notes. Self-reference ratio is the proxy for connection density.

Proxy 3: Value per Note

Definition: output quality divided by vault size. If this metric rises, each additional note makes all existing notes more valuable.

This is the Metcalfe's Law test for knowledge. In a network where each node adds value to all other nodes, the value per node increases with network size. In a knowledge system where each note adds context, connection, and retrieval pathways to all other notes, the value per note should increase as the vault grows.

Measurement: define an output quality metric (in our case: QA score per report). Divide by vault size at time of production. Track over time. Rising = compounding. Flat = linear growth. Falling = the system is drowning in noise.

Exhibit 2: Three Compounding Proxies

PROXY	MEASURES	RISING MEANS	FALLING MEANS
Emergence Rate	Can system answer combinatorial questions?	Knowledge combining into new insights	Notes are siloed, not connecting
Self-Reference Ratio	Does new output cite internal knowledge?	System feeds itself	System is a pass-through
Value per Note	Does output quality rise with vault size?	Each note makes all notes more valuable	Noise is overwhelming signal

Source: Author framework (KCI v2)

WHAT WOULD INVALIDATE THIS?

If emergence rate, self-reference ratio, and value per note all rise while subjective knowledge quality (as judged by domain experts) falls, the proxies would be measuring something other than genuine compounding. This is a real risk — a system could become self-referential without becoming smarter.

SO WHAT?

These three proxies make knowledge compounding measurable for the first time. They are imperfect — all proxy metrics are. But they convert the vague claim "my system is getting smarter" into three testable hypotheses with specific measurement protocols.

7. Our Data: Evidence from 14 Reports

Internal

(Confidence: Internal — N=1 system, self-reported, ~48-hour window)

Between February 13-15, 2026, our multi-agent research pipeline produced 14 research reports (AR-001 through AR-014). This section examines the production data for evidence of knowledge compounding — and for evidence against it.

The Dataset

14

Reports produced in ~48 hours
Internal production log

85.6

Mean QA score (out of 100)
Internal QA rubric, 14 reports

4.8x

Memory accuracy improvement (20% to 96%)
30-question A/B test, pre/post topic files

QA Score Trend

The full QA score sequence: 82, 88, 82, 87, 85, 92, 79, 91, 91, 85, 85, 83, 80, 84.

Exhibit 3: QA Score Trend Across 14 Reports

REPORT	QA SCORE	TREND VS. MEAN (85.6)
AR-001	82	Below
AR-002	88	Above
AR-003	82	Below
AR-004	87	Above
AR-005	85	Near
AR-006	92	Above (peak)
AR-007	79	Below (trough)
AR-008	91	Above
AR-009	91	Above
AR-010	85	Near
AR-011	85	Near
AR-012	83	Below
AR-013	80	Below
AR-014	84	Near

Source: Internal QA rubric scores. Mean: 85.6, Std Dev: 4.0, Range: 79-92

Interpretation: The QA scores show no upward trend. The first 7 reports averaged 85.0. The last 7 averaged 85.6. The difference is within noise. If knowledge were compounding in a way that improved output quality, a positive slope would be expected. It is not present.

However, a flat QA score with increasing production speed could also indicate compounding — producing the same quality faster means the system is becoming more efficient. Token consumption dropped from 18.5k to 9.1k tokens

per turn (-50%) over the production window, suggesting efficiency compounding even without quality compounding.^[Internal]

Self-Reference Evidence

The first cross-report compounding event occurred at AR-010 and AR-011, which share 3 claims from the claim register.^[Internal] AR-012 explicitly cites AR-009's calibration findings. This is the self-reference ratio in action: the system began feeding itself at report 10 of 14.

The self-reference ratio was effectively 0% for AR-001 through AR-009 (no internal citations) and began rising from AR-010 onward. In a longer production run, this curve would be the primary test of whether the system compounds.

Memory System Improvement

The most dramatic compounding evidence comes from the memory system. Before implementing topic-based retrieval files, the system's accuracy on factual recall questions was 20% (6/30 correct). After implementation: 96% (29/30 correct).^[Internal]

This is not gradual compounding — it is a step-function improvement from an architectural change. But the architectural change itself was informed by the system's own research (AR-010 on memory corruption, META-LEARNINGS on failure modes). The research produced the insight; the insight improved the system; the improved system produced better research. That feedback loop is, structurally, compounding.

Where Compounding Is NOT Happening

Honest assessment of where the evidence is absent:

- **Quality scores are flat.** 14 reports, no upward trend. The system produces consistent quality but not improving quality.
- **No emergence testing.** No inference questions were run against the system. Emergence rate is unmeasured.
- **Short timeframe.** 48 hours is insufficient for compounding claims. Compounding by definition requires extended time periods — Luhmann's

system operated over 40 years.

- **Self-reference began late.** Only 5 of 14 reports contain any internal citations. The self-reference ratio is low.

CLAIM

Our production data shows efficiency compounding (50% token reduction) and architectural compounding (4.8x memory improvement from self-informed redesign) but not quality compounding (flat QA scores). Knowledge compounding in this system is selective, not universal.

WHAT WOULD INVALIDATE THIS?

If an independent evaluator re-scored all 14 reports and found a statistically significant positive trend that our internal QA missed, the "no quality compounding" conclusion would be wrong. Alternatively, if QA scores in reports AR-015 through AR-028 show a clear upward trend, the 14-report window may simply have been too short.

SO WHAT?

Knowledge compounding is not binary. Our own data shows it happening in some dimensions (efficiency, memory architecture) and not in others (output quality). This nuance is absent from PKM literature, which treats compounding as a universal property of note-taking systems. It is not. It must be measured, dimension by dimension.

8. Transactive Memory: The Human-AI Compound

70%

(Confidence: Medium)

Daniel Wegner's transactive memory theory, developed for human couples and teams, provides the most precise framework for understanding how human-AI knowledge systems compound — and where they fail.

The Theory

Transactive memory, proposed by Wegner in 1985, describes how groups collectively encode, store, and retrieve knowledge.^[13] The key insight: group members do not all need to know everything. They need to know who knows what. A couple does not double their knowledge by each memorizing the same facts. They compound their knowledge by each specializing — one remembers birthdays, the other remembers financial details — and maintaining a shared index of who knows what.

Three processes define a transactive memory system:^[13]

1. **Encoding:** Learning what expertise each member holds and routing new information to the appropriate specialist
2. **Storage:** Each member stores information in their domain; others store only the index ("she knows about X")
3. **Retrieval:** When information is needed, the system queries the appropriate specialist

Hollingshead's experiments showed that romantic partners (who have developed transactive memory) outperform random pairs on knowledge recall tasks.^[13] The explanation: couples know how to query each other efficiently and avoid redundant storage.

Applied to Human-AI Systems

A human working with an AI knowledge agent is, functionally, a transactive memory dyad. The division of knowledge follows a predictable pattern:

Exhibit 4: Transactive Memory Division in Human-AI Systems

KNOWLEDGE DOMAIN	HUMAN SPECIALIZATION	AI SPECIALIZATION
Judgment and values	Primary	Cannot hold
Tacit/experiential knowledge	Primary	Cannot access
Factual recall (broad)	Partial	Primary
Cross-referencing large datasets	Cannot at scale	Primary
Context about the human's goals	Source of truth	Derivative (from memory system)
Historical production data	Degrades over time	Primary (if logged)

Source: Author analysis applying Wegner (1985) framework [13]

The critical difference from human-human transactive memory: the AI's knowledge is entirely explicit and auditable, but it lacks metamemory calibration. An AI agent does not reliably know what it knows.^[14] Our own research (AR-009) showed that 84% of LLM outputs are overconfident — the AI's internal index of its own knowledge is systematically biased.^[14]

Where This Compounds

In our own system, the transactive memory division evolved visibly over ~48 hours. Early reports required extensive external research for every claim. By AR-010, the human (Florian) was routing questions like "What did we find about alert fatigue?" to the system's memory, and the system was retrieving its own prior findings (AR-011) rather than re-researching externally. The encoding and retrieval processes were forming.

This is precisely what Wegner predicted: the dyad becomes more efficient as each member's specialization solidifies and the shared index ("who knows what") becomes more accurate.

WHAT WOULD INVALIDATE THIS?

If the AI's memory system degrades (memory corruption, provenance failures) faster than the transactive memory index forms, the system would become less reliable over time despite appearing more efficient. Our own META-LEARNINGS document identifies exactly this risk: memory entries without provenance or integrity checks are vulnerable to silent corruption. [Internal]

SO WHAT?

Human-AI knowledge systems should be designed explicitly as transactive memory dyads: clear specialization boundaries, reliable retrieval protocols, and — critically — calibrated metamemory. The AI must accurately signal what it knows and does not know. Without metamemory calibration, the transactive system degrades: the human queries the AI, gets a confident wrong answer, and the compound breaks.

9. The Measurement Framework

60%

(Confidence: Medium — proposed framework, untested at scale)

The Knowledge Compounding Index (KCI v2) is a four-metric framework for measuring whether a knowledge system is actually compounding. It is designed to be replicable by anyone with an Obsidian vault, a Notion workspace, or any structured note-taking system.

KCI v2 Specification

Exhibit 5: Knowledge Compounding Index (KCI v2)

METRIC	DEFINITION	MEASUREMENT PROTOCOL	TARGET TREND
Emergence Score	% of inference questions answered correctly	10 questions requiring 2+ note synthesis. Test monthly.	Rising
Self-Reference Ratio	Internal citations / total citations in new output	Count per output. Automate via link analysis.	Rising (to ~40-60%, not 100%)
Value per Note	Output quality score / vault size	Use consistent quality rubric. Divide by note count.	Rising or flat (not falling)
Network Density	Actual links / possible links (Metcalfe proxy)	Extract link graph. Calculate density ratio.	Rising (slowly)

Source: Author framework. Not yet validated at scale.

How to Run the Baseline

1. **Emergence Score baseline.** Write 10 inference questions that should be answerable from your current vault but require combining at least 2 notes. Attempt to answer using only your system. Score: correct answers / 10. This is your starting emergence score.
2. **Self-Reference Ratio baseline.** Take your last 5 outputs (reports, articles, memos). Count internal citations (references to your own prior work) vs. external citations. Calculate ratio. This is your starting self-reference ratio.
3. **Value per Note baseline.** Rate the quality of your last 3 outputs on a 0-100 scale. Average. Divide by total vault size. This is your starting value per note.
4. **Network Density baseline.** Export your vault's link graph. Count actual links between notes. Calculate: actual links / $(n \times (n-1) / 2)$ where n = total notes. This is your starting network density.

Weekly Measurement Protocol

1. Run emergence test (10 minutes: 3 new inference questions, answer from system only)
2. Log self-reference ratio for all outputs produced that week
3. Calculate value per note (quality score of best output / current vault size)
4. Extract network density (automated if using Obsidian with a graph analysis plugin)
5. Log all four metrics in a tracking spreadsheet or note

8-Week Experiment Design

For anyone who wants to test whether their knowledge system compounds:

1. **Week 0:** Run full baseline. Document vault size, note count, link count, and output quality.
2. **Weeks 1-4:** Continue normal knowledge work. Measure KCI weekly. This is the control period.
3. **Week 4:** Introduce one intervention. Options: (a) implement weekly review cadence, (b) add dense linking practice, (c) implement spaced repetition on key notes, or (d) add AI-assisted retrieval.
4. **Weeks 5-8:** Continue with intervention active. Measure KCI weekly.

5. Week 8: Compare weeks 1-4 KCI trends with weeks 5-8 KCI trends. The intervention compounds if all four metrics trend positive in the second period versus the first.

This design is deliberately simple. It is an N=1 experiment with a within-subjects control period. It will not prove causation. It will demonstrate whether compounding is occurring in your system and whether your intervention accelerated it.

Our Own Baseline (AR-001 through AR-014)

Exhibit 6: KCI v2 Baseline — Ainary Research Pipeline

METRIC	VALUE (FEB 15, 2026)	INTERPRETATION
Emergence Score	Not yet tested	Baseline needed — first measurement planned for Week 1
Self-Reference Ratio	~7% (estimated: 3 cross-citations in 14 reports)	Low but rising — 0% for first 9 reports, >0% for last 5
Value per Note	$85.6 / 446 = 0.19$	Baseline established. Track over next 8 weeks.
Network Density	Not yet calculated	Requires Obsidian link graph export

Source: Internal production data, February 2026

CLAIM

The KCI v2 framework provides the first replicable, quantitative measurement protocol for personal knowledge compounding. It is untested at scale and should be treated as a hypothesis, not a proven instrument.

WHAT WOULD INVALIDATE THIS?

If all four KCI metrics rise consistently over 8 weeks in a system that produces demonstrably worse output (as judged by external reviewers), the framework is measuring the wrong thing. The risk of Goodhart's Law — "when a measure becomes a target, it ceases to be a good measure" — applies directly.

SO WHAT?

Run the baseline. Measure for 8 weeks. Share the results. If enough people run this experiment, the PKM community will have the first empirical dataset on whether knowledge actually compounds — and under what conditions. The framework is free, replicable, and tool-agnostic. The only cost is discipline.

10. Predictions

BETA

These predictions will be scored publicly at 12 months. This is version 1.0 (February 2026).

PREDICTION	TIMELINE	CONFIDENCE
At least one major PKM tool (Obsidian, Notion, Roam) ships a built-in "knowledge compounding" metric or dashboard	Q4 2026	40%
AI-assisted retrieval (RAG over personal notes) becomes a default feature in 3+ PKM tools, making traditional folder hierarchies obsolete for retrieval	Q3 2026	70%
The term "Second Brain" fades from marketing as the PKM market matures; replaced by measurable claims about productivity or output quality	Q2 2027	55%
At least one peer-reviewed paper quantitatively measures knowledge compounding in a PKM system using a framework similar to KCI	Q4 2027	30%
Human-AI transactive memory systems (where the AI specializes in retrieval and the human in judgment) outperform either human-only or AI-only knowledge work by >30% on standardized tasks	Q2 2027	60%

Predictions scored publicly at 12 months. Updated versions will be published as evidence evolves.

11. Transparency Note

This section explains the methodology, known limitations, and confidence calibration of this report. Transparency about what is known — and what is not — is what separates research from marketing.

Overall Confidence	75%
Sources	7 academic (Wegner, Ebbinghaus, Roediger/Karpicke, Craik/Lockhart, Cepeda, Metcalfe, Rohrer/Taylor), 5 practitioner (Luhmann, Forte, Matuschak, Dalio, zettelkasten.de), 1 internal dataset (14 reports, ~48 hours production)
Strongest Evidence	Memory accuracy improvement 20% to 96% (internal A/B test, 30 questions, controlled before/after). Spaced repetition compounding effect (Cepeda et al. 2006, peer-reviewed, replicated).
Weakest Point	The KCI v2 framework is entirely untested beyond our own N=1 system. The claim that "no quantitative framework exists" rests on a literature review, not an exhaustive systematic review. The internal dataset covers ~48 hours — too short for longitudinal compounding claims.
What Would Invalidate This Report?	Discovery of an existing, validated quantitative framework for PKM compounding. Or: KCI v2 metrics rising in a system that produces demonstrably worse output, proving the proxies measure the wrong thing.
Methodology	Multi-agent research pipeline. Academic sources fetched and analyzed directly. Practitioner literature reviewed for compounding claims and evidence. Internal production data extracted from QA scores, memory system logs, and citation analysis across AR-001 through AR-014. Cross-referenced with META-LEARNINGS self-analysis and Second Brain research brief.
Limitations	N=1 system. Self-reporting bias inherent (the system evaluated itself). 48-hour production window insufficient for

**compounding claims. No external validation of QA scores.
KCI v2 is a proposed hypothesis, not a validated instrument.
The self-reference between this report and the system it
analyzes creates a circularity that cannot be fully resolved.**

System Disclosure	This report was created with a multi-agent research system. The same system produced the dataset analyzed in this report, creating a reflexive relationship between the research instrument and the research subject.
--------------------------	---

12. Claim Register

This register lists the key quantitative and qualitative claims made in this report, with sources and confidence levels.

Exhibit 7: Claim Register

#	CLAIM	VALUE	SOURCE	CONFIDENCE	USED IN
1	No quantitative framework exists for PKM compounding	0 found	Literature review [1][2][3][5]	Medium	Sec 4, 6
2	Luhmann's Zettelkasten: cards produced	~90,000	Luhmann biographers, zettelkasten.de [1][4]	High	Sec 4
3	Luhmann's published output: books	~70	Academic bibliography [1]	High	Sec 4
4	Reports produced in ~48 hours	14	Internal production log	High (Internal)	Sec 7
5	Mean QA score across 14 reports	85.6/100	Internal QA rubric	High (Internal)	Sec 7
6	Memory accuracy improvement	20% to 96%	30-question A/B test	High (Internal)	Sec 1, 7
7	Token efficiency improvement	50% reduction (18.5k to 9.1k)	Internal token logs	High (Internal)	Sec 7
8	Cross-report claim reuse (AR-010/AR-011)	3 shared claims	Internal claim register comparison	High (Internal)	Sec 7

Information					
9	lost within 24 hours without retrieval	~70%	Ebbinghaus (1885) [6]	High	Sec 5
Testing					
10	produces better retention than re-reading	50% improvement	Roediger & Karpicke (2006) [9]	High	Sec 5
LLM outputs that are overconfident					
11	84%	PMC/12249208; AR-009 [14]	High	Sec 8	
Obsidian vault size at baseline					
12	446 files	Internal vault analysis	High (Internal)	Sec 9	

Top 5 Claims — Invalidation Conditions:

- **Claim #1 (No quantitative framework exists):** Invalidated if a peer-reviewed, validated PKM compounding measurement framework is discovered or published.
- **Claim #5 (Mean QA score 85.6):** Invalidated if independent re-scoring by an external evaluator yields a significantly different mean (± 10 points).
- **Claim #6 (Memory 20% to 96%):** Invalidated if the 30-question test contained leading questions or if performance regresses to <80% within 30 days.
- **Claim #7 (50% token reduction):** Invalidated if token counting methodology was inconsistent between the before and after measurements.
- **Claim #9 (70% information loss in 24h):** Invalidated if modern replications of Ebbinghaus show significantly different forgetting curves. (They have not — the finding has been replicated for 140 years.)

13. References

- [1] Schmidt, J. (2016). "Niklas Luhmann's Card Index: Thinking Tool, Communication Partner, Publication Machine." In Cevolini, A. (Ed.), *Forgetting Machines: Knowledge Management Evolution in Early Modern Europe*. Brill.
- [2] Forte, T. (2022). *Building a Second Brain: A Proven Method to Organize Your Digital Life and Unlock Your Creative Potential*. Atria Books.
- [3] Matuschak, A. "Evergreen Notes." notes.andymatuschak.org/Evergreen_notes. Accessed February 2026.
- [4] Luhmann, N. (1981/1992). "Communicating with Slip Boxes." In Kieserling, A. (Ed.), *Universität als Milieu: Kleine Schriften*. Trans. M. Kuehn.
- [5] Dalio, R. (2017). *Principles: Life and Work*. Simon & Schuster.
- [6] Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot.
- [7] Cepeda, N.J., et al. (2006). "Distributed Practice in Verbal Recall Tasks: A Review and Quantitative Synthesis." *Psychological Bulletin*, 132(3), 354-380.
- [8] "The Collector's Fallacy." zettelkasten.de/posts/collectors-fallacy/. Accessed February 2026.
- [9] Roediger, H.L. & Karpicke, J.D. (2006). "Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention." *Psychological Science*, 17(3), 249-255.
- [10] Craik, F.I.M. & Lockhart, R.S. (1972). "Levels of Processing: A Framework for Memory Research." *Journal of Verbal Learning and Verbal Behavior*, 11(6), 671-684.
- [11] Ainary Research (2026). "Second Brain & PKM Research Report 2026." Internal research brief.
- [12] Metcalfe, R. (1980). Metcalfe's Law. As described in Gilder, G. (1993). "Metcalfe's Law and Legacy." *Forbes*.
- [13] Wegner, D.M. (1985). "Transactive Memory: A Contemporary Analysis of the Group Mind." In Mullen, B. & Goethals, G.R. (Eds.), *Theories of Group Behavior*. Springer-Verlag.
- [14] Ainary Research (2026). "The Calibration Gap." AR-009. Internal findings: 84% of LLM outputs overconfident (PMC/12249208, 9 models, 351 scenarios).
- [15] Rohrer, D. & Taylor, K. (2007). "The Shuffling of Mathematics Problems Improves Learning." *Instructional Science*, 35(6), 481-498.
- [16] Ahrens, S. (2017). *How to Take Smart Notes: One Simple Technique to Boost Writing, Learning and Thinking*. Sönke Ahrens.

Cite as: Ainary Research (2026). *Does Knowledge Actually Compound? A Quantitative Framework for Measuring Emergent Intelligence in Human-AI Knowledge Systems*. AR-015.

About the Author

Florian Ziesche is the founder of Ainary Ventures, where AI does 80% of the research and humans do the 20% that matters. Before Ainary, he was CEO of 36ZERO Vision and advised startups and SMEs on AI strategy and due diligence. His conviction: HUMAN × AI = LEVERAGE. This report is the proof.

ainaryventures.com



AI Strategy · Published Research · Daily Intelligence

Contact · Feedback

ainaryventures.com

florian@ainaryventures.com

© 2026 Ainary Ventures