● Ainary

AR-029 Confidence: 62%

# Does AI Quality Actually Compound?

A 25-Report Longitudinal Study of Our Own Pipeline. The Answer Is More Interesting Than We Expected.

February 2026

v1.0

Florian Ziesche · Ainary Ventures

*"Diminishing improvements in single-step accuracy can compound, leading to exponential growth in the length of task a model can complete."*

— Sinha et al., "The Illusion of Diminishing Returns," University of Cambridge / MPI, 2025

## CONTENTS

FOUNDATION

ANALYSIS

ACTION

# 1. How to Read This Report

This report examines our own pipeline — the system that produced AR-001 through AR-025. We are grading our own homework. That bias is the point: the meta-question of whether AI quality self-assessment is even valid is as important as the quality data itself.

| RATING | MEANING | EXAMPLE |
|---|---|---|
| High | 3+ independent sources, peer-reviewed or primary data | Diminishing returns in LLM scaling (PNAS, ScienceDirect, JMLR) |
| Medium | 1-2 sources, plausible but not independently confirmed | Fine-tuning saturation at ~6,500 samples (arXiv 2024) |
| Low | Single secondary source, methodology unclear | AI prototype-to-production takes 8 months average (Gartner via Medium) |
| Internal | Our own data — TRUST-LEDGER, pipeline metrics. N=1 system, self-assessed. | Average QA score 85.3 across 15 reports (TRUST-LEDGER.json) |

This report was produced using a **multi-agent research pipeline** — the same pipeline being studied. The circularity is acknowledged and discussed in the Adversarial Self-Review (Section 8).

## 2. Executive Summary

**We measured our own AI pipeline across 25 reports. Quality does not compound. Efficiency does. The distinction matters more than most AI narratives acknowledge.**

- **QA scores are flat, not rising:** average 85.3 across 15 measured reports. First five: 84.8. Last five: 83.4. The trend is slightly negative, not compounding.[Internal]
- **Template compliance improved from 7/10 to 9/10** — the locked template (Decision D-157) created a quality floor, not a quality ceiling. Formatting got better. Thinking did not.[Internal]
- **Token costs dropped 50.8%** (18,500 → 9,100 context tokens) via architectural optimization — a genuine efficiency compound, not a quality compound.[Internal]
- **Source quality remained flat:** peer-reviewed percentage stayed at 20-30% across all reports. Source count fluctuated with topic (11-21), not with pipeline maturity.[Internal]
- **External research confirms the pattern:** LLM scaling shows diminishing returns (PNAS 2025)[1], fine-tuning saturates at ~6,500 samples (arXiv 2024)[2], and AI temporal quality degradation is empirically documented (Nature 2022)[3]

***Keywords:*** *AI Quality Measurement, Longitudinal Study, Diminishing Returns, Pipeline Assessment, Self-Evaluation Bias, Template Compliance, Knowledge Compounding*

## 3. Methodology

This report synthesizes three data sources: (1) internal pipeline metrics from TRUST-LEDGER.json (15 reports with QA scores, confidence ratings, source counts, claim counts, runtimes); (2) structural analysis of report HTML files comparing AR-001, AR-005, and AR-025 against five quality dimensions; (3) external research on LLM quality improvement, diminishing returns, and temporal degradation. The pipeline producing this report is the same pipeline being studied — a limitation that is analyzed rather than hidden.

**Limitations:** QA scores are self-assessed by the pipeline (no external validation). Only 15 of 25 reports have TRUST-LEDGER entries. The "blind comparison" is conducted by the same AI system, making true blindness impossible. The sample size (N=1 pipeline, 25 reports) is insufficient for statistical significance. These limitations are not disclaimers — they are findings.

## 4. The Data: 15 Reports, 15 QA Scores, One Uncomfortable Truth Internal

*(Confidence: Internal — self-reported QA scores from TRUST-LEDGER.json)*

**The TRUST-LEDGER records 15 reports with QA scores ranging from 79 to 92. The average is 85.3. The trend is flat — or slightly declining.**

85.3

Average QA score across 15 reports

TRUST-LEDGER.json (verified: sum 1279 / 15 = 85.27)

79–92

QA score range (AR-007 lowest, AR-006 highest)

13-point spread, no convergence trend

–1.4

Average QA decline, first 5 vs last 5 reports

84.8 avg (AR-001–005) → 83.4 avg (AR-011–015)

### Exhibit 1: QA Score Timeline — All 15 Reports

| REPORT | QA SCORE | CONFIDENCE | SOURCES | CLAIMS | KNOWN ISSUES |
|---|---|---|---|---|---|
| AR-001 | 82 | 72 | 12 | 15 | Web search limited; transitions weak |
| AR-002 | 88 | 78 | 14 | 18 | Economic model visualization needed |
| AR-003 | 82 | 75 | 16 | 22 | Fast-moving regulatory updates |
| AR-004 | 87 | 80 | 15 | 20 | Maturity scoring needs validation |
| AR-005 | 85 | 77 | 13 | 19 | Jurisdiction variance |
| AR-006 | 92 | 85 | 18 | 24 | None |
| AR-007 | 79 | 70 | 11 | 16 | Weak transitions; poor flow |
| AR-008 | 91 | 82 | 17 | 21 | Board-level language needs audience validation |
| AR-009 | 91 | 84 | 19 | 23 | None |
| AR-010 | 85 | 72 | 21 | 27 | High claim count, evidence pressure |
| AR-011 | 85 | 75 | 15 | 20 | Alert fatigue data from security domain |
| AR-012 | 83 | 75 | 14 | 18 | Klarna reversal unverified; McKinsey synthesized |

| | | | | | |
|---|---|---|---|---|---|
| AR-013 | 80 | 72 | 13 | 17 | LangChain adoption approximated |
| AR-014 | 84 | 82 | 16 | 19 | None |
| AR-015 | 85 | 78 | 16 | 12 | N=1 system; self-reporting bias |

*Source: TRUST-LEDGER.json, verified calculation. Agent_reputation.avg_qa reports 85.3.*

**Exhibit 2: QA Score by Cohort**

| COHORT | REPORTS | AVG QA | AVG CONFIDENCE | AVG SOURCES |
|---|---|---|---|---|
| Early (AR-001–005) | 5 | 84.8 | 76.4 | 14.0 |
| Middle (AR-006–010) | 5 | 87.6 | 78.6 | 17.2 |
| Late (AR-011–015) | 5 | 83.4 | 76.4 | 14.8 |

*Source: Author calculation from TRUST-LEDGER.json data*

The pattern is not a steady climb. It is a rise (84.8 → 87.6) followed by a decline (87.6 → 83.4). The middle cohort benefited from strong topics (Security Playbook at 92, Calibration Gap and Governance at 91 each). The late cohort regressed. This is **topic dependence**, not quality compounding.

CLAIM

QA scores across 15 reports show no statistically significant upward trend. The average is 85.3 with a standard deviation of approximately 4.0. The pattern is consistent with random variation around a fixed mean, not with compounding improvement.

### WHAT WOULD INVALIDATE THIS?

External review of the same 15 reports producing QA scores that show a clear upward trend ($r > 0.5$, $p < 0.05$). Alternatively, if AR-016 through AR-025 (not in TRUST-LEDGER) show consistently higher scores, the late-cohort decline could be an artifact of incomplete data.

### SO WHAT?

A pipeline that produces 85/100 quality reports is valuable. A pipeline that claims to compound quality over time but actually doesn't is making a false promise. The honest framing: this is a consistent-quality production system, not a learning system.

## 5. The Blind Test: Report #1 vs Report #25 Internal

*(Confidence: Internal — self-assessment, no true blinding possible)*

**We compared AR-001 (first report) and AR-025 (latest report) across five quality dimensions. AR-025 scores 17% higher overall — but the improvement is almost entirely in template compliance and honesty, not in research depth.**

**Exhibit 3: Blind Comparison — AR-001 vs AR-005 vs AR-025**

| DIMENSION | AR-001 (FIRST) | AR-005 (MIDDLE) | AR-025 (LATEST) | DELTA |
|---|---|---|---|---|
| Structure Quality | 8/10 | 8/10 | 9/10 | +1 |
| Source Quality | 6/10 | 6/10 | 7/10 | +1 |
| Claim Precision | 7/10 | 7/10 | 8/10 | +1 |
| Honesty / Limitations | 8/10 | 7/10 | 9/10 | +1 |
| Template Compliance | 7/10 | 8/10 | 9/10 | +2 |
| **Total** | **36/50** | **36/50** | **42/50** | **+6 (+17%)** |

*Source: Internal structural analysis of report HTML files. Self-assessed — bias acknowledged.*

### What Improved

**Template compliance (+2 points):** AR-001 used "What Must Change" instead of "Recommendations" as heading. Quote was from "This Report" (later ruled must be external). Section ordering was pre-standardization. AR-025 follows the locked template almost perfectly.

**Honesty (+1 point):** AR-025 marks its own experiment as "Internal" confidence, discloses N=1 limitations three times, labels its 10x claim as "hypothesis, not measurement." AR-001 was honest but less systematic about it.

**Claim precision (+1 point):** AR-025 uses specific numbers with clear provenance ("73% higher emergence, 250 data points"). AR-001 was also specific ("84% of scenarios, 9 models, 351 scenarios") but mixed in more vague claims ("95% fail" from secondary source).

### What Did Not Improve

**Source quality (+1 point, marginal):** AR-001 used 12 sources, ~20% peer-reviewed. AR-025 used 21 sources, ~30% peer-reviewed. The improvement is real but modest — and source count depends on topic availability, not pipeline maturity.

**Originality (not scored):** AR-001 introduced a novel framing (trust as three-layer problem). AR-025 introduced a novel framework (KCI). Both are original. This dimension varies by topic, not by report number.

CLAIM

The 17% improvement from AR-001 to AR-025 is real but primarily structural. Template compliance accounts for 2 of 6 gained points. Honesty disclosure accounts for 1 point. Source quality and claim precision each gained 1 marginal point. No dimension shows exponential improvement.

WHAT WOULD INVALIDATE THIS?

An external blind review where 10 readers rate AR-001 and AR-025 without knowing which came first. If AR-025 is rated significantly higher on content quality (not formatting), the "only formatting improved" thesis is wrong.

SO WHAT?

The pipeline learned to format better. It did not learn to think better. This is the critical distinction between process improvement and intellectual improvement — and most AI quality claims conflate the two.

## 6. What Actually Compounds (And What Doesn't)

70%

*(Confidence: Medium — internal data + external research)*

**Three things compound in an AI content pipeline. Three things do not. The industry narrative conflates all six.**

### Compounds ✅

**1. Template compliance.** The locked template (Decision D-157, TRUST-LEDGER) ensures every new report inherits all structural learnings. Kintsugi principle: "Every correction now compounds. Future reports inherit all learnings."[Internal] This is genuine compounding — but of format, not content.

**2. Operational efficiency.** Context tokens dropped 50.8% (18,500 → 9,100) via progressive disclosure architecture.[Internal] Runtime stabilized. Cost per report dropped from an estimated $3.50 to $2.75. These are real, measurable efficiency gains that compound with each report produced.

**3. Process discipline.** Confidence badges on every section, invalidation-before-so-what ordering, claim registers, transparency notes — all became standard through iterative feedback. Decision D-152 ("Invalidation BEFORE So What") was a Florian correction that now applies to all reports permanently. This is process learning, captured in the template.

### Does Not Compound ❌

**4. Research depth.** QA scores are flat at 85.3. The LLM does not "remember" insights from prior reports. Each report starts a fresh research session. Report #25 has no more domain expertise than report #1 — it just wears a better suit.

**5. Source quality.** Peer-reviewed percentage stays at 20-30%. Source count fluctuates with topic (11-21). The pipeline cannot access paywalled journals, does not build a citation network across reports, and does not accumulate domain-specific source relationships.

**6. Originality.** Novel framings (AR-001's three-layer trust stack, AR-009's calibration gap, AR-025's KCI framework) are topic-dependent, not iteration-dependent. The pipeline does not build on prior insights — it generates new ones from scratch each time.

**Exhibit 4: Compounding Assessment — Six Dimensions**

| DIMENSION | COMPOUNDS? | MECHANISM | EVIDENCE |
|---|---|---|---|
| Template Compliance | Yes | Locked template inherits all corrections | 7/10 → 9/10 across reports |
| Operational Efficiency | Yes | Architecture optimization (INDEX.md pattern) | 50.8% token reduction |
| Process Discipline | Yes | Decisions encoded in TEMPLATE-RULES.md | 12 locked decisions (D-148 to D-157) |
| Research Depth | No | Stateless LLM — no session memory | QA flat at 85.3 (σ ≈ 4.0) |
| Source Quality | No | Web search limits; no citation accumulation | Peer-reviewed % flat at 20-30% |
| Originality | No | Topic-dependent, not iteration-dependent | Novel framings appear randomly |

*Source: Author analysis of TRUST-LEDGER data and report HTML files*

SO WHAT?

The honest framing: we built a formatting machine that got better at formatting. The content quality is consistently good (85/100) but not improving. This is not a failure — consistent 85% quality at decreasing cost is valuable. But calling it "compounding" is misleading. Compounding implies exponential growth. What we have is linear consistency with logarithmic efficiency gains.

## 7. External Evidence: Does Anyone's AI Quality Compound? 72%

*(Confidence: Medium-High — peer-reviewed sources)*

**The external literature confirms our finding: AI systems show diminishing returns on quality while showing compounding returns on efficiency. This pattern is consistent across domains.**

### Diminishing Returns in LLM Scaling

**PNAS (March 2025):** "Scaling language model size yields diminishing returns for single-message political persuasion." Fine-tuning on just 10,000 examples was sufficient to match models fine-tuned with extensive proprietary procedures. Beyond a threshold, more compute does not mean better output.[1]

**ScienceDirect (December 2025):** "The development of LLMs is characterized by non-linear scaling and target scaling, with diminishing returns as models grow larger. The relationship between size and capability varies by specific task."[4]

**arXiv (July 2024):** Fine-tuning with 200 samples improved accuracy from 70% to 88%. But a saturation point was reached at approximately 6,500 samples, "beyond which additional data yields diminishing returns."[2]

### The Compounding Illusion

**Sinha et al. (September 2025):** "Diminishing improvements in single-step accuracy can compound, leading to exponential growth in the length of task a model can complete."[5] This is a critical finding — but note what compounds: task *length*, not task *quality*. The model can do more steps, not better steps.

The paper also reveals a "self-conditioning effect" — models become more likely to make mistakes when the context contains their prior errors. This directly maps to our pipeline: if a report contains a flawed framing, future reports that reference it compound the error, not the insight.

### Temporal Quality Degradation

**Nature (July 2022):** "Temporal quality degradation in AI models" documents that trained models degrade over time even under minimal data drift. The study tested 32 datasets with 4 standard AI models and found degradation patterns are consistent and predictable.[3] This is the opposite of compounding — it is decay.

### The Developer Learning Curve

**Gartner (via Medium, June 2025):** "It takes an average of 8 months just to move from AI prototype to production, with 30% of generative AI projects expected to be abandoned after proof of concept."[6] The learning curve for AI systems is real — but it's the *human operators* who learn, not the AI. Our template improvements (human decisions encoded) confirm this pattern.

CLAIM

External evidence consistently shows that AI systems exhibit diminishing returns on quality metrics while showing compounding returns on efficiency and scope metrics. Our pipeline data is consistent with this broader pattern: efficiency compounds (50.8% token reduction), quality does not (85.3 average, flat).

WHAT WOULD INVALIDATE THIS?

A longitudinal study (12+ months, 100+ outputs) of an AI content pipeline showing statistically significant quality improvement (not just efficiency) over time, with external evaluation (not self-assessment). We found no such study.

SO WHAT?

The AI industry narrative of "compound improvement" is partially true and partially marketing. What compounds: process, efficiency, scope. What doesn't: depth, originality, source quality. Organizations investing in AI pipelines should optimize for the dimensions that actually compound (template, process, cost) and invest human effort in the dimensions that don't (research depth, critical thinking, source curation).

## 8. Adversarial Self-Review 55%

*(Confidence: Medium — self-critique, inherently limited)*

**This section subjects our own study to four adversarial perspectives. The result: the study design has serious limitations that we cannot resolve from inside the system.**

### As Scientist: "Is the Study Design Valid?"

**No, not by academic standards.** The same system that produced the reports is grading them. There is no inter-rater reliability. The QA scores in TRUST-LEDGER were assigned by the pipeline itself — they are self-reported grades, not external validation. TRUST-LEDGER hypothesis H-002 explicitly flags this: "Unsere QA-Scores sind overconfident (79-92 Range zu hoch?)" with a proposed fix (external review at $500-1,000). That hypothesis remains untested after 25 reports.

The sample size (N=1 pipeline, 15 scored reports) is insufficient for regression analysis. The confounding variable — topic difficulty — is not controlled. AR-006 (Security Playbook) scored 92 not because the pipeline improved but because security is a well-documented domain with strong sources. AR-007 (Orchestration) scored 79 because multi-agent orchestration had fewer quality sources available.

### As Skeptiker: "Are You Measuring Quality or Compliance?"

**Mostly compliance.** The QA score weights template adherence, source count, word count, and claim count. These are measurable proxies. A perfectly formatted report with shallow analysis would score 80+. A brilliant report with formatting errors would score lower. The metric optimizes for what is easy to measure (format) not what matters (insight).

The 85.3 average may mean "consistently adequate" rather than "consistently excellent." Without external benchmarking, we cannot distinguish between the two.

### As Investor: "Is 'Efficiency Improves' Enough?"

**No.** Producing reports 50% cheaper is table stakes. If the reports are 85/100 quality at report #1 and 85/100 quality at report #25, the system is a commodity text generator with a professional template — not a compounding intelligence system. The moat question: what prevents a competitor from achieving the same quality on day one by copying the template?

Answer: nothing. The template is the moat, and templates are copyable. True compounding would require accumulated domain expertise, proprietary source networks, or calibration data — none of which our pipeline builds.

### "What Is the Difference Between 'Getting Better' and 'Getting Different'?"

The pipeline didn't get better — it got more consistent. The template locked the quality floor, not the ceiling. Reports became more alike (convergence toward template), not better (improvement toward excellence). Standardization is not the same as quality. A McDonald's hamburger is standardized. That doesn't make it a good hamburger.

> SO WHAT?
>
> This adversarial review reveals that the study's biggest finding may be its own limitations. The pipeline cannot objectively measure its own quality. The QA scores are probably overconfident (H-002). The blind comparison is not truly blind. And the most interesting question — does the content actually help anyone make better decisions? — is not measured at all. That's the real quality metric, and we don't have it.

# 9. Recommendations

**Scope:** These recommendations apply to any team operating an AI content pipeline and wondering whether quality is improving over time.

## For AI Pipeline Operators

1. **Separate efficiency metrics from quality metrics.** Token cost, runtime, and template compliance compound. Research depth, source quality, and originality do not. Report them separately. Do not let improving efficiency scores mask flat quality scores.
2. **Get external validation.** Self-assessed QA scores are unreliable (AR-009: 84% of LLM outputs are overconfident). Budget $500-1,000 for external review of 5 reports. This is the single highest-ROI quality investment available and remains untested in our own system (H-002).
3. **Build a citation accumulation layer.** The pipeline is stateless — report #25 cannot cite report #1. Implement a vector database of prior research, a claim registry that prevents re-citing debunked claims, and a source relationship graph. This is the infrastructure that would enable actual quality compounding.

## For AI Quality Researchers

1. **Distinguish between process compounding and content compounding.** Most "AI improvement over time" studies measure process metrics (speed, cost, compliance). Content quality metrics (accuracy, depth, originality) should be tracked separately with external evaluation.
2. **Publish longitudinal quality data.** We found zero longitudinal studies of AI content pipeline quality with external evaluation. This is a significant gap in the literature.
3. **Test the self-conditioning effect.** Sinha et al. (2025) showed models make more errors when context contains prior errors.[5] This has direct implications

for pipelines that self-reference: compounding errors may be more likely than compounding insights.

## For Anyone Claiming "AI Quality Compounds"

1. **Show the data.** What specific metric improved? Over what time period? With what evaluation methodology? "It got better" is not evidence.
2. **Control for topic difficulty.** A pipeline producing security reports (well-documented domain) will score higher than one producing novel framework reports (poorly documented domain). Without controlling for topic, quality trends are meaningless.
3. **Distinguish template from thinking.** If the improvement is "reports look more professional now," that is template compounding, not quality compounding. These are different claims with different value propositions.

## 10. Predictions BETA

These predictions will be scored publicly at 12 months. Version 1.0 (February 2026).

| PREDICTION | TIMELINE | CONFIDENCE |
|---|---|---|
| Our own QA scores remain in the 80-90 range for the next 25 reports without architectural changes to the pipeline | 6 months | 80% |
| Implementing a citation accumulation layer (vector DB of prior research) improves self-reference ratio from ~0% to >20% within 3 months | Q2 2026 | 65% |
| External review of our reports produces QA scores 10-15 points lower than self-assessed scores | When tested | 60% |
| At least one major AI lab publishes longitudinal quality data (not just benchmark scores) for their content pipeline | 12 months | 30% |

# 11. Transparency Note

This section discloses the methodology, confidence calibration, and known limitations of this report.

| | |
|---|---|
| **Overall Confidence** | 62% — Medium. The internal data is complete but self-assessed. The external research is strong but not directly comparable. The study design has fundamental limitations (self-grading, no control, N=1) that cannot be resolved from within the system. |
| **Sources** | 6 external sources (peer-reviewed: PNAS, Nature, ScienceDirect; preprints: arXiv ×2; industry: Gartner) + internal pipeline data (TRUST-LEDGER.json, 15 report entries, 25 report HTML files). Full citations in Section 13. |
| **Strongest Evidence** | The TRUST-LEDGER QA score data (15 entries, verifiable calculation) and the external diminishing returns research (PNAS peer-reviewed, Nature peer-reviewed). |
| **Weakest Point** | The blind comparison (Section 5) is self-assessed. The QA scores themselves may be overconfident (H-002 untested). Only 15 of 25 reports have TRUST-LEDGER entries — AR-016 through AR-025 data is missing. |
| **What Would Invalidate** | External review showing clear quality improvement trend across the 25 reports. Or: demonstrating that the self-assessed QA scores accurately reflect external quality perception (validating the self-assessment methodology). |
| **Methodology** | This report was produced using the same multi-agent research pipeline being studied. The circularity is a limitation, not a feature. The pipeline used web search for external sources, file reads for internal data, and structural analysis of HTML files for the blind comparison. |

**System Disclosure**

This report was created with a multi-agent research system. It is the 29th report produced by this system, making it simultaneously the subject and product of its own study.

## 12. Claim Register

### Exhibit 5: Claim Register — Top 12 Claims

| # | CLAIM | VALUE | SOURCE | CONFIDENCE | USED IN |
|---|---|---|---|---|---|
| 1 | Average QA score across 15 reports is 85.3 | 85.3 | TRUST-LEDGER.json (verified) | Internal | Sec 2, 4 |
| 2 | QA scores show no upward trend (late cohort lower than middle) | 84.8→87.6→83.4 | TRUST-LEDGER.json | Internal | Sec 4 |
| 3 | Token costs dropped 50.8% | 50.8% | TRUST-LEDGER economics | Internal (verified) | Sec 2, 6 |
| 4 | Template compliance improved from 7/10 to 9/10 | +2 points | Report HTML analysis | Internal | Sec 5 |
| 5 | LLM scaling yields diminishing returns on persuasion | Diminishing | PNAS 2025 (peer-reviewed) | High | Sec 7 |
| 6 | Fine-tuning saturates at ~6,500 samples | ~6,500 | arXiv 2407.13906 | Medium | Sec 7 |
| 7 | AI models show temporal quality degradation | Degradation | Nature 2022 (peer-reviewed, 32 datasets) | High | Sec 7 |

| | | | | | |
|---|---|---|---|---|---|
| 8 | Single-step accuracy gains compound into task length, not quality | Length ≠ Quality | Sinha et al. arXiv 2025 | Medium | Sec 7 |
| 9 | Self-conditioning: models make more errors when context contains prior errors | Directional | Sinha et al. arXiv 2025 | Medium | Sec 7 |
| 10 | 84% of LLM outputs are overconfident | 84% | PMC/12249208 (9 models, 351 scenarios) | High | Sec 8 |
| 11 | QA self-assessment is likely overconfident (H-002 untested) | Hypothesis | TRUST-LEDGER hypotheses | Low (untested) | Sec 8 |
| 12 | Prototype-to-production takes 8 months average | 8 months | Gartner via Medium (644 respondents) | Medium | Sec 7 |

**Top 5 Claims — Invalidation Conditions:**

1. **Claim #1 (85.3 average):** Invalidated if recalculation of TRUST-LEDGER data produces a different average (verified: 1279/15 = 85.27, rounded to 85.3).
2. **Claim #2 (No upward trend):** Invalidated if AR-016–025 TRUST-LEDGER entries show consistent scores >88, shifting the trend upward.

3. **Claim #3 (50.8% token reduction):** Invalidated if actual token usage measurements differ from TRUST-LEDGER economics data.
4. **Claim #5 (Diminishing returns):** Invalidated if subsequent research demonstrates sustained linear quality scaling with model size (contradicting PNAS finding).
5. **Claim #11 (Self-assessment overconfident):** Validated or invalidated by implementing H-002 (external review at $500-1,000). Currently untested.

# 13. References

[1] Bai, H. et al. (2025). "Scaling language model size yields diminishing returns for single-message political persuasion." *Proceedings of the National Academy of Sciences (PNAS)*. https://www.pnas.org/doi/10.1073/pnas.2413443122

[2] Patel, A. et al. (2024). "Crafting Efficient Fine-Tuning Strategies for Large Language Models." *arXiv:2407.13906*. https://arxiv.org/html/2407.13906v1

[3] Vela, D. et al. (2022). "Temporal quality degradation in AI models." *Scientific Reports (Nature)*, 12, 11654. https://www.nature.com/articles/s41598-022-15245-z

[4] Chen, X. et al. (2025). "Breaking Myths in LLM scaling and emergent abilities with a comprehensive statistical analysis." *ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S092523122503214X

[5] Sinha, A. et al. (2025). "The Illusion of Diminishing Returns: Measuring Long Horizon Execution in LLMs." *arXiv:2509.09677*. University of Cambridge / MPI. https://arxiv.org/html/2509.09677v1

[6] Schiff, S. (2025). "The AI Learning Curve: Why Benefits Will Lag Behind Capabilities." *Medium*, citing Gartner research (644 respondents). https://medium.com/@sschiff/the-ai-learning-curve-why-benefits-will-lag-behind-capabilities-c5fcca5b27c2

[7] Muennighoff, N. et al. (2025). "Scaling Data-Constrained Language Models." *Journal of Machine Learning Research*, 26, 1-66. https://www.jmlr.org/papers/volume26/24-1000/24-1000.pdf

[8] ACM/IEEE ESEM (2024). "Continuous Quality Improvement of AI-based Systems: the QualAI Project." *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. https://dl.acm.org/doi/10.1145/3674805.3695393

[9] Ainary Research (2026). *State of AI Agent Trust 2026.* AR-001.

[10] Ainary Research (2026). *The Knowledge Compounding Flywheel.* AR-025.

[11] Ainary Research (2026). *META-LEARNINGS: What Our Own Research Teaches Us About Building Better AI Agents.*

*Ainary Research (2026). Does AI Quality Actually Compound? A 25-Report Longitudinal Study. AR-029.*

**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

● Ainary

AI Strategy · Published Research · Daily Intelligence

Contact · Feedback

ainaryventures.com
florian@ainaryventures.com