What AI Found, What We Already Knew

Chungil Chae

Fri, 3 April 2026

What AI Found, What We Already Knew

Interpreting a Paper Autonomously Generated from Data by an AI-Scientist

Date written: 2026-04-03

Opening: We Handed the Data to a Machine

We ran 36 models. Across 5 methodological paradigms, drawing on 16 theories, we asked the same question of 282 cases three times over. We organized the results into two reports (a strategy edition and an interpretation edition).

Then we gave the same data to an **AI-Scientist**. "Analyze it yourself, and write a paper."

The AI chose K-means clustering on its own. It autonomously carried out 5 experiments. It wrote an 8-page paper. The cost was $8, the time 20 minutes.

What that AI found **overlaps exactly** with one piece of the conclusion we had already reached.

Chapter 1. What the AI Saw

The Choice of Clustering

The AI had no theoretical guidance whatsoever. It knew nothing of TOE, of P-E Fit, of absorptive capacity. All we gave it were two seed_ideas (nonlinear extension of TOE, dose-response separation) as hints.

The AI generated a third idea on its own: **"Classify firms with K-means, and analyze each cluster separately."** Its scores were Interestingness 9, Feasibility 9, Novelty 8. Among the four ideas, it gave this one the highest rating of all—to itself.

This is the same intuition we had pursued with T2 (LPA) in our first-round analysis. The sense that there are **hidden types** within the data. The suspicion that beneath the average lies heterogeneity.

Three Profiles

The 3 clusters the AI classified:

Cluster	n	DT awareness	Smart systems	Firm size (log)	Q3_difference
0 (low readiness)	122	1.84	16.1	2.51	1.525
1 (high readiness)	51	3.24	25.4	3.34	1.498
2 (medium readiness)	70	1.89	21.3	3.73	1.247

Here the first odd point appears. **The Q3_difference of the high-readiness firms (cluster 1) is almost identical to that of the low-readiness firms (cluster 0).** 1.498 vs 1.525. DT awareness is twice as high and smart systems are 60% greater, yet the training effect is the same.

This is exactly what we found in T2 (LPA) of the first-round analysis: *"A more prepared firm does not necessarily learn better."* It is the very pattern the interpretation report explained as the "large-firm paradox" and the "ceiling effect."

The Dramatic Gap of R^2 .002 → .336

The AI's most impactful finding:

- **Cluster 0** (low readiness, n=122): $R^2 = .002$. No variable explains the training effect.
- **Cluster 1** (high readiness, n=51): $R^2 = .336$. Smart systems (+) and firm size (-) are strong predictors.

- **Cluster 2** (medium readiness, n=70): R^2 = .018. Again, nothing explains anything.

What was R^2 = .029 across the entire dataset jumps to .336 once only the high-readiness firms are isolated. **12 times higher.** The AI called this "dramatic heterogeneity."

Chapter 2. What We Already Knew

When the AI's findings are set against our 36 analyses, the puzzle pieces fall into place.

Comparison 1: "Predictors Work Only in High Readiness"

The AI's finding: R^2 is meaningful only in cluster 1.

Our findings: - **T2 (LPA)**: The difference in training effect across profiles was not significant, but the **mechanism** differed by profile. - **T3/T20 (QCA)**: Equifinality. There are multiple paths to a high training effect. ~DT_AWARE (low awareness) is also a valid path. - **combo1 (RSM)**: The P-E Fit interaction term (p=.002) is significant only when "awareness and infrastructure are **both** high at the same time."

What the AI found with K-means is **another expression** of what we had reached through QCA's equifinality and RSM's interaction term. "It works only in high readiness" and "the awareness × infrastructure interaction term is significant" are different angles on the same phenomenon.

Comparison 2: "The Negative Effect of Firm Size"

The AI's finding: in cluster 1, firm size = -0.839 (p<.001). Even among high-readiness firms, the larger the size, the **lower** the training effect.

Our findings: - **T14 (ANOVA)**: The large-firm paradox. An adverse effect in firms of large size. - **T15**: SF + multiple-participation reverse synergy. The limits of additional input in firms that already have a lot. - **RSM (B1)**: An inverted-U for firm size (p=.029), optimal at ~13 people.

The AI's = -0.839 corresponds to the **right-hand declining segment of the inverted-U**. Because high-readiness (cluster 1) firms have a large average size (log 3.34 28 people), they have already passed the optimum of the inverted-U (~13 people). They are in the region of diminishing returns.

This is **one side** of what the interpretation report explained as "the dual mechanism of Liability of Smallness + ceiling effect"—and the AI captured it.

Comparison 3: "A Difference in Slopes, Not Intercepts"

The AI's most refined finding: adding cluster dummy variables yields non-significance (p=.856), but adding cluster × variable interaction terms yields significance (p=.043).

This means that while the **mean training effect of the three groups is similar** (no difference in intercepts), **the mechanism that determines the training effect differs** (a difference in slopes).

The integrative proposition of the interpretation report was precisely this:

> "The effect of SME digital transformation training is not a matter of 'present or absent,' but appears 'when awareness and infrastructure meet.' "

Not "present or absent" (difference in intercepts) → consistent with the AI's non-significant dummy (p=.856). "When they meet" (conditional) → consistent with the AI's significant interaction term (p=.043).

Chapter 3. What the AI Missed

The reason the AI's paper received a Reject (an internal review—it reviews what it itself wrote, based on a model built from CS-side papers) is not simply a journal mismatch. **The things essential to a social-science paper are missing.**

Missing 1: Theory

The AI's paper has no theory. There is a **description**—"there is heterogeneity"—but no **explanation** of "why heterogeneity arises."

We drew on 16 theories. P-E Fit explained the interaction term, diminishing returns the inverted-U, equifinality the multiple paths, experiential learning the dose-response. **A finding without theory is no more than a pattern.** That the AI found R^2 .002 → .336 is impressive, but without a theoretical answer to "why do predictors work only in high readiness?" it remains an **unfinished finding**.

The answer: it is P-E Fit. Only in firms where the fit between awareness (P) and infrastructure (E) is high does the mechanism activate by which additional organizational characteristics (smart systems, firm size) predict the training effect. In firms with low fit, the mechanism itself is "switched off."

Missing 2: Triangulation

The AI reached its conclusion with K-means alone. We applied 5 methodologies to the same question: - **LPA** (person-centered) → profile types - **QCA** (set-theoretic) → condition combinations - **RSM** (nonlinear) → interaction terms / inverted-U - **DID/PSM** (causal inference) → treatment effects - **CLPM** (longitudinal) → causal direction

A single K-means silhouette of .293 may be called "reasonable," but **is that pattern confirmed by other methodologies too?** Our answer was "yes, it converges across all 5." The AI's answer remained at the level of "I'm not sure—I changed the seed and it came out similar."

Missing 3: The "Limits of Measurement" Insight

One of our most important findings: **Training is working. It is only that Q3_difference fails to capture it.**

- Smart systems as DV → $R^2 = .517$
- Q3_difference as DV → $R^2 = .136$

With the same IVs, simply changing the DV makes a **fourfold** difference in explanatory power. This is the "dose-response separation" phenomenon that converged through 5-methodology triangulation (OLS, PSM, DID, CLPM, RSM).

The AI was unable even to pose this question. It used only Q3_difference as the dependent variable and never conceived the idea, "what if we change to a different DV?" This is the **absence of domain knowledge**. Its ability to optimize within the data is outstanding, but the ability to **imagine outside the data** is not yet there.

Missing 4: Discussion

The heart of a social-science paper is the Discussion. "What does this finding add to existing theory?" "What does it mean for practitioners?" "What are the policy implications?"

The AI's Conclusion stopped at a summary of results plus a list of limitations. The **prescriptive conclusion** our interpretation report reached—the word "tailored," the argument that we must diagnose a firm's current position and design interventions that supplement the missing dimensions—was absent from the AI.

Chapter 4. So What Does This Mean?

The Real Value of the AI-Scientist

The R^2 .002 → .336 that the AI found in 20 minutes and $8 is one piece of the conclusion we reached in 3 months with 36 models. About one-fifth of the whole. But **the direction was exactly right**.

What this means: 1. **Value as an accelerator of exploration**: A workflow is possible in which the AI first scans for heterogeneity through clustering, and the researcher then layers on theory and triangulates. 2. **Value as a hypothesis-generation tool**: The AI's finding that "it works only in high readiness" provokes the next question, "why?" Answering that question with P-E Fit is the human's part. 3. **Value as a pattern-confirmation tool**: The AI independently confirms the conclusion we reached through 36 analyses. This is a kind of **methodological triangulation**.

The Limits of the AI-Scientist

1. **Absence of theory**: It finds patterns but cannot explain them.
2. **Absence of DV imagination**: It optimizes only within the given DV.
3. **Absence of literature**: There is no dialogue with prior research.
4. **Absence of Discussion**: It cannot answer "so what?"
5. **Absence of triangulation**: It relies on a single methodology.

All five of these are problems of **domain knowledge**. Its capacity for statistical pattern recognition already approaches or exceeds the human (20 minutes vs 3 months). What it lacks is knowing "what matters in this field."

The Next Step: AI-Scholar

Filling this gap is the development goal of the AI-Scholar.

- Theory Engine: theory DB + pattern–theory matching
- Literature Engine: SSCI literature search + citation generation
- Analysis Engine: automated triangulation (QCA + LPA + RSM + SEM)
- Writing Engine: a Theory → Hypotheses → Discussion structure

What the AI-Scientist demonstrated is **possibility**. In 20 minutes and $8, it can make a finding that points in the right direction. Layer theory and context onto that, and the automatic generation of a social-science paper is not impossible.

Closing: The Machine's Eye and the Researcher's Eye

The AI saw R^2 .002 → .336. We saw "when awareness and infrastructure meet."

The AI wrote "dramatic heterogeneity." We arrived at the word "tailored."

The AI found the patterns. We named them, gave them meaning, and proposed what to do next.

Both of us looked at the same data, but **we look at different depths**. The AI's depth is deepening fast. Yet the questions "why?" and "so what?" still belong to the human domain.

This preprint is the first social-science test of the AI-Scientist, and the starting point for developing the AI-Scholar. How the machine's eye and the researcher's eye can collaborate is what we must now design.