

When the Method Disagrees with the Plan: AI-Moderated Interviews Surface an Emergent Disconfirmation, and a Belief/ Reflex Gap in What People Inherit from Caregivers

Justin Hoang,¹ Benjamin Lo,¹ and Nicole Zeng¹

¹ *The Observatory and Dialogue AI*

Corresponding author: [email placeholder]

Abstract

Standardized, AI-moderated interviewing promises survey-scale qualitative data without surrendering depth, but it raises a validity worry: does a scripted machine moderator simply confirm whatever the study set out to find? We report a small (N = 26) AI-moderated interview study on what adults inherit from their caregivers, designed to probe contradiction and emergent meaning. The method produced two results that bear on its own credibility. First, it surfaced an emergent disconfirmation of the study's framing: a substantial subset of participants kept their inherited rules, reported no contradiction, and in one case rejected the study's therapeutic premise outright, and in another the participant refused the moderator's attempt to impose a say-do contradiction across four turns. Second, it yielded a substantive pattern stated here at descriptive strength: participants update their stated beliefs about an inherited rule far faster than they change the behavior, a dissociation between an updatable value system and a slower behavioral reflex. We document a three-component interview-quality score (signal, moderator, usability; weights 50/30/20; study mean 3.96) and treat the moderator-over-steer case as a measured feature, not a flaw. The contribution is primarily methodological (an instrument that can disconfirm its own design and expose its own leading) and secondarily theoretical (the belief/reflex dissociation as a transmission mechanism). We make no causal or population claims.

Keywords: AI-moderated interviews, qualitative methods, large language models, intergenerational transmission, say-do gap, research credibility

1. Introduction

Qualitative interviewing buys depth at the cost of scale, and surveys buy scale at the cost of depth. AI-moderated interviewing, in which a language-model moderator runs a standardized but adaptive conversation, is being adopted rapidly as a way to have both [CITATION NEEDED: industry adoption figures for AI-moderated interviewing]. The promise is depth at survey scale with emergent themes a fixed-choice instrument cannot pre-list. The worry is the mirror image: a standardized machine moderator, tuned to "surface contradictions," may manufacture the very contradictions the study hoped to find, a confirmation engine wearing the costume of a method.

This paper uses a single small study to ask whether an AI-moderated instrument can fail honestly: can it return results that cut against its own design, and can it expose its own leading? We find that it can, on both counts, and that documenting those failures is what licenses the substantive finding the same study produced.

Contribution statement. This paper contributes, first, a methodological observation: a standardized AI-moderated interview protocol surfaced an emergent disconfirmation of its own guiding frame and, in a documented case, a participant refusing a moderator-imposed interpretation, both captured by a three-component interview-quality score that localizes where an interview is weak. Second, it contributes a descriptive substantive finding: a dissociation between rapidly-updating stated belief and slow-changing behavioral reflex in how adults carry inherited caregiving rules forward. Both claims are bounded to a single-platform, cross-sectional sample of 26; we make no causal or generalization claim.

The remainder positions the work against AI-interview and qualitative-method scholarship (§2), details the design, sample, coding, and quality scoring (§3), reports the priors-vs-emergence ledger and the substantive findings (§4), and discusses what an honestly-failing instrument implies for the method (§5), before stating the limitations that bound every claim (§6).

2. Related work and background

A fast-growing literature applies large language models to qualitative work, mostly to *analysis*: inductive thematic coding and codebook development with LLMs (De Paoli, 2024; Gao et al., 2024), with recurring cautions that the approach can flatten interpretive nuance and requires human oversight (De Paoli, 2024). A smaller strand examines LLMs in the *data-collection* role, as adaptive conversational interviewers (Kreuter, 2025), alongside evidence that generative AI in open-ended response settings can homogenize what participants produce (Zhang, Xu, and Alvero, 2025). Our study sits in this second strand but turns the lens onto a specific validity threat: not whether the AI moderator can elicit data, but whether it can elicit data that *disconfirms the researcher's frame*.

Methodologically we build on emergent, blind-to-hypothesis coding in the grounded-theory tradition (Glaser and Strauss, 1967; Braun and Clarke, 2006), and on the front-

stage/back-stage distinction (Goffman, 1959) that motivates coding behavior over stated belief. Substantively, the belief/reflex gap we describe is adjacent to long-standing distinctions between explicit and implicit or automatic determinants of behavior [CITATION NEEDED: dual-process / implicit-explicit attitude-behavior literature] and to work on the intergenerational transmission of parenting and family rules [CITATION NEEDED: intergenerational transmission of parenting practices]. We do not test those theories; we report a pattern consistent with them and leave the causal question open. The gap this study fills is narrow and specific: an existence proof that a standardized AI-moderated instrument can return an emergent disconfirmation and surface its own leading, with a quality measure that makes both legible.

3. Methods

3.1 Design

Twenty-six AI-moderated conversational interviews (single recurring model moderator), roughly 10 to 13 minutes each, on a six-section guide: a story anchor, the meaning drawn from it, household rules, transmission (kept / rejected / repeated), a believe-but-struggle-to-practice probe, and a present-day reinterpretation. The moderator was instructed to seek concrete examples, probe emotional language, surface contradictions, and track before-versus-now. The reinterpretation section (asking explicitly "what do you understand now") is the one place the guide invites after-the-fact reflection, which we flag where it bears on a finding.

3.2 Sample and recruitment

The 26 participants are all parents or caregivers, recruited as a convenience sample; the study's own plan targeted 300 across six demographic cells, which this dataset does not fill. The realized sample skews female, White, suburban, two-parent, and "consistent" in self-reported upbringing (full composition in Appendix D). These skews and the small N bound every claim below.

3.3 Analysis

Every interview was read in full and coded emergently (open to axial to synthesis) and, critically, *blind to the study's predictions*. Those predictions were extracted verbatim into a priors ledger before coding and scored only afterward (Appendix A; §4.1), a discipline intended to prevent the analyst from "finding" the hypothesis. Moderator turns were separated from participant turns so that only participant speech was treated as evidence and so that moderator behavior could be scored independently.

3.4 Quality and credibility (IQS)

Each interview received an Interview Quality Score with three components, each 0 to 5: participant Signal depth, Moderator effectiveness (scored from the full dialogue, never from participant content), and Data Usability (where authenticity concerns live). Weights were 50/30/20 (Exhibit 1). Reporting the components separately is deliberate:

the composite hides the diagnostic split, and one interview (P22) scores high on signal while scoring low on moderator effectiveness precisely because the moderator led, the case we treat as a feature in §4.

Exhibit 1. Interview Quality Score, component means (0 to 5), N = 25 usable.



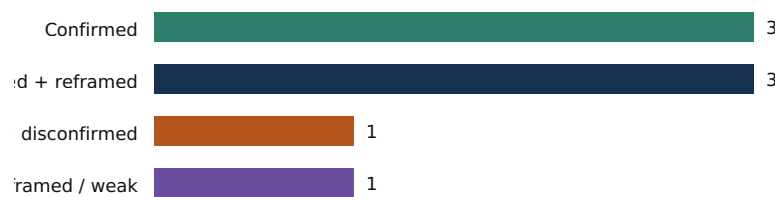
Weights: Signal 50%, Moderator 30%, Usability 20%. Study composite mean 3.96 (range 2.0 to 4.7). One void/dropoff interview (P03) is excluded from means. Source: IQS deliverable.

4. Results

4.1 Plan versus emergence: the instrument disconfirmed its own frame

Of eight predictions extracted from the study plan, the data confirmed some (rules are inherited; rules carry a help/harm duality; the same word splits in meaning), reframed others, and *partly disconfirmed* the central one (Exhibit 2). The plan assumed participants would reject inherited rules yet reproduce them, treating the contradiction as universal. A distinct subset instead kept their rules, reported no contradiction, and passed them on as discipline. One participant rejected the study's implicit therapeutic frame outright (P12, Appendix C). This is the result a confirmation engine could not produce.

Exhibit 2. Priors ledger outcomes (8 predictions scored after blind coding).



Counts of plan predictions by post-coding status. "Partly disconfirmed" is the central reject-but-repeat prediction (H4), which held for many but not all. Source: priors ledger (Appendix A).

4.2 The moderator over-steer, surfaced and measured

In one interview the moderator repeatedly offered a say-do contradiction the participant did not feel, and she declined it across four turns, including: "But I didn't say that I didn't want to repeat my parents' angry reactions." (P22, *I didn't say that I didn't want to repeat my parents' angry reactions*). The interview scores high on signal and low on moderator effectiveness; the IQS localizes the problem rather than hiding it in a composite. Two consequences follow. The refusal is itself evidence the broader findings

are not merely moderator-suggested, and it caps how far any moderator-led agreement elsewhere can anchor a claim.

4.3 *The belief/reflex dissociation*

The most consistent substantive pattern is a gap between updated belief and unchanged behavior. Participants rejected an inherited rule in clear cognitive terms while reporting that they could not stop enacting it: “I no longer believe in stuffing down my emotions or sweeping conflict under the rug, but I do still struggle mightily with reversing those things, even though I believe that I should reverse them.” (P05, *I no longer believe in stuffing down my emotions*). A participant who teaches her own children the opposite rule described reverting under stress: “I know in my head that it's not the best way to handle emotion, but the eight-year-old child in me still exists and that's how I tend to... where I go whenever something negative happens.” (P24, *the eight-year-old child in me still exists*). We state this at descriptive strength: the value system updates on contact with new information, while the behavioral reflex, learned in childhood under pressure, persists. The most frequently described inherited rule was emotional suppression (“Our family is a very much put on a happy face and don't admit anything's wrong kind of family.”, P11), which also carried the clearest help/harm duality. Prevalence is reported as presence in this set, never as an incidence rate (Appendix B).

4.4 *Orientation, not demographics, is the operative split*

Participants clustered behaviorally into Reckoners (reparent and fight the reflex), Inheritors (keep and transmit the rules), and Negotiators (keep the value, adapt the rule), a division that cut across age, gender, and class and did not map onto the planned demographic cells. We present these as descriptive types, not sized segments.

5. Discussion

The methodological reading is the load-bearing one. A standardized AI-moderated instrument returned a result against its own design and exposed an instance of its own leading. That an instrument can fail honestly is precisely what a skeptic of AI-moderated research should want demonstrated, and it is more persuasive than any single clean finding. The blind priors-ledger discipline and the component-separated quality score are the mechanisms that made the disconfirmation and the over-steer visible rather than absorbed.

The substantive reading revises a common framing. Intergenerational “cycles” are often discussed as if insight were the lever: understand the pattern and you change it. Our participants suggest the lever is misplaced. They had the insight and still could not move the behavior, which points interventions toward the moment of reproduction (the catch) rather than the moment of reflection (the lesson). We offer this as a hypothesis the data is consistent with, not a tested claim.

6. Limitations

The threats to validity are real and bound every claim. (1) *Sample*. $N = 26$, a convenience sample skewed female, White, suburban, and two-parent; the planned demographic cells are unfilled, so no segment or population estimate is defensible, and we report no finding as a percentage. (2) *Cross-sectional*. One conversation per person supports no causal claim. (3) *Self-report and say-do*. The belief/reflex gap is reconstructed from what people said about what they do; behavioral observation would test it. (4) *AI-moderation effects*. A single model moderator on one platform; moderator over-steer is documented in at least one case, and register effects may inflate textual authenticity signals for some participants. (5) *Elicitation*. The reinterpretation finding's presence is partly guide-driven. (6) *Scale untested*. The depth and emergence claims are demonstrated at $N = 26$; the plan's scale ambition (300 interviews) is explicitly not tested here.

7. Conclusion

A small AI-moderated study disagreed with its own plan and showed where its moderator led, and in doing so earned the descriptive finding it also produced: people change what they believe about an inherited rule long before they change what they do. The contribution is an existence proof (the method can fail honestly) and a bounded mechanism (belief/reflex dissociation), offered for replication at larger scale and with behavioral measures.

References

- Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101.
- De Paoli, S. (2024). Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review*, 42(4), 997-1019.
- Gao, J., Guo, Y., Lim, G., Zhang, T., Zhang, Z., Li, T. J.-J., and Perrault, S. T. (2024). CollabCoder: A lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
- Glaser, B. G., and Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine.
- Goffman, E. (1959). *The Presentation of Self in Everyday Life*. Doubleday.
- Kreuter, F. (2025). AI conversational interviewing: Transforming surveys with LLMs as adaptive interviewers. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, 179-204. Association for Computational Linguistics.
- Zhang, S., Xu, J., and Alvero, A. J. (2025). Generative AI meets open-ended survey responses: Research participant use of AI and homogenization. *Sociological Methods & Research*, 54(3), 1197-1242.
- [CITATION NEEDED] placeholders above mark claims the authors must ground before submission (AI-interview adoption statistics; dual-process attitude-behavior theory; intergenerational transmission of parenting). No reference here is fabricated; unverifiable claims are flagged, not invented.

Appendix A. Interview protocol (probe areas)

Probe areas and intent (not the verbatim moderator script, not participant answers). Q1 Story anchor: elicit one concrete formative caregiving moment. Q2 Meaning: what it taught, explicit and implicit. Q3 Rules: spoken and unspoken household rules, including emotions, money, conflict. Q4 Transmission: which rules kept, rejected, or repeated. Q5 Tension: a believe-but-struggle-to-practice example. Q6 Reinterpretation: present-day understanding of caregivers (note: invites reflection by design).

Appendix B. Codebook summary

Synthesis codes (prevalence not sized at N = 26).

Code	Definition	Prevalence
Emotional suppression	'Keep everything in', 'happy face', 'we never talked about emotions'. The most frequently described inherited rule; clearest help/harm duality.	not sized (presence-based)
Belief / reflex dissociation	Cognitive rejection of a rule coexists with automatic behavioral reproduction under stress. The belief updates; the reflex does not.	not sized
Orientation: Reckoner / Inheritor / Negotiator	Behavioral stance toward the inheritance: reparent and fight the reflex; keep and transmit as discipline; or keep the value and adapt the rule.	not sized
Reinterpretation-as-sorting	Adult reflection grants understanding to one caregiver and withdraws it from another; socioeconomic reframe the most common warming form.	not sized; presence partly script-elicited
Meaning split: independence / respect	Same word, opposed lived meanings (competence vs. economic armor vs. cage; mutual regard vs. non-questioning deference).	not sized
Loss anchor / parentification	Formative memory is often a parent's death or illness; several participants were caregivers as children.	not sized; emergent, unpredicted

Prevalence is presence-based at small N and reported as "not sized" rather than as a percentage.

Appendix C. Representative verbatims (curated selection, exact and located)

A confidentiality-safe selection, not the full bank. Every quote exact and located.

P##	Demographics	Verbatim	Locator
P05	35-44 F Orlando, FL	"I no longer believe in stuffing down my emotions or sweeping conflict under the rug, but I do still struggle mightily with reversing those things, even though I believe that I should reverse them."	<i>I no longer believe in stuffing down my emotions</i>
P24	55-64 F San Diego, CA	"I know in my head that it's not the best way to handle emotion, but the eight-year-old child in me still exists and that's how I tend to... where I go whenever something negative happens."	<i>the eight-year-old child in me still exists</i>
P08	35-44 F Leander, TX	"I know that I'm not being the best example. I'm not practicing what I am preaching to him."	<i>I'm not practicing what I am preaching to him</i>
P12	45-54 F North Royalton, OH	"They all want to go see a therapist. Then when they go see the therapist, they blame the parents. [...] And I just think it's kind of dumb."	<i>They all want to go see a therapist</i>
P15	45-54 M Lexington, KY	"I still think they're a waste of time, actually. So I've adopted my parents' position on that. But I do them anyway."	<i>I still think they're a waste of time, actually</i>
P22	55-64 F Thornton, CO	"But I didn't say that I didn't want to repeat my parents' angry reactions."	<i>I didn't say that I didn't want to repeat my parents' angry reactions</i>
P20	35-44 F Carol Stream, IL	"My dad I will fault a lot too, but my mom I will not fault anything to."	<i>my mom I will not fault anything</i>
P17	25-34 F Virginia Beach, VA	"They were absent. That was what they said, is they were absent. They were completely gone. We had no parents."	<i>They were absent. That was what they said</i>
P20	35-44 F Carol Stream, IL	"I am still the first person that my siblings will call when they need something."	<i>I am still the first person that my siblings will call when they need something</i>
P11	35-44 F Napa, CA	"Our family is a very much put on a happy face and don't admit anything's wrong kind of family."	<i>put on a happy face and don't admit anything's wrong</i>
P08	35-44 F Leander, TX	"I have learned that not all adults deserve respect."	<i>not all adults deserve respect</i>
P25	35-44 M Bakersfield, CA	"What I observed is [...] parenting is more about sacrifice than love I guess."	<i>parenting is more about sacrifice than love</i>

Full transcripts are withheld to protect participant confidentiality; curated quotes remain exact and located, so the verbatim oath holds. One participant record (P15) arrived with a raw system ID instead of a name and is flagged for scrubbing.

Appendix D. Interview Quality Score, summary

IQS summary (N = 25 usable; P03 void excluded from means).

Component	Weight	Mean (0-5)
Signal depth	50%	3.92
Moderator effectiveness	30%	3.68
Data usability	20%	4.48
Composite IQS		3.96

Band distribution (of 26): high 15, medium 6, low 4, void 1. Range 2.0 to 4.7. Flags: 1 dropoff (P03), 1 truncated recording (P13), 1 PII raw-ID (P15), 1 moderator over-steer (P22).

Appendix E. Demographics

Sample composition (N = 26). Not a population estimate.

Dimension	Distribution (count)
Gender	Female 15, Male 10, Other 1
Age	25-34 9, 35-44 11, 45-54 3, 55-64 3
Race/ethnicity	White 18, Asian 3, Latino or Hispanic 2, Black or African American 2, Mixed or multiracial 1
Geography	Suburban 14, Residential (small town) 5, Urban 4, Rural 3
Upbringing SES	Less well-off 10, About the same 7, Better off 5, Much less well-off 2, Much better off 2
Caregiving consistency	Very consistent 13, Mostly consistent 11, Changed a few times 1, Changed frequently 1