

Recipients of Caregiving

The capability question, answered at 26 interviews

The Observatory × Dialogue AI · June 2026

Is this the flagship capability story? Yes, on depth and emergence. Not yet on scale.

The call: Lead with this study as proof that AI-moderated interviews deliver real depth and surface findings a survey never could, and that the method is honest enough to disagree with its own brief. Make the scale claim the next study, not this one.

Why now: The study set a clear bar: a skeptical buyer should say "we could not have gotten this from a survey," "the patterns are real, not cherry-picked." We cleared the first two. The interviews produced a genuine mechanism (people update the belief about an inherited rule fast and cannot change the reflex) and, more persuasively, the method surfaced a finding that cut against our own framing and a moment where the AI moderator over-stepped and the participant pushed back on the record. That self-correction is the most sellable thing here. It is exactly what a skeptic of AI research wants to see.

What we're doing: Position depth, emergence, and honesty as the proven capability, anchored on two artifacts: the emergent disconfirmation (a real group kept their inherited rules and rejected the "healing" frame we expected) and the documented moderator over-steer that the quality score caught rather than hid.

What we're explicitly not doing: Claiming scale. This is 26 interviews, not the 300 the plan targeted, and all 26 are parents, so the planned demographic cells are unfilled. We do not present any finding as a population percentage, and we do not say the scale hypothesis is validated. Saying so would be the one move that hands a skeptic the win.

Tradeoffs we're accepting: A smaller, honest claim now over a bigger, unprovable one. We trade "AI interviews at scale, proven" (which we cannot defend yet) for "AI interviews with human-grade depth and built-in self-correction, proven at 26, scaling next" (which we can). The narrower claim is the more credible one to the exact audience that matters.

Next: Research to run the same instrument at the planned 300, across the full demographic design, to test the scale half of the claim (owner: research team; target: next cycle). Two fixes ride along: tune the moderator to stop imposing contradictions (the P22 case), and scrub the one record that arrived with a raw system ID instead of a name (P15) before anything leaves the building.

Evidence: 26 AI-moderated interviews (25 usable, one dropoff), IQS 3.96/5. Confidence flag: small N, depth and emergence demonstrated; scale and population prevalence explicitly not claimed.