

Can a large language model judge a person’s stage of conscious human development from ordinary conversation?

Carole Griggs^a, Damon Abraham^b, Julia Mossbridge^c, Raiya Kind^d, Ted Strauss^a, Arnaud Delorme^{e,f,*}

^a*iConscious, Portola Valley, CA, USA*

^b*Independent researcher, Denver, CO, USA*

^c*Mossbridge Institute, Sebastopol, CA, USA*

^d*Independent researcher, San Francisco, CA, USA*

^e*University of California San Diego, La Jolla, CA, USA*

^f*Institute of Noetic Sciences, Novato, CA, USA*

Abstract

This pilot study asked whether a large language model (LLM) can judge a person’s stage of spiritual development from a short, lightly guided chat. An AI assessor named Kairos, built on a Gemini family model, rated chat transcripts using the 14 stage iConscious developmental model. We compared its ratings with the consensus of trained human experts. The experts rated each transcript independently and could not see each other’s ratings. On the held out test set of 40 participants, which the AI never saw during calibration, agreement between the AI and the human panel reached a weighted kappa of 0.828. That is more than twice our pre-registered minimum of 0.40. Across the full group of 50 participants, agreement stayed strong at 0.765. The AI’s

*Corresponding author.

Email address: arnodelorme@gmail.com (Arnaud Delorme)

rating fell within one stage of the human consensus 90% of the time in both groups. The AI’s confidence scores also carried real information. When the AI reported higher confidence, its ratings were more accurate. Its confidence also tracked the human panel’s confidence on hard cases, giving a useful signal about when its own ratings could be trusted. To our knowledge, this is the first pre-registered validation of LLM based developmental stage assessment against blinded expert consensus. These results are early evidence that LLMs can pick up a real developmental signal from everyday conversation. This task has usually required hours of structured interviews or expert scored written tests. The study also has clear limits. Several stages had too few participants for stable estimates. Agreement depended on the quality of the conversation. Findings on personality, self-transcendence, and sentiment scores were exploratory and should be treated as ideas for future testing, not conclusions.

Keywords: large language model, developmental stage assessment, spiritual development, ego development, AI and human agreement, natural language processing

1. Introduction

Adults can keep growing psychologically long after childhood. By adult psychological development we mean the step by step reorganization of how a person makes meaning. This shows up in how they take perspectives, manage their emotions, reason about ethics, and handle complexity. This capacity has been studied formally since Loevinger’s foundational work on ego development [1]. Later frameworks extended this line of work. Cook-

8 Greuter extended it with the Maturity Assessment Profile, describing nine
9 levels of growing ego maturity [2, 3]. Kegan described how modern life de-
10 mands more complex minds [4]. Torbert and colleagues built an action logics
11 framework for leadership [5]. Commons and colleagues built a model of hier-
12 archical complexity [6]. More recently, contemplative and integral perspec-
13 tives have extended this inquiry into experiential and transpersonal territory
14 [7, 8, 9, 10].

15 Across this literature, higher stages of adult development go along with
16 more prosocial behavior, greater emotional stability, and more openness to
17 experience. These capacities predict individual and group performance in or-
18 ganizations [5, 3, 4, 11, 12]. Because of this, there is strong applied interest in
19 developmental assessment for leadership programs, executive coaching, orga-
20 nizational diagnostics, and education. Yet the measurement tools themselves
21 have held the field back.

22 The main validated tools are the Sentence Completion Test (SCT) [1, 13]
23 and structured developmental interviews. Both need expert training to ad-
24 minister. Both take substantial time per person. Both need costly manual
25 scoring by certified raters. This makes large scale, repeated, or embedded
26 assessment impractical in most organizations and schools. In short, the de-
27 velopmental signal is real and well described in theory, but the tools for
28 measuring it are too slow and costly for wide use.

29 Recent advances in large language models (LLMs) open a new possibil-
30 ity. Theory suggests that developmental stage shows up in the structure
31 and quality of everyday language. It appears in the perspectives a person
32 takes, the complexity they tolerate, the distance they keep from their own

self talk, and the reasoning strategies they use. If that is true, then LLMs trained on huge amounts of human language may be able to detect that signal in ordinary, unstructured text. Developmental psychology has long held that language does not just report internal states. Language also shapes and reflects those states [4, 14, 15]. A capable language model, given the right developmental scaffolding, might therefore read the underlying developmental structure directly from conversation.

This idea had not been tested systematically against validated ground truth standards. The present study is a first test of it. We used a pre-registered design with a held out test set, an independent panel of trained human expert raters as ground truth, and a primary outcome measure chosen in advance.

The study uses the iConscious developmental stage model [16] as its framework. The iConscious model is a scale of conscious, or spiritual, development. It builds directly on Cook-Greuter’s Maturity Assessment Profile and the Washington University Sentence Completion Test tradition [1, 13, 2, 3], on O’Fallon’s STAGES model [9], and on Wilber’s integral framework [8]. It also draws on hierarchical complexity models and contemplative traditions. It defines a sequence of stages marked by growing capacity for multi-perspective awareness, self-reflective regulation, post-conventional ethical reasoning, and stable engagement with paradox and uncertainty.

Higher iConscious stages are theorized to match the general capacities described across the literature. These include holding several conflicting perspectives at once, observing and regulating one’s own identity, beliefs, and emotional reactions, acting from internalized principles rather than so-

58 cial convention or approval, and making decisions grounded in systems level
59 awareness and long term consequences. These capacities are theorized to
60 produce the prosociality, stability, and openness that the broader literature
61 links to later stage functioning [3, 4, 5].

62 The central question is simple to state. Can a large language model, work-
63 ing from ordinary chatbot transcripts and calibrated to a validated spiritual
64 development framework, produce stage ratings that agree with trained hu-
65 man experts well enough to count as a meaningful and reliable assessment
66 tool? The primary objective of this study was to answer that question, testing
67 whether an LLM based AI assessor (Kairos, referred to as ConsciousAI in the
68 pre-registration) calibrated to the iConscious model, could agree with expert
69 human judges on developmental stage ratings taken from natural conversa-
70 tion transcripts, at a level that is both statistically significant and practically
71 meaningful. A secondary objective was to test whether the AI’s self-reported
72 confidence scores were well calibrated. Four exploratory research questions
73 examined links between assessed stage and self-report measures of personal-
74 ity, self-transcendence, unconditional love, and written sentiment.

75 **2. Materials and methods**

76 *2.1. Ethics statement*

77 The study was approved by the Institutional Review Board of the In-
78 stitute of Noetic Sciences under protocol DELA_2026_01. All participants
79 gave informed consent. Participants were compensated 25 US dollars. Data
80 were collected between April 1 and July 1, 2026.

81 2.2. Study design

82 The study had four phases. First, we developed and calibrated the AI
83 assessment model on a held out validation subset. Second, we collected
84 natural language transcripts from participants who chatted freely with a
85 chatbot. Third, the AI assessor and a panel of trained human expert raters
86 assessed each transcript independently and blind to each other. Fourth,
87 we ran the pre-registered statistical analysis of agreement, calibration, and
88 exploratory secondary outcomes.

89 2.3. Terminology and stage descriptions

90 The iConscious model is a scale of conscious, or spiritual, development.
91 Its lower stages track the psychological maturation described in the adult
92 development literature. Its higher stages extend into spiritual and transper-
93 sonal territory. In this report, the terms consciousness, conscious develop-
94 ment, and conscious developmental stage refer to a person’s position on this
95 scale, as operationalized by the iConscious stage model. We do not use these
96 terms in the neuroscientific or philosophical sense of subjective or phenom-
97 enal consciousness, meaning the felt, first person quality of experience. We
98 make no claim about subjective awareness, sentience, or the so called hard
99 problem of consciousness [17]. The construct under study is developmental
100 maturity as assessed from conversational language, spanning psychological
101 and spiritual growth.

102 S2 Table (supplementary material) lists the stages that appear in this
103 study, with a short description of each, for readers who are new to the iCon-
104 scious framework. Stages are numbered on the iConscious scale, which spans
105 stages 1 to 14. Only stages 3 to 11 were observed here.

106 2.4. *Participants and data collection*

107 Participants had unstructured, naturally paced conversations with a stan-
108 dard chatbot (gemini-3.5-flash-high model). Sessions lasted 30 minutes. Par-
109 ticipants were encouraged to talk naturally about topics such as their rela-
110 tionship with themselves, with others, and with their life direction. The
111 study instructions stressed that there are no right or wrong answers and that
112 we were simply interested in how they engage. No specific topics were re-
113 quired. No participant was told their assessed or expected developmental
114 level during the study. This was done to keep self-presentation bias from
115 shaping the conversations.

116 Participants saw this message when they entered the session: “Welcome
117 to this research session. The researchers appreciate you taking your time to
118 participate for the next 30 minutes. We’ll talk about your relationship with
119 yourself, others, and life.”

120 The chatbot ran with this instruction: “You are a lightly modified version
121 of Gemini. Your job is to interact with people while keeping them confined
122 to personal topics (e.g. around the self, relationships, and life direction).
123 Respond with 1-2 sentences (sparingly 3), and vary your responses.”

124 Transcripts were generated automatically as text only, with no audio or
125 video. After the conversation, participants shared their transcript through
126 a survey. The survey also collected age and gender and a set of self-report
127 instruments used in the exploratory analyses. These were the Ten-Item Per-
128 sonality Inventory (TIPI) [18], the 15 item Self-Transcendence Scale (STS)
129 [19], and the Unconditional Love Scale (ULS) [20, 21, 22]. Gender was col-
130 lected using an open text field and age using a field with the example “e.g.

25.” The TIPI used seven response categories, from 1 (Disagree Strongly) to 7 (Agree Strongly). The STS retained the 15 original items but used modified response labels: 1 (Not Very), 2 (Some), 3 (Somewhat), and 4 (Very). The four ULS items asked about unconditional love toward oneself, other humans, animals, and the computer used to complete the survey. Its response labels were 1 (never), 2 (a few times), 3 (sometimes), 4 (often), and 5 (a great deal). The survey instructed participants to leave an answer blank to indicate N/A. The full administered questionnaire, including all instructions and response options, is provided as supplementary material.

Short written check-ins were given right before and right after the session. At both time points, participants were asked: “Please describe in one sentence how you feel right now?” We converted the text of each check-in into a sentiment score with NLTK VADER [23], a compound score that runs from -1 (very negative wording) to $+1$ (very positive wording), to index the change from before to after.

The original research proposal set a recruitment target of 30 participants. Participants were recruited by email through subscribers of the Institute of Noetic Sciences (IONS), iConscious subscribers, students at the University of Denver, and members of The Portal, a wellness club in Mill Valley, California. Because this was a pilot study and more data could be collected, the final cohort was 50 participants. Participants ranged in age from 19 to 86 years (mean 51.9, SD 18.4). Thirty-two identified as female, 16 as male, and 2 as nonbinary. This cohort was split into a 10 participant validation set used for model calibration and a 40 participant held out test set used only for confirmatory out-of-sample analysis. All primary confirmatory results are

156 reported for the strict test set ($N = 40$). The full cohort ($N = 50$) is reported
157 as a secondary check.

158 *2.5. Human expert assessment (ground truth)*

159 Each transcript was rated independently by a blinded panel of two trained
160 expert judges (co-authors CG and TS). Both are certified in the iConscious
161 developmental model. During their initial ratings they had no access to
162 the AI assessor’s output or to each other’s ratings. Each judge spent 1
163 to 2 hours per transcript. When the two judges disagreed, they resolved the
164 difference through a structured calibration discussion. This produced a single
165 Human Consensus Ground Truth (GT) rating per transcript. The panel also
166 produced a confidence score for each transcript, which was used in the H3
167 calibration analysis. Because the judges consulted each other whenever they
168 disagreed, their independent initial ratings were not saved.

169 *2.6. AI assessor development and calibration*

170 The Kairos assessment engine was built on a large language model in the
171 Gemini family, accessed through Google’s public API. The engine consists of
172 prompt scaffolding around the publicly available model. The final validated
173 configuration (named Variant F Median in the internal analysis plan) used
174 the gemini-3.5-flash-high model. It aggregated ratings across 21 independent
175 replicates per transcript and took the median as the final stage assignment.
176 This replicate and median procedure was designed to reduce the random
177 variability of any single model call and to make the ratings more stable. The
178 spread across the 21 replicates was not retained in the analysis dataset.

179 For each replicate, the model also reported a confidence percentage as
180 part of its structured output. The final confidence score for a transcript was
181 the aggregate of these self-reported values across replicates.

182 During Phase 1, the research team operationalized the iConscious stage
183 model by defining linguistic, semantic, emotional regulation, and perspective
184 taking markers theorized to match each developmental level. These markers
185 were used to build the assessment prompt that scaffolds the AI’s evaluation
186 process. The full text of the assessment prompt is proprietary and is available
187 from iConscious under a research agreement. S1 Table describes its structure,
188 including the marker categories, the output format, and how confidence was
189 elicited. Prompt calibration used only the 10 participant validation subset.
190 After calibration the model configuration was locked and no further changes
191 were allowed. All assessments of the 40 participant test set were made by
192 this locked configuration of Kairos.

193 During the assessment phase, Kairos independently assigned a develop-
194 mental stage and a self-reported confidence score to each transcript. It used
195 only the conversational text as input. It had no access to questionnaire data,
196 demographic information, or human panel ratings. The AI assessor and the
197 human expert panel were kept operationally isolated throughout.

198 *2.7. Pre-registration*

199 To rule out after the fact reinterpretation of results, the confirmatory hy-
200 potheses, success thresholds, primary outcome measure, and analytic meth-
201 ods were written into a Statistical Analysis Plan (SAP) and locked before
202 data analysis. The plan was registered on the Open Science Framework and
203 is publicly available (<https://osf.io/eu6sk/>). The split between confirma-

204 tory and exploratory analyses in this report follows that plan. Confirmatory
205 results address the hypotheses set in advance. Exploratory analyses were run
206 transparently alongside them, but they carry a higher risk of false positives
207 because they involve multiple uncorrected tests in a modest sample.

208 The SAP names the Big Five Inventory-10 (BFI-10) [24] as the personal-
209 ity instrument. The questionnaire actually administered contained the TIPI,
210 a different ten item measure of the Big Five with seven response categories.
211 We report this discrepancy from the registered instrument specification for
212 exploratory RQ1. The confirmatory hypotheses H1 to H3 do not use ques-
213 tionnaire scores.

214 2.8. Confirmatory hypotheses

215 2.8.1. H1: AI and human agreement (primary outcome)

216 The Kairos assessor will show statistically significant agreement with ex-
217 pert human judges on developmental stage ratings from natural language
218 transcripts, measured by quadratic weighted kappa (κ_w). The pre-registered
219 success criterion requires both the point estimate and the lower bound of the
220 95% confidence interval to exceed $\kappa_w \geq 0.40$. This matches the substantial
221 agreement threshold on Fleiss’s benchmark scale [25].

222 2.8.2. H2: Confidence and accuracy calibration

223 The AI assessor’s self-reported confidence scores will show a statistically
224 significant negative Spearman rank correlation with the AI’s own absolute
225 rating error relative to human consensus ground truth. This would confirm
226 that higher confidence goes with lower error. The pre-registered success
227 criterion is a one-tailed $p < 0.10$.

228 *2.8.3. H3: Confidence convergence*

229 The AI assessor’s self-reported confidence will show a statistically signifi-
230 cant positive Spearman rank correlation with the human expert panel’s own
231 confidence ratings. This would confirm that the AI’s uncertainty tracks a
232 shared sense of how hard each case is. The pre-registered success criterion is
233 a one-tailed $p < 0.10$.

234 For these two secondary calibration hypotheses, the pre-registered thresh-
235 old of one-tailed $p < 0.10$ is more lenient than the conventional 0.05. This
236 bar was set in advance because H2 and H3 are directional secondary checks in
237 a pilot study, intended to guide the design of a larger replication rather than
238 to establish the calibration effects on their own. The primary hypothesis
239 (H1) uses a much stricter criterion that does not depend on this threshold.

240 *2.9. Exploratory research questions*

241 The following four research questions were pre-registered as exploratory.
242 No corrections were applied to the reported p values, and results are read
243 accordingly.

- 244 1. RQ1: Do Big Five personality subscale scores show the theoretically
245 predicted correlations with human consensus developmental stage?
246 The administered measure was the TIPI, whereas the SAP specified
247 the BFI-10, as noted above.
- 248 2. RQ2: Does the 15 item Self-Transcendence Scale (STS) correlate posi-
249 tively with assessed developmental stage, in line with the theorized link
250 between later stage development and self-transcendent experience?

- 251 3. RQ3: Does the Unconditional Love Scale (ULS), including the item
252 about unconditional love for a computer or device, correlate with as-
253 sessed developmental stage?
- 254 4. RQ4: Does the sentiment score change from before to after the conver-
255 sation, and does the baseline score relate to developmental stage?

256 *2.10. Statistical analysis*

257 The primary outcome was the agreement between Kairos’s stage assign-
258 ments and the Human Consensus Ground Truth. We quantified it with a
259 quadratic weighted kappa (κ_w). Because this compares two sets of ratings,
260 the AI’s stage assignments and the panel’s single consensus rating per tran-
261 script, it is Cohen’s weighted kappa. Standard errors and confidence intervals
262 used the exact asymptotic method of Conger [26], which generalizes kappa
263 across raters. The quadratic weighting penalizes disagreements in propor-
264 tion to their size on the ordinal stage scale. Rating a stage 8 participant as
265 stage 4 is penalized much more than rating a stage 5 participant as stage 4.
266 This makes weighted kappa a better fit than unweighted kappa for ordered
267 categories. The pre-registered success criterion required both the κ_w point
268 estimate and the lower bound of its 95% confidence interval to exceed 0.40.

269 Secondary agreement metrics included percent exact agreement, percent
270 adjacent agreement (within one stage), mean absolute error (MAE), and
271 signed mean error (bias). Confidence calibration was tested with Spearman
272 rank correlations. H2 tested the negative link between AI confidence and AI
273 absolute error. H3 tested the positive link between AI confidence and human
274 panel confidence.

275 Stage specific sensitivity and specificity (one versus rest) were computed

and reported only for stages with at least 5 participants in the relevant cohort,
per the pre-registered plan. Estimates based on 5 to 8 participants are flagged
as highly unstable and descriptive only. All confirmatory analyses were run
on both the strict held out test set ($N = 40$) and the whole cohort ($N = 50$).
The test set is the primary confirmatory result.

Exploratory analyses (RQ1 to RQ4) were run without correction for multiple comparisons, in line with their pre-registered exploratory status. They are read as hypothesis generating, not confirmatory. One participant did not complete the questionnaire, so all exploratory analyses use $N = 39$ in the test set and $N = 49$ in the whole cohort.

3. Results

3.1. Confirmatory hypotheses

All three pre-registered confirmatory hypotheses were supported in both the strict out-of-sample test set ($N = 40$) and the whole cohort ($N = 50$). Table 1 summarizes the confirmatory outcomes.

Table 1: **Summary of pre-registered confirmatory hypothesis tests.** All three hypotheses were supported in both cohorts. All p values for H2 and H3 are one-tailed.

Hypothesis	Test set ($N = 40$)	Whole cohort ($N = 50$)	Threshold
H1 Agreement (κ_w)	$\kappa_w = 0.828$ 95% CI [0.670, 0.985] $p < .001$	$\kappa_w = 0.765$ 95% CI [0.562, 0.969] $p < .001$	$\kappa_w \geq 0.40$ (lower bound ≥ 0.40)
H2 Calibration: accuracy	$\rho = -0.248$ $p = .062$	$\rho = -0.241$ $p = .046$	$\rho < 0$ $p < .10$
H3 Calibration: convergence	$\rho = 0.231$ $p = .076$	$\rho = 0.343$ $p = .007$	$\rho > 0$ $p < .10$

291 *3.1.1. H1: AI and human agreement*

292 The primary confirmatory analysis measured agreement between Kairos’s
293 stage assignments and the Human Consensus Ground Truth using quadratic
294 weighted kappa with Conger’s exact asymptotic standard error [26]. On the
295 held out test set, agreement reached $\kappa_w = 0.828$ (SE = 0.078, 95% CI [0.670,
296 0.985], $z = 10.65$, $p < .001$). On the whole cohort, agreement stayed strong
297 at $\kappa_w = 0.765$ (SE = 0.101, 95% CI [0.562, 0.969], $z = 7.57$, $p < .001$). Both
298 point estimates were far above the pre-registered minimum of 0.40. Critically,
299 the lower bounds of both 95% confidence intervals also exceeded 0.40. This
300 satisfies the pre-registered criterion in its most conservative form.

301 On the Landis and Koch interpretive scale [27], a κ_w of 0.828 falls in
302 the almost perfect range (0.81 to 1.00), well above the substantial agreement
303 band the study was designed to reach. Two notes help put this number
304 in context. First, published agreement between trained human scorers on
305 instruments such as the SCT typically falls between about 0.7 and 0.9 [13, 2].
306 The agreement observed here sits inside that band. In other words, the AI
307 agreed with the expert consensus about as well as trained experts typically
308 agree with each other. Second, the Landis and Koch bands were derived for
309 unweighted kappa. Applying them to a quadratic weighted kappa is common
310 but informal, so the almost perfect label should be read as a descriptive gloss
311 rather than a validated threshold.

312 Secondary agreement metrics are shown in Table 2. The AI and the hu-
313 man panel assigned exactly the same stage in about 53% of cases. Adjacent
314 agreement, where the AI’s rating fell within one stage of the human consen-
315 sus, held at 90.0% in both cohorts. On an ordered scale like this one, being

off by one stage is a qualitatively different error than being off by many stages. The quadratic weighted kappa reflects this by weighting errors in proportion to their size. The signed mean error shows a small positive bias in both cohorts. The AI assigned higher stages than the humans by 0.16 to 0.35 levels on average. This bias is small relative to the scale range and is weaker in the whole cohort estimate.

Table 2: **Secondary agreement metrics for Kairos versus human consensus ground truth.**

Metric	Test set (N = 40)	Whole cohort (N = 50)
Percent exact agreement	52.5%	54.0%
Percent adjacent agreement (≤ 1 stage)	90.0%	90.0%
Mean absolute error (MAE)	0.650 levels	0.680 levels
Signed mean error (bias)	+0.350 levels	+0.160 levels

3.1.2. H2: Confidence and accuracy calibration

To test whether the AI’s self-reported confidence scores carried real information about rating reliability, we computed the Spearman rank correlation between AI confidence and absolute rating error (the absolute difference between the AI stage and the ground truth stage). On the test set, $\rho = -0.248$ (one-tailed $p = .062$). On the whole cohort, $\rho = -0.241$ (one-tailed $p = .046$). Both correlations were negative and significant at the pre-registered threshold of $p < .10$ (one-tailed). Higher AI confidence went with lower rating error. This supports using AI confidence as a quality control signal. Transcripts where the AI reports low confidence should be expected to yield less reliable stage assignments than transcripts where it reports high confidence.

333 3.1.3. H3: Confidence convergence

334 To test whether the AI’s confidence tracked the same sense of difficulty
335 that the human panel felt, we computed the Spearman rank correlation be-
336 tween AI confidence and human panel confidence. On the test set, $\rho = 0.231$
337 (one-tailed $p = .076$). On the whole cohort, $\rho = 0.343$ (one-tailed $p = .007$).
338 Both correlations were positive and significant at the pre-registered thresh-
339 old. The clearly stronger whole cohort result suggests the AI’s uncertainty
340 is not just self-referential. It reflects real properties of the transcripts that
341 trained human judges also find hard to rate. This convergence is evidence
342 that the AI and the human panel respond to a partly shared signal in the
343 transcripts. For transparency, the whole cohort results for both calibration
344 hypotheses also clear the conventional 0.05 level (H2 $p = .046$, H3 $p = .007$),
345 while the test set values ($p = .062$ and $p = .076$) pass only the pre-registered
346 0.10 bar.

347 As shown in Table 3, agreement rates split by confidence tertile confirm
348 the calibration pattern. Exact match accuracy roughly doubles from the
349 low confidence group (30.8% to 35.3%) to the medium and high confidence
350 groups (61.5% to 64.7%). The medium confidence group reached 100% ad-
351 jacent agreement in both cohorts. Even when the AI’s confidence was only
352 moderate, large misclassifications were rare.

353 3.2. Stage specific sensitivity and specificity

354 Per the pre-registered plan, stage specific sensitivity and specificity (one
355 versus rest) are reported only for stages with at least 5 participants in the
356 relevant cohort. Estimates based on 5 to 8 participants are flagged as highly
357 unstable and treated as descriptive only. Results appear in Table 4.

Table 3: **Exact and adjacent agreement rates by AI confidence tertile (test set and whole cohort).**

Confidence group	Test N	Exact	Adjacent	Whole N	Exact	Adjacent
Low (58.7% to 79.7%)	13	30.8%	84.6%	17	35.3%	82.4%
Medium (80.0% to 85.6%)	14	64.3%	100.0%	16	62.5%	100.0%
High (86.0% to 95.5%)	13	61.5%	84.6%	17	64.7%	88.2%

Table 4: **Stage specific one versus rest sensitivity and specificity.** *n = 5 in the whole cohort. Estimates for Stage 3 are highly unstable and descriptive only. Stages with fewer than 5 participants are not reportable (n.r.).

iConscious stage	Test N	Test sens.	Test spec.	Whole N	Whole sens.	Whole spec.
Stage 3	3	n.r.	n.r.	5	0.400*	0.933*
Stage 4	18	0.611	0.818	21	0.667	0.759
Stage 5	14	0.500	0.846	17	0.471	0.848
Stage 6	2	n.r.	n.r.	3	n.r.	n.r.
Stage 9	1	n.r.	n.r.	1	n.r.	n.r.
Stage 10	1	n.r.	n.r.	2	n.r.	n.r.
Stage 11	1	n.r.	n.r.	1	n.r.	n.r.

As shown in Table 4, stages 4 and 5 were the only stages with enough participants (14 to 21) to support even partly stable estimates, and even these should be read with the overall sample size in mind. Within these limits, Kairos identified Stage 4 participants with a sensitivity of 0.611 to 0.667 and Stage 5 participants with a sensitivity of 0.471 to 0.500. Specificity was consistently higher, from 0.759 to 0.848. The model's errors leaned toward false negatives (missing the correct stage) rather than false positives (assigning a stage to people who were not at it). The Stage 3 whole cohort estimate (n = 5) sits exactly at the minimum reporting threshold and should be treated with matching caution. Stages 6, 9, 10, and 11 each had one to

three participants and cannot yield meaningful estimates under any standard psychometric criterion.

Figure 1 shows the full confusion matrices for the test set and the whole cohort as heat maps. Rows are the Human Consensus Ground Truth stage. Columns are the AI assigned stage. Diagonal cells are exact matches. Off diagonal cells are misclassifications. Stage 7 is included on both axes to preserve the ordinal integrity of the scale. No participant was rated Stage 7 by either the human panel or the AI in this sample. Its absence from the sample does not mean it does not exist in the iConscious model. Stage 8 likewise had no ground truth participants. One AI assigned Stage 8 rating appears for a participant whose ground truth was Stage 5.

Inspection of both panels shows that most misclassifications were adjacent. When the human and AI ratings disagreed, they most often differed by one stage in either direction rather than by a large jump. This pattern matches the 90% adjacent agreement rates reported above and the strong quadratic weighted kappa, which discounts small errors while heavily penalizing large ones.

3.3. AI citation profile analysis

The AI assessor’s outputs included citations to the iConscious stage or stages it invoked while forming each rating. The mean number of stages cited per transcript was 2.42 in the test set (whole cohort 2.44). The mean citation range (maximum minus minimum cited stage) was 1.75 levels in the test set (whole cohort 1.80 levels). The citation range did not correlate significantly with the ground truth stage ($\rho = 0.141$, $p = .385$ in the test set and $\rho = 0.241$, $p = .091$ in the whole cohort) or with ground truth panel confidence

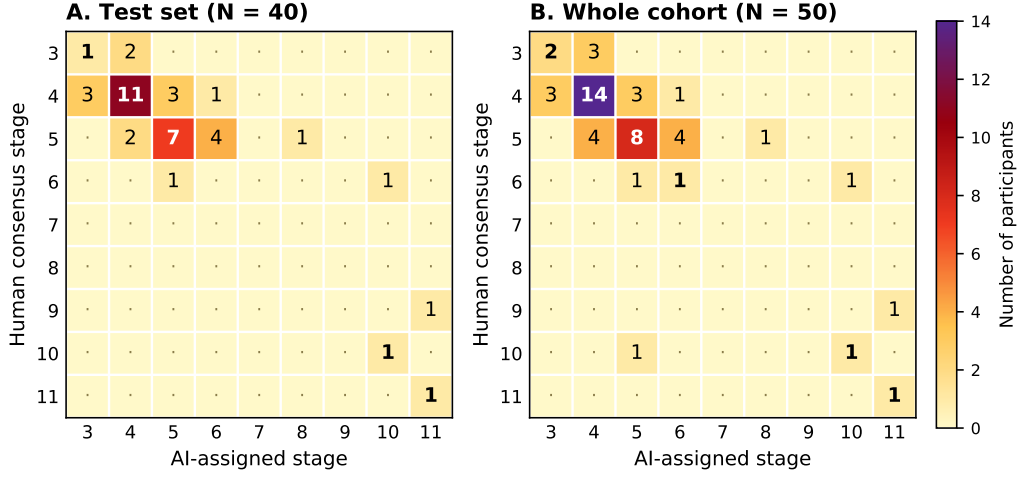


Figure 1: **Confusion matrix heat maps for the test set and the whole cohort.** A: Test set (N = 40). B: Whole cohort (N = 50). Rows show the human consensus ground truth stage. Columns show the AI assigned stage. Each cell shows the number of participants with that combination. Bold numbers on the diagonal are exact matches. Color runs from pale yellow (few participants) through orange and red to deep purple (many participants). Cells with a centered dot had no observed participants, including all of Stage 7, which was not represented in this sample.

393 ($\rho = -0.036$, $p = .827$ in the test set and $\rho = 0.016$, $p = .914$ in the whole
394 cohort). This null result means that citation breadth, the range of stages the
395 AI referenced while reasoning, does not act as a proxy for participant stage
396 or for human perceived rating difficulty. It is therefore not a useful secondary
397 signal for quality filtering.

398 3.4. Transcript helpfulness as a moderator of agreement

399 A supplementary analysis examined whether transcript quality moder-
400 ated AI and human agreement. The helpfulness score was produced by the
401 AI assessor itself. For each transcript, the model rated how much useful

developmental evidence the conversation contained, on a 0 to 100% scale. Transcripts were split at the median helpfulness score, which was 90.0% in both cohorts. Because many transcripts sat exactly at the median value, the split is 29 versus 11 in the test set rather than 20 versus 20. Transcripts at the median were grouped with the high helpfulness group. Weighted kappa and Spearman correlations were computed separately for the high helpfulness (HH) and low helpfulness (LH) subgroups. Results appear in Table 5.

Table 5: **AI and human agreement by transcript helpfulness (median split at 90.0%).** HH means high helpfulness. LH means low helpfulness.

Metric	Test HH (n = 29)	Test LH (n = 11)	Whole HH (n = 36)	Whole LH (n = 14)
Mean absolute error	0.793 levels	0.273 levels	0.833 levels	0.286 levels
Weighted κ_w	0.827	0.593	0.754	0.615
95% CI	[0.657, 0.996]	[0.040, 1.000]	[0.526, 0.983]	[0.153, 1.000]
Spearman ρ (p)	0.778 (p < .001)	0.737 (p = .010)	0.751 (p < .001)	0.713 (p = .004)

Weighted kappa was clearly higher in the HH subgroup ($\kappa_w = 0.827$ test, 0.754 whole) than in the LH subgroup ($\kappa_w = 0.593$ test, 0.615 whole). The apparently lower MAE in the LH subgroup is a statistical artifact of range restriction. LH transcripts were concentrated among lower stage participants (Stages 3 to 5), where the numeric range of possible errors is smaller. HH transcripts spanned the full range (Stages 3 to 11), including rare higher stage participants who inflate MAE without weakening the kappa estimate. The higher and more robust κ_w in the HH group confirms this reading. Transcript helpfulness is therefore a meaningful moderator of agreement, with direct implications for how future AI based developmental assessment tools should be designed.

420 3.5. *Exploratory research questions*

421 3.5.1. *RQ1: Big Five personality subscales versus developmental stage*

422 Correlations between TIPI subscales and human consensus stage appear
 423 in Table 6. No Big Five subscale reached statistical significance in either
 424 cohort. Openness to experience, the trait closest in theory to developmental
 425 complexity, showed a trend in the unexpected negative direction in the test
 426 set ($\rho = -0.264$, $p = .105$) and was essentially null in the whole cohort.
 427 These data do not support a simple direct link between self-reported Big
 428 Five traits and assessed developmental stage in this sample. The modest
 429 sample limits the power to detect small effects, so this null result should not
 430 be read strongly.

Table 6: **Spearman correlations between TIPI subscales and human consensus developmental stage.** Exploratory $N = 39$ (test) and 49 (whole cohort) because one participant had missing questionnaire data (see Statistical analysis).

TIPI trait	Test ($N = 39$) ρ (p)	Whole ($N = 49$) ρ (p)	Expected direction
Extraversion	0.013 (.938)	-0.015 (.917)	Exploratory
Agreeableness	0.028 (.867)	0.040 (.785)	Exploratory
Conscientiousness	-0.065 (.696)	-0.074 (.616)	Exploratory
Neuroticism	-0.166 (.314)	-0.163 (.262)	Negative
Openness to experience	-0.264 (.105)	-0.015 (.922)	Positive

431 3.5.2. *RQ2: Self-transcendence versus assessed developmental stage*

432 The Self-Transcendence Scale (STS) total score (range 15 to 60) corre-
 433 lated significantly with the human consensus stage in both cohorts. In the
 434 test set, $\rho = 0.382$, $p = .016$. In the whole cohort, $\rho = 0.407$, $p = .004$. Cor-
 435 relations with the AI's own ratings pointed the same way but were weaker.

436 In the test set, $\rho = 0.306$, $p = .058$. In the whole cohort, $\rho = 0.321$, $p = .025$.
437 These results fit theoretical accounts linking later stage development to an
438 orientation toward self-transcendence [7, 10, 9]. This is the most theoreti-
439 cally coherent and statistically consistent pattern in the exploratory data. It
440 is a natural candidate for promotion to a confirmatory hypothesis in a larger,
441 adequately powered study.

442 3.5.3. RQ3: Unconditional love versus developmental stage

443 The Unconditional Love Scale (ULS) composite showed a positive but
444 non-significant correlation with ground truth stage in the test set ($\rho = 0.215$,
445 $p = .188$) and reached significance in the whole cohort ($\rho = 0.316$, $p = .027$).
446 The ULS item that asks about unconditional love for a computer or device
447 showed weak, non-significant correlations in both cohorts (test $\rho = 0.081$,
448 $p = .625$, whole $\rho = 0.197$, $p = .175$). These data give no support to the
449 idea that assessed developmental stage predicts stronger affection toward AI
450 systems specifically.

451 3.5.4. RQ4: Change in sentiment score from before to after the session

452 Sentiment results appear in Table 7. The mean sentiment score after
453 the session was slightly higher than before in both cohorts, but individual
454 changes went both ways. About 41% to 43% of participants improved (change
455 greater than $+0.05$), 15% to 18% stayed neutral, and 39% to 44% decreased.
456 The sentiment score before the session did not correlate significantly with
457 developmental stage in either cohort. These results do not support claims
458 that brief chatbot interactions reliably improve emotional state. They also
459 do not support a link between baseline sentiment and developmental level in

460 this sample.

Table 7: **Change in sentiment score from before to after the session (NLTK VADER compound scores).** Scores run from -1 to $+1$. One participant had missing questionnaire data, so $N = 39$ (test) and 49 (whole cohort), as noted in the Statistical analysis section.

Sentiment metric	Test set (N = 39)	Whole cohort (N = 49)
Mean sentiment score before session	0.233	0.258
Mean sentiment score after session	0.287	0.338
Mean change	+0.053	+0.080
Improved (change $> +0.05$)	16 (41.0%)	21 (42.9%)
Neutral (within ± 0.05)	6 (15.4%)	9 (18.4%)
Decreased (change < -0.05)	17 (43.6%)	19 (38.8%)
Score before vs. GT stage (ρ , p)	0.217 (.184)	0.106 (.471)

461 4. Discussion

462 The primary confirmatory finding is that Kairos, an LLM based assessor
 463 calibrated to the iConscious developmental stage model, agreed substantially
 464 with a panel of trained human expert judges on stage ratings from natural,
 465 unstructured chatbot conversations. On the strictly held out test set ($N =$
 466 40), weighted kappa reached 0.828. Under standard interpretive conventions
 467 this is near perfect agreement [27], and it is more than twice the pre-registered
 468 minimum. The result held out-of-sample, on data the model never saw during
 469 calibration. That is the more stringent and more practically informative test.

470 The calibration findings reinforce the primary result, and they matter for
 471 deployment. H2 and H3 look like two modest correlation tests, but together

472 they show a property that any real world use of this technology would de-
473 pend on. The system knows when it does not know. Its confidence dropped
474 on exactly the transcripts where its error rose (H2), and it flagged the same
475 transcripts that human experts also found hard to rate (H3). This conver-
476 gence suggests that the AI and the human experts were responding to partly
477 shared information in the transcripts, not to independent signals that hap-
478 pened to line up. The AI’s uncertainty appears to reflect genuine transcript
479 level ambiguity of the kind trained raters also struggle with, rather than an
480 artifact of its scoring procedure.

481 *4.1. Interpretation in the context of the developmental assessment literature*

482 Developmental stage assessment has always depended on demanding in-
483 struments. The Sentence Completion Test [1, 3] requires structured written
484 responses and certified expert scoring. Developmental interviews require
485 trained administrators and long, skill intensive sessions. These demands
486 have kept developmental assessment inside research settings and high invest-
487 ment programs. They have ruled out use at the scale of whole organizations,
488 repeated tracking, or consumer applications.

489 The present results suggest that a meaningful developmental signal, one
490 that trained judges using a validated framework can detect and agree on, may
491 also be detectable by an LLM from ordinary conversational text produced
492 during a routine chatbot interaction. If this replicates, the measurement
493 burden drops sharply. The cost difference is easy to state from our own
494 methods. Human ground truth took 1 to 2 hours per transcript per judge.
495 With two judges and 50 transcripts, that is roughly 100 to 200 expert hours.
496 The AI produced its ratings in minutes of computer time. Embedded de-

497 velopmental sensing in coaching platforms, educational software, leadership
498 programs, and widely used conversational AI tools would become technically
499 feasible without dedicated interview sessions or expert scoring pipelines.

500 *4.2. What this study does not establish*

501 These claims must be scoped carefully. This study shows detection of a
502 developmental signal in existing conversational text. It does not show the
503 ability to cause developmental change. It makes no claim about what an AI
504 system should do with a developmental assessment once it has one. Whether
505 AI facilitated developmental interventions are safe, effective, or ethically ap-
506 propriate is a separate question that these data cannot answer.

507 *4.3. The helpfulness moderation effect*

508 The finding that transcript helpfulness strongly moderated agreement
509 ($\kappa_w = 0.827$ versus 0.593 in the test set) has direct design implications. The
510 reliability of LLM based developmental assessment is not independent of the
511 conversational input. Transcripts with richer, more substantive engagement
512 yield more reliable ratings. In one sense this is unsurprising. Short or shal-
513 low answers give any rater, human or AI, less to work with. But it matters
514 in practice. Any deployment should include minimum conversational depth
515 criteria, adaptive prompting that draws out substantive engagement, or con-
516 fidence based filtering before treating AI ratings as actionable.

517 *4.4. Self-transcendence and developmental stage*

518 Among the four exploratory questions, the link between the Self-
519 Transcendence Scale and assessed stage (RQ2) was the most coherent and

520 consistent result. It reached significance in both cohorts for the human
 521 consensus ratings ($\rho = 0.38$ to 0.41) and pointed the same way for the
 522 AI’s ratings. Multiple frameworks predict this link, connecting later stage
 523 development to expanded awareness, reduced ego centered processing,
 524 and orientation toward transcendent meaning [7, 10, 9]. On its own this
 525 result is not confirmatory. It was exploratory and uncorrected for multiple
 526 comparisons. But it is a suitable basis for a pre-registered confirmatory
 527 hypothesis in a properly powered follow up study.

528 *4.5. Null and mixed exploratory findings*

529 The absence of significant links between Big Five subscales and devel-
 530 opmental stage (RQ1, Table 6) is compatible with arguments in the devel-
 531 opmental literature that personality traits and developmental stage measure
 532 different things [3, 4]. Traits measured by instruments like the TIPI capture
 533 dispositions. Developmental stage captures the structural complexity of a
 534 person’s meaning making. However, the small sample and the limited mea-
 535 surement precision of a two item per trait inventory [18] mean that these
 536 null correlations do not establish that the constructs are independent.

537 The mixed results for the Unconditional Love Scale (RQ3) and the het-
 538 erogeneous sentiment findings (RQ4) give no basis for inferring systematic
 539 links between these measures and developmental stage in this sample. The
 540 null result for the device love item is worth noting. It constrains speculative
 541 claims about developmental stage and emotional attachment to AI systems.

542 4.6. Limitations

543 Several limitations bound these findings. First, the sample was small
544 relative to the stage space. Most participants fell in Stages 4 and 5, and
545 Stages 3, 6, 9, 10, and 11 had only one to five participants each, so per stage
546 estimates are unstable. Recruitment channels also skewed toward contem-
547 plative, development interested populations, and the sample skewed toward
548 older adults (mean age 52 years) and toward women. Results may not gen-
549 eralize to broader populations.

550 Second, the reference standard is not fully independent. Two of the
551 study’s authors served on the expert panel, including a co-creator of the
552 iConscious model, and both hold a declared professional interest in its val-
553 idation (see Declaration of competing interests). Blinding and operational
554 isolation during the assessment phase prevented copying, but the same people
555 wrote the Kairos assessment prompt and produced the ground truth ratings,
556 so both reflect one reading of the framework. If other certified raters would
557 apply the model differently, the AI and the panel would err in the same direc-
558 tion and agreement would look higher than it is. Kappa measures agreement
559 with the chosen criterion, not whether the criterion is right. The judges also
560 consulted each other whenever they disagreed, so their independent initial
561 ratings were not saved and inter-rater reliability before consensus cannot be
562 reported. Future studies should log each judge’s initial rating before any
563 discussion.

564 Third, several measurement caveats apply. The AI showed a small posi-
565 tive bias (signed mean error of +0.16 to +0.35 stages). Agreement was clearly
566 higher for high helpfulness transcripts ($\kappa_w = 0.83$ versus 0.59 on the test set),

567 so the headline figures are conditional on this sample’s transcript quality.
568 The whole cohort analyses include the 10 calibration transcripts, making the
569 held out test set the more conservative estimate. The exploratory analyses
570 were not corrected for multiple comparisons.

571 The administered personality instrument was the TIPI rather than the
572 BFI-10 named in the SAP, so RQ1 deviates from the registered instrument
573 specification. The modified STS response labels may also affect comparability
574 with studies using the original response format. These instrument differences
575 should be considered when interpreting the exploratory questionnaire results.

576 Finally, both the conversations and the assessments relied on a single
577 model family (Gemini, gemini-3.5-flash-high with 21 replicate median aggre-
578 gation). Whether the procedure generalizes to other platforms, text formats,
579 models, prompts, or aggregation methods is unknown.

580 *4.7. Future directions*

581 This pilot establishes a proof of concept for LLM based developmental
582 stage detection and points to concrete next steps.

583 The most pressing need is replication with a much larger sample, deliber-
584 ately stratified for adequate coverage across the full iConscious stage range.
585 At least 15 to 20 participants per stage would be needed for stable per stage
586 sensitivity and specificity estimates. Deliberate stratification during recruit-
587 ment, using the assessment bot to over-sample under-represented stages, is
588 the practical way to get there at a reasonable total sample size.

589 Later studies should test whether the calibrated assessor generalizes to
590 transcripts from other conversational AI platforms, to written formats such as
591 journaling or asynchronous messaging, and to transcribed in-person dialogue.

592 Cross-platform testing would show whether the observed agreement reflects
593 a general property of the developmental signal in language or something
594 specific to Gemini style chatbot sessions.

595 The exploratory link between self-transcendence and assessed stage
596 (RQ2) is well grounded in theory and consistent across cohorts and rating
597 sources. A follow up study should pre-register this link as a confirmatory
598 hypothesis, with a power analysis based on the effect size seen here ($\rho \approx$
599 0.32 to 0.41) at conventional power (0.80) and significance ($\alpha = .05$).

600 The moderation of agreement by transcript helpfulness should be tested
601 experimentally. Randomly assigning different conversational prompting
602 strategies would show whether the link is causal and how large the effect on
603 reliability is. That would directly inform design specifications for deployed
604 tools.

605 Future studies should also report formal inter-rater reliability among
606 panel members before consensus. This establishes the quality of the hu-
607 man ground truth itself and reveals any systematic differences among raters
608 that could shape the consensus score.

609 This study measured stage at a single time point. Longitudinal designs
610 tracking the same people over time would show whether the AI assessor is
611 sensitive to developmental change. That sensitivity is a necessary condition
612 for any future use in programs that aim to support growth. Whether AI
613 mediated conversation can itself be a developmental intervention is a distinct
614 and much more complex scientific and ethical question. It requires separate
615 investigation with appropriate institutional review, purpose specific informed
616 consent, and safety protocols.

617 4.8. Conclusion

618 This pilot study provides early evidence that a large language model,
619 evaluated against blinded expert consensus, can detect developmental stage
620 from ordinary conversation. Agreement reached $\kappa_w = 0.828$ on the held out
621 test set, and the AI’s confidence scores tracked both its own accuracy and
622 human judgments of case difficulty. These results are a proof of concept for
623 scalable developmental assessment without dedicated interviews or expert
624 scoring.

625 The results do not yet generalize across all stages or conversational
626 contexts. The sample was modest and concentrated in the developmental
627 midrange, transcript quality moderated agreement, and the secondary
628 findings were exploratory. A larger stratified replication and generalizability
629 testing are the next steps.

630 The contribution of this work is to show that the path is worth taking.
631 The developmental signal that trained experts detect through structured
632 assessment appears to be present, and accessible to AI, in everyday conver-
633 sation. That finding matters for developmental psychology, organizational
634 science, educational technology, and the design of AI systems meant to sup-
635 port the people who use them.

636 **Supplementary material**

637 Two supplementary tables accompany this article in a separate file. S1 Ta-
638 ble describes the structure of the Kairos assessment prompt, including marker
639 categories, output format, confidence elicitation, and aggregation. The full
640 prompt text is proprietary and available from iConscious under a research

641 agreement. S2 Table lists the iConscious developmental stages observed in
642 this study with a short description of each.

643 A second supplementary file contains the full questionnaire administered
644 to participants, including the written check-ins before and after the con-
645 versation, demographic questions, ULS items, STS items with the modified
646 response labels, and TIPI items. It reproduces the survey screens with their
647 original instructions, item order, wording, and response options. Browser
648 address bars containing session identifiers have been removed.

649 **CRedit authorship contribution statement**

650 Carole Griggs: Conceptualization, Investigation, Methodology, Project
651 administration, Writing (original draft). Conceived the study, co-led the
652 project, and oversaw study design, participant recruitment, data collection,
653 project management, and manuscript preparation.

654 Damon Abraham: Conceptualization, Data curation, Methodology, Su-
655 pervision, Writing (review and editing). Co-led the scientific design, devel-
656 oped the analysis strategy, supervised the scientific direction, and critically
657 revised the manuscript for important intellectual content.

658 Julia Mossbridge: Conceptualization, Methodology, Writing (review and
659 editing). Contributed to qualitative measurement and study design, inter-
660 preted the data, and critically revised the manuscript for important intellec-
661 tual content.

662 Raiya Kind: Conceptualization, Methodology, Project administration,
663 Writing (review and editing). Coordinated project operations, contributed
664 to study design and prompt engineering, and critically revised the manuscript

665 for important intellectual content.

666 Ted Strauss: Conceptualization, Data curation, Investigation, Writing
667 (review and editing). Contributed expertise in developmental psychology
668 and the iConscious framework, independently rated the conversational tran-
669 scripts as a blinded expert judge, participated in the discussions that estab-
670 lished consensus developmental stage ratings, interpreted the psychological
671 assessments, and critically revised the manuscript for important intellectual
672 content.

673 Arnaud Delorme: Methodology, Project administration, Writing (original
674 draft), Writing (review and editing). Managed the institutional review board
675 process, contributed to the study design, and participated in writing and
676 revising the manuscript.

677 Damon Abraham, Julia Mossbridge, Raiya Kind, and Ted Strauss ap-
678 proved the version submitted and agreed to be accountable for all aspects
679 of the work, including ensuring that questions related to the accuracy or
680 integrity of any part of the work are appropriately investigated and resolved.

681 **Funding**

682 This research was funded by iConscious with its own funds and with
683 anonymous donations made to iConscious. Two of the authors lead iCon-
684 scious, and their roles in the study are described in the CRediT statement
685 and the Declaration of competing interests.

686 **Declaration of competing interests**

687 Carole Griggs and Ted Strauss lead iConscious and are the creators of
688 the iConscious developmental model. They hold a professional interest in its
689 validation and report financial support from iConscious. This research was
690 funded by iConscious using its own funds and anonymous donations made
691 to iConscious. The other authors declare that they have no known compet-
692 ing financial interests or personal relationships that could have appeared to
693 influence the work reported in this paper.

694 **Data availability**

695 The pre-registered analysis plan is available on the Open Science Frame-
696 work (<https://osf.io/eu6sk/>). The transcript data and questionnaire re-
697 sults are available from the corresponding author on reasonable request, sub-
698 ject to participant privacy constraints. Analysis code is available from the
699 corresponding author on reasonable request. The full text of the AI assess-
700 ment prompt is proprietary and is available from iConscious under a research
701 agreement (see S1 Table in the supplementary material).

702 **References**

- 703 [1] J. Loevinger, Ego development: Conceptions and theories, Jossey-Bass,
704 San Francisco, 1976.
- 705 [2] S. R. Cook-Greuter, Postautonomous ego development: A study of its
706 nature and measurement, Ph.D. thesis, Harvard Graduate School of
707 Education (1999).

- 708 [3] S. R. Cook-Greuter, Making the case for a developmental perspective,
709 Industrial and Commercial Training 36 (7) (2004) 275–281.
- 710 [4] R. Kegan, In over our heads: The mental demands of modern life, Har-
711 vard University Press, Cambridge, MA, 1994.
- 712 [5] W. R. Torbert, S. Cook-Greuter, D. Fisher, E. Foldy, A. Gauthier,
713 J. Keeley, D. Rooke, S. Ross, C. Royce, J. Rudolph, S. Taylor, M. Tran,
714 Action inquiry: The secret of timely and transforming leadership,
715 Berrett-Koehler, San Francisco, 2004.
- 716 [6] M. L. Commons, E. J. Trudeau, S. A. Stein, F. A. Richards, S. R.
717 Krause, Hierarchical complexity of tasks shows the existence of devel-
718 opmental stages, Behavioral Development Bulletin 4 (1) (1998) 1–26.
- 719 [7] A. H. Maslow, Religions, values, and peak experiences, Ohio State Uni-
720 versity Press, Columbus, 1964.
- 721 [8] K. Wilber, Integral psychology: Consciousness, spirit, psychology, ther-
722 apy, Shambhala, Boston, 2000.
- 723 [9] T. O’Fallon, N. Polissar, M. B. Neradilek, T. Murray, The validation
724 of a new scoring method for assessing ego development based on three
725 dimensions of language, Heliyon 6 (3) (2020) e03472.
- 726 [10] D. B. Yaden, J. Haidt, R. W. Hood, D. R. Vago, A. B. Newberg, The
727 varieties of self-transcendent experience, Review of General Psychology
728 21 (2) (2017) 143–160.

- 729 [11] D. S. Chiaburu, I.-S. Oh, C. M. Berry, N. Li, R. G. Gardner, The
730 five-factor model of personality traits and organizational citizenship be-
731 haviors: A meta-analysis, *Journal of Applied Psychology* 96 (6) (2011)
732 1140–1166.
- 733 [12] H. Zhao, S. E. Seibert, G. T. Lumpkin, The relationship of personality
734 to entrepreneurial intentions and performance: A meta-analytic review,
735 *Journal of Management* 36 (2) (2010) 381–404.
- 736 [13] L. X. Hy, J. Loevinger, *Measuring ego development*, 2nd Edition,
737 Lawrence Erlbaum Associates, Mahwah, NJ, 1996.
- 738 [14] M. Basseches, *Dialectical thinking and adult development*, Ablex, Nor-
739 wood, NJ, 1984.
- 740 [15] T. L. Dawson, Z. Stein, *Grow up: How taking development seriously*
741 *can transform your organization and your world*, Integral Publishers,
742 2011.
- 743 [16] C. Griggs, T. Strauss, *iConscious: Accelerating human potential*, iCon-
744 scious LLC, 2019.
- 745 [17] D. J. Chalmers, Facing up to the problem of consciousness, *Journal of*
746 *Consciousness Studies* 2 (3) (1995) 200–219.
- 747 [18] S. D. Gosling, P. J. Rentfrow, W. B. Swann, Jr., A very brief measure
748 of the Big-Five personality domains, *Journal of Research in Personality*
749 37 (6) (2003) 504–528. doi:10.1016/S0092-6566(03)00046-1.

- 750 [19] P. G. Reed, Self-transcendence and mental health in oldest-old adults,
751 Nursing Research 40 (1) (1991) 5–8.
- 752 [20] J. Mossbridge, K. Johnson, P. Washburn, A. Williams, M. Sapiro,
753 Smartphone time machine: Tech-supported improvements in time per-
754 spective and wellbeing measures, Frontiers in Psychology 12 (2021)
755 744209.
- 756 [21] J. Mossbridge, B. Goertzel, R. Mayet, E. Monroe, G. Nehat, D. Hanson,
757 G. Yu, Emotionally-sensitive AI-driven android interactions improve so-
758 cial welfare through helping people access self-transcendent states, in:
759 AI for Social Good Workshop, Neural Information Processing Systems
760 2018, 2018.
- 761 [22] J. Mossbridge, K. Cameron, M. Boccuzzi, State, trait, and target pa-
762 rameters associated with accuracy in two online tests of precognitive
763 remote viewing, Journal of Anomalous Experience and Cognition 4 (1)
764 (2024) 88–121.
- 765 [23] C. J. Hutto, E. Gilbert, VADER: A parsimonious rule-based model for
766 sentiment analysis of social media text, in: Proceedings of the Eighth
767 International AAAI Conference on Weblogs and Social Media, 2014, pp.
768 216–225.
- 769 [24] B. Rammstedt, O. P. John, Measuring personality in one minute or
770 less: A 10-item short version of the Big Five Inventory in English and
771 German, Journal of Research in Personality 41 (1) (2007) 203–212.

- 772 [25] J. L. Fleiss, Measuring nominal scale agreement among many raters,
773 Psychological Bulletin 76 (5) (1971) 378–382.
- 774 [26] A. J. Conger, Integration and generalization of kappas for multiple
775 raters, Psychological Bulletin 88 (2) (1980) 322–328.
- 776 [27] J. R. Landis, G. G. Koch, The measurement of observer agreement for
777 categorical data, Biometrics 33 (1) (1977) 159–174.