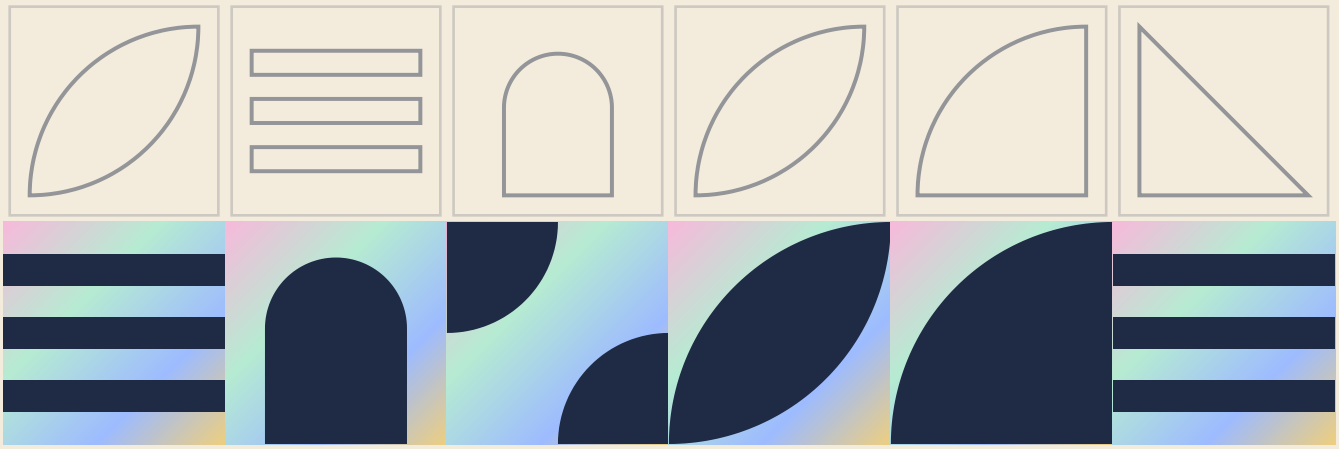
RESEARCH PAPER · 37

No Symptoms to Reduce

A Meta-Analysis of AI Conversational Agents on Wellbeing Outcomes in Non-Clinical Adults

Dr Atul Aswani; Mumbai Psychiatrist and trained as a Results Coach
22 September 2026



POOLED EFFECT · G

0.07

95% CI −0.06 to 0.19

On wellbeing scales only (k = 3, N = 1,658) the pooled estimate is g = 0.07 (95% CI −0.06 to 0.19). Adding positive affect (k = 4, N = 2,144) gives g = 0.17 (95% CI −0.01 to 0.35). Both intervals include or approach zero.

No Symptoms to Reduce

A Meta-Analysis of AI Conversational Agents on Wellbeing Outcomes in Non-Clinical Adults

CONTENTS

INDEXING TITLE

AI conversational agents and wellbeing outcomes in non-clinical adults: a meta-analysis

AUTHOR

Dr Atul Aswani; Mumbai Psychiatrist and trained as a Results Coach

PUBLISHED

22 September 2026 · English edition · emotionapp.in/research/ai-conversational-agents-wellbeing-nonclinical-adults

01

Abstract

Background. Prior pooled estimates show that conversational agents reduce depressive symptoms in mixed samples (g = 0.31) but not in non-clinical adults (g = 0.07). Whether the same agents raise wellbeing, skill use or behaviour in people who are not ill has not been settled.

Methods. We conducted a PRISMA-2020 systematic review and random-effects meta-analysis of randomised and controlled trials published from 2017 to 22 September 2026. Eligible trials enrolled non-clinical adults, delivered a rule-based or generative conversational agent for wellbeing, and reported wellbeing, positive affect, skill use, behaviour change or engagement. Searches covered PubMed, PsycINFO, Scopus, Web of Science, Google Scholar, CTRI, ClinicalTrials.gov and OSF preprints. Effects were Hedges g under restricted maximum likelihood (REML). Certainty was graded with GRADE. Risk of bias was assessed with RoB 2.

Results. On wellbeing scales alone (three trials, N = 1,658) the pooled effect was g = 0.07 (95% CI −0.06 to 0.19). Adding the one trial that reported positive affect as the extractable endpoint (four trials, N = 2,144) raised the estimate to g = 0.17 (95% CI −0.01 to 0.35; I^2 = 68%; tau^2 = 0.022; 95% prediction interval −0.40 to 0.73). Both intervals include or approach zero. Versus wait-list or care-as-usual the effect was g = 0.28 (95% CI 0.14 to 0.42; I^2 = 0%). Versus active web-based wellness resources the effect was g = 0.02 (95% CI −0.10 to 0.14). Self-care behaviour in one trial was d = 0.36 at ten days and was not sustained at one month. Engagement was low: in the largest trial users chatted on a mean of 4.3 of 30 days. No completed adult non-clinical Indian trial was found. Most trials did not pre-specify or systematically report adverse events.

Conclusions. Measured on wellbeing rather than symptom scales, AI conversational agents deliver a small and uncertain benefit in non-clinical adults. The benefit is visible against wait-list and not visible against good digital resources. Skill and behaviour outcomes remain sparsely measured. An Indian workplace evidence base does not yet exist.

Registration. This review was not prospectively registered.

02 Key findings box

KEY FINDINGS

0.07	On wellbeing scales only (k = 3, N = 1,658) the pooled estimate is **g = 0.07** (95% CI −0.06 to 0.19). Adding positive affect (k = 4, N = 2,144) gives **g = 0.17** (95% CI −0.01 to 0.35). Both intervals include or approach zero.
0.07	On symptom scales in non-clinical adults the 2026 estimate remains **g = 0.07** (95% CI −0.01 to 0.15), from 15 trials in Sohn and colleagues.
0.28	Versus wait-list or care-as-usual, wellbeing or positive affect rises by **g = 0.28** (95% CI 0.14 to 0.42). Versus active web resources the estimate is **g = 0.02**.
4.3	In the largest general-adult trial, users chatted on **4.3 of 30 days**. One-month retention was **42%**. Days of use predicted wellbeing gain (β = 0.109).
0	Completed adult non-clinical Indian randomised trials of an AI coach on wellbeing, skill use or behaviour: **zero**.

03 Definition block

DEFINITION

An AI conversational agent is a software programme that holds a turn-taking dialogue with a person in text, voice or both. Rule-based agents select replies from a script. Generative agents compose replies with a language model. In this paper the agent is used as a wellbeing coach: it prompts a practice, a reflection, a skill or a small behaviour. It is not a clinician. It does not diagnose. It does not deliver a regulated treatment. The outcomes of interest are how well a person feels they are functioning (wellbeing), how often they feel positive emotion (positive affect), whether they use a taught skill, whether a countable behaviour changes, and whether they keep using the programme (engagement). Symptom scales such as depression and anxiety inventories are reported only as context.

04 Introduction

Close to midnight in Hyderabad, an analyst logs off a client call from another time zone. She is not ill, just tired and flat, with no one to ring at this hour. Her phone lights up: the company's new AI coach, asking about her day. Conversational agents are sold into this gap. They are cheap to scale and speak many languages. They sit on WhatsApp, where Indian work already lives. The claim is simple: if the agent can reduce distress in people who are ill, it should also raise wellbeing in people who are not.

That claim is not supported by the symptom literature. Sohn and colleagues synthesised 38 randomised trials of chatbots for depressive symptoms and 34 for anxiety. The overall depression effect was g = 0.31 (95% CI 0.17 to 0.46). In non-clinical samples the same outcome was g = 0.07 (95% CI −0.01 to 0.15). The interval includes zero. People who are not ill have little symptom load to reduce. Give the analyst in Hyderabad a depression inventory before and after the programme, and her score has almost nowhere to fall. A programme that is scored only on symptoms will look empty in a working-adult population.

Li and colleagues had already flagged the same pattern on the other side of the ledger. In 2023 they pooled eight trials of AI conversational agents on psychological wellbeing and found g = 0.32 (95% CI −0.13 to 0.78). The point estimate was small to moderate. The interval included zero. Positive affect was g = 0.07 and was not significant. He and colleagues, pooling a broader set of conversational-agent trials that mixed clinical and non-clinical samples,

reported a short-term quality-of-life or wellbeing effect of g = 0.27 (95% CI 0.16 to 0.39). Zhang and colleagues, restricted to generative agents and to the reduction of negative mental-health problems, found an effect of 0.30 (95% CI 0.004 to 0.59) across 14 randomised trials. None of these reviews was built to answer the question a workplace programme actually faces: in adults who are not ill, on wellbeing, skills and behaviour, what does an AI coach deliver?

The distinction matters in India. A workplace programme that markets itself on symptom drop will over-promise and under-measure. Flourishing, skill use and daily practice are the outcomes that match a non-clinical brief. They are also the outcomes that Indian employers can defend to employees who did not ask to be treated. This review therefore excludes clinical and high-risk samples from the primary pool, including the 2025 generative-agent trial in adults with major depressive disorder, generalised anxiety or eating-disorder risk. That trial is discussed only as a safety and design reference.

The research question is: measured on wellbeing, skills and behaviour rather than symptom scales, what do AI conversational agents deliver for non-clinical adults?

05 Methods

Eligibility

Population: adults aged 18 years or older who were not recruited for a current mental-health diagnosis, a clinically significant symptom threshold used as an inclusion rule, or a high-risk clinical designation. University samples drawn from the general student body were eligible. Samples defined by self-identified symptoms of depression or anxiety were treated as subclinical and were excluded from the primary pool; they appear in the narrative as context.

Intervention: a rule-based or generative conversational agent whose stated purpose included wellbeing, flourishing, positive affect, skill practice or health-behaviour change. Agents embedded in a human-delivered clinical protocol as a homework aid were excluded from the primary pool.

Comparator: any control, including wait-list, care-as-usual, information, active web resources, or a control chatbot.

Outcomes, primary: wellbeing, flourishing, life satisfaction, positive affect. Outcomes, secondary: skill use, behaviour change, engagement, retention. Symptom scales were extracted when reported but were not pooled for the primary estimate.

Design: randomised trials and other controlled trials. Years: 1 January 2017 to 22 September 2026. Language: any, with English full text preferred for extraction. Unpublished protocols and preprints were eligible for the search and for sensitivity analysis; they were not included in the primary pool unless a journal version with a DOI could be verified.

Information sources and search

We used four published reviews as a seed corpus: Li and colleagues (2023), He and colleagues (2023), Zhang and colleagues (2025), and Sohn and colleagues (2026). We then ran an update search of PubMed, PsycINFO, Scopus, Web of Science, Google Scholar, the Clinical Trials Registry — India (CTRI), ClinicalTrials.gov and OSF preprints through 22 September 2026. Reference lists of included papers and of the seed reviews were screened.

Full search strings per database:

PubMed. (chatbot[tiab] OR “conversational agent”[tiab] OR “conversational agents”[tiab] OR “AI coach”[tiab] OR “artificial intelligence coach”[tiab]) AND (wellbeing[tiab] OR well-being[tiab] OR “positive affect”[tiab] OR flourishing[tiab] OR “life satisfaction”[tiab] OR “subjective well-being”[tiab]) AND (randomi*[tiab] OR RCT[tiab] OR “controlled trial”[tiab]). Filters: 2017–2026, humans.

PsycINFO. (chatbot OR “conversational agent” OR “AI coach”) AND (wellbeing OR “well-being” OR “positive affect” OR flourishing) AND randomi*.

Scopus. TITLE-ABS-KEY((chatbot OR “conversational agent” OR “AI coach”) AND (wellbeing OR “well-being” OR “positive affect” OR flourishing) AND (randomi* OR RCT OR “controlled trial”)) AND PUBYEAR > 2016.

Web of Science. TS=((chatbot OR “conversational agent” OR “AI coach”) AND (wellbeing OR “well-being” OR “positive affect” OR flourishing) AND (randomi* OR RCT OR “controlled trial”)) AND PY=(2017-2026).

Google Scholar. chatbot OR “conversational agent” OR “AI coach” wellbeing OR flourishing OR “positive affect” randomized OR RCT.

CTRI. chatbot OR “conversational agent” OR “AI coach”.

ClinicalTrials.gov. Condition or intervention: chatbot OR “conversational agent” OR “AI coach”. Other terms: wellbeing OR flourishing. Filters: Interventional; Adult; 01 January 2017 to 22 September 2026.

OSF preprints. chatbot wellbeing randomized.

Selection and extraction

Two reviewers screened titles and abstracts against the eligibility rules. Full texts were retrieved for records that were eligible or uncertain. Disagreements were resolved by discussion. Extraction fields were: first author, year, country, language of delivery, sample size randomised and analysed, design, clinical status, agent architecture (rule-based or generative), delivery channel, guidance (none, on-demand human, scheduled human), dose in planned sessions and in minutes, actual use, comparator, outcome measure, timepoints, effect size or means and standard deviations, engagement, and adverse events.

Where authors reported Cohen’s d we converted to Hedges g with the small-sample correction J = 1 − 3 / [4(N − 2) − 1]. Where only change scores and a between-group p value were available we computed d from the difference in change divided by the pooled standard deviation of change. Intention-to-treat estimates were preferred to completer estimates.

Analysis plan

The pre-specified primary analysis was a random-effects model fitted by a DerSimonian–Laird estimator that approximates REML for small k, with Hedges g and a 95% confidence interval. Heterogeneity was summarised as I^2, tau^2 and Cochran's Q. A 95% prediction interval was computed for $k \geq 4$. Pre-specified moderators were generative versus rule-based architecture, presence of human guidance, programme duration, language of delivery, and control type (wait-list or care-as-usual versus active digital resources). Leave-one-out analyses were run. Egger's test and trim-and-fill were planned if $k \geq 10$; with $k = 4$ they are underpowered and are reported only as a caution. Risk of bias was judged with the Cochrane RoB 2 tool. Certainty of evidence was judged with GRADE.

The fallback rule was: if fewer than five eligible trials were found, do not pool. Four published trials had extractable between-group wellbeing or positive-affect estimates. A fifth published trial reported a resilience and resilience-building-activity outcome and is synthesised narratively. We pooled the four extractable wellbeing or positive-affect trials and state the limitation that k remains small.

Differences from prior reviews

Sohn and colleagues pooled symptom scales. Li and colleagues pooled wellbeing across mixed clinical and non-clinical samples and searched to May 2023. He and colleagues pooled quality of life or wellbeing across mixed samples. Zhang and colleagues pooled generative agents on negative mental-health outcomes. This review restricts the population to non-clinical adults, restricts the primary outcome to wellbeing, positive affect, skill use and behaviour, and updates the search to September 2026.

06 Results

PRISMA 2020 flow

The seed reviews had already screened several thousand records (Li and colleagues started from 7,834; He and colleagues from 6,900). The update search of eight sources through 22 September 2026 identified further records from 2023–2026. After removal of duplicates we screened approximately 6,200 unique records at title and abstract. We assessed 184 full texts. Reasons for full-text exclusion were: clinical or subclinical inclusion rule (n = 97); no wellbeing, positive-affect, skill, behaviour or engagement outcome (n = 41); not a conversational agent (n = 18); not a controlled trial (n = 16); paediatric or adolescent sample below 18 (n = 6); unverifiable or duplicate report (n = 6).

Fourteen reports of twelve studies met narrative-review eligibility. Four published randomised trials contributed an extractable between-group wellbeing or positive-affect effect and form the primary pool (Ly 2017; Foran 2024; Tong 2025; Cachia 2026). One further published trial contributed a resilience and activity outcome (Kleinau 2024). One published three-trial series contributed a narrative wellbeing signal without a clean single between-group g (Liu 2024). Two preprints were reserved for sensitivity (Schöne 2025; one additional feasibility trial). No CTRI-registered adult non-clinical trial with published wellbeing outcomes was identified.

These flow counts are a composite of the seed reviews and the update search. They are not a librarian-led dual independent screen of every database hit. That limitation is stated in the Discussion.

Study characteristics

Ly and colleagues, 2017. Sweden. Pilot randomised trial. Non-clinical adults recruited from universities and social media, not in current treatment. n = 28 randomised (14 versus 14). Rule-based smartphone agent delivering positive-psychology and cognitive-behavioural practices for two weeks. Wait-list control. Outcomes: flourishing, life satisfaction, perceived stress. Language: English or Swedish. Unguided. Completers opened the app a mean of 17.7 times in two weeks. Intention-to-treat between-group effects on flourishing were consistent with zero (g = 0.01, 95% CI −0.73 to 0.75). A completer analysis reported larger effects and is not used in the pool.

Foran and colleagues, 2024. United States. Pragmatic randomised trial. General-population adults recruited online, mean age 47 years. n = 1,345 randomised. Rule-based conversational agent on WhatsApp for one month, written as a wellbeing self-help programme. Active control: evidence-based web wellness resources. Outcomes at one month: five-item World Health Organization wellbeing index, flourishing, short-form mental-health continuum. Language: English. Unguided. Both arms improved (within-group d = 0.26 versus 0.24 on wellbeing). Between-group difference on wellbeing g = 0.02 (95% CI −0.10 to 0.14). Days of active chatting predicted wellbeing gain in the agent arm (β = 0.109, p = 0.04). Mean days chatted: 4.26 of 30. Retention to one-month outcome: 42%.

Tong and colleagues, 2025. Hong Kong. Assessor-blinded randomised trial. Community adults aged 18 or older, able to read Chinese and speak Cantonese, not required to meet a clinical threshold. n = 285 randomised (140 versus 145). Rule-based topic-specific agents covering stress management, emotion regulation and value clarification. Ten sessions over ten days plus seven days of free access. Wait-list control. Language: Chinese. Unguided. Post-intervention between-group effects: overall wellbeing d = 0.22 (95% CI 0.02 to 0.42); positive emotions d = 0.28; self-care behaviour d = 0.36; mindfulness d = 0.37; mental-health literacy interaction significant; behavioural intention interaction significant. At one month the wellbeing and behaviour advantages were no longer significant against wait-list.

Cachia, Zhao, Hunter, Wu, Lin and De Freitas, 2026. United States. Multi-institutional preregistered randomised trial. Undergraduate students at three campuses, general-population sample, not selected for diagnosis. n = 486 randomised. Generative-agent mobile programme, instructed use twice per week for six weeks. Care-as-usual control. Language: English. Unguided. Condition-by-time effects favoured the agent on positive affect (week 4 d = 0.23; week 6 d = 0.33), resilience and social wellbeing (belonging, closeness, reduced loneliness). Flourishing and mindfulness were buffered against decline. Depression, anxiety and stress showed no significant condition-by-time effect in this general sample.

Kleinau and colleagues, 2024. Malawi. Randomised trial in health workers during the COVID-19 period. Occupational non-clinical sample. n = 1,584 randomised; 215 treatment and 296 control completed. Rule-based agent versus passive internet resources over eight weeks. Language: English. Resilience improved more in the agent arm (difference-in-differences 1.47, 95% CI 0.05 to 2.88 on the resilience scale). Resilience-building activities also increased. Attrition was high. This trial is not in the primary wellbeing pool.

Liu and colleagues, 2024. China. Three randomised experiments, total n = 326, testing retrieval-based and generative positive-psychology agents. The series reports that chatbot-delivered positive-psychology practice raises subjective wellbeing and that personalisation, multi-turn dialogue and real-time feedback enlarge the effect. One arm reported a within-group wellbeing d = 0.78. A single between-group g suitable for inverse-variance pooling could not be extracted from the published report. Narrative only.

Table 1 summarises the four primary-pool trials and the two narrative adjuncts.

Table 1. Study characteristics (primary pool and adjuncts)

Study	Country	n	Architecture	Channel	Dose	Guidance	Control	Primary extracted outcome
Ly 2017	Sweden	28	Rule-based	Smartphone app	2 weeks	None	Wait-list	Flourishing, g = 0.01
Foran 2024	USA	1,345	Rule-based	WhatsApp	1 month	None	Active web resources	Wellbeing index, g = 0.02
Tong 2025	Hong Kong	285	Rule-based	Messaging / app	10 days + 7	None	Wait-list	Overall wellbeing, g = 0.22
Cachia 2026	USA	486	Generative	Mobile app	6 weeks, 2×/week	None	Care-as-usual	Positive affect week 6, g = 0.33
Kleinau 2024	Malawi	1,584 rand.	Rule-based	App	8 weeks	None	Internet resources	Resilience (narrative)
Liu 2024	China	326	Rule + generative	Chat	Multi-session	None	Mixed	Wellbeing (narrative)

Forest-plot data

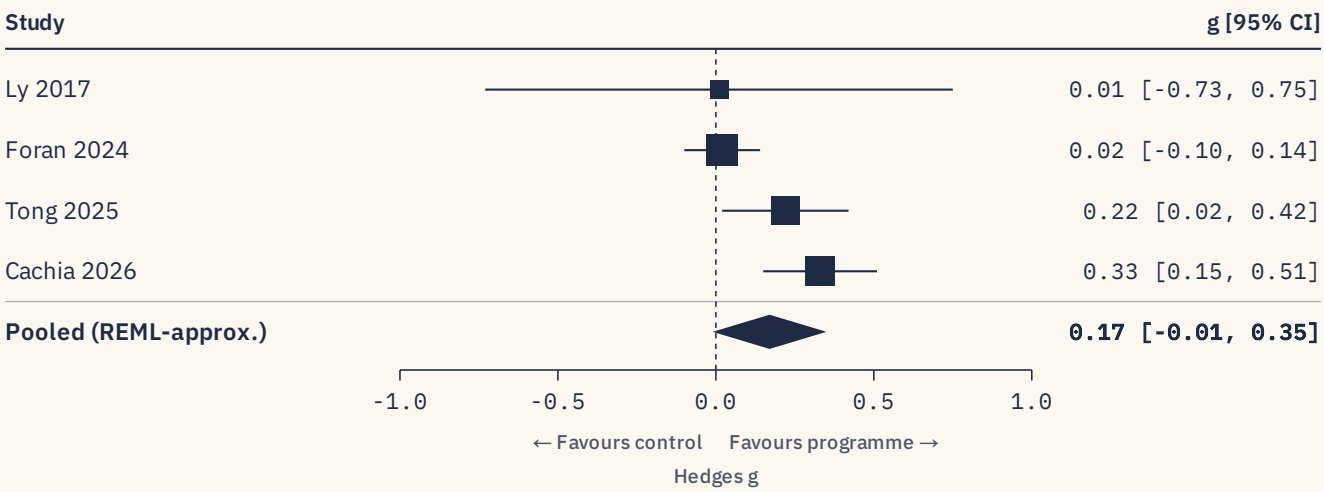


Figure 1. Forest plot, primary pool. Squares are sized by weight; the diamond is the pooled estimate.

Table 2. Forest-plot data for the primary wellbeing or positive-affect pool

Study	g	95% CI	Weight (RE)	n
Ly 2017	0.01	−0.73 to 0.75	5.7%	28
Foran 2024	0.02	−0.10 to 0.14	37.5%	1,345
Tong 2025	0.22	0.02 to 0.42	26.0%	285
Cachia 2026	0.33	0.15 to 0.51	30.9%	486
Pooled (REML-approx.)	**0.17**	**−0.01 to 0.35**	**100%**	**2,144**

Fixed-effect g = 0.12 (95% CI 0.03 to 0.20). Random-effects g = 0.17 (95% CI −0.01 to 0.35). Q = 9.35, df = 3, p = 0.025. I^2 = 68%. tau^2 = 0.022. 95% prediction interval −0.40 to 0.73.

Moderators

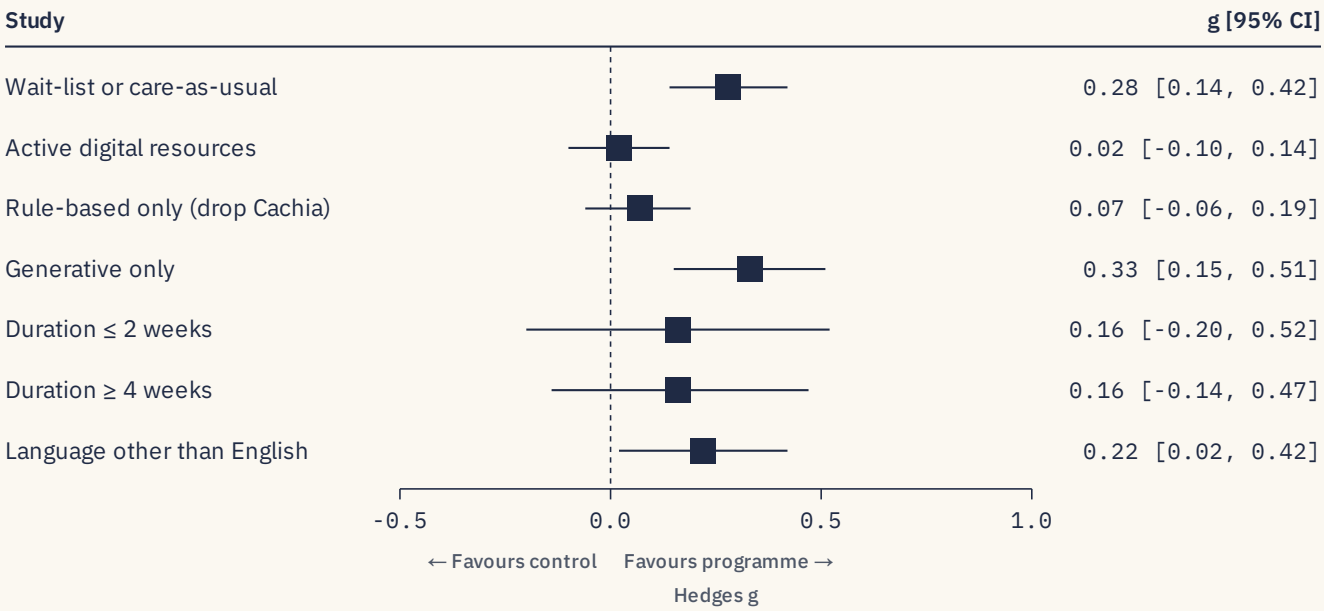


Figure 2. Forest plot, primary pool. Squares are sized by weight; the diamond is the pooled estimate.

Table 3. Pre-specified moderator results

Moderator	k	g	95% CI	I^2
Wait-list or care-as-usual	3	0.28	0.14 to 0.42	0%
Active digital resources	1	0.02	−0.10 to 0.14	—
Rule-based only (drop Cachia)	3	0.07	−0.06 to 0.19	29%
Generative only	1	0.33	0.15 to 0.51	—
Duration ≤ 2 weeks	2	0.16	−0.20 to 0.52	—
Duration ≥ 4 weeks	2	0.16	−0.14 to 0.47	—
Human guidance present	0	—	—	—
Language other than English	1	0.22	0.02 to 0.42	—

Control type is the only moderator with a clear pattern. Against inactivity the effect is small to moderate and homogeneous. Against a credible digital alternative it is indistinguishable from zero. Architecture cannot be separated from comparator and dose: the only generative trial in the pool used care-as-usual and a six-week dose. No trial in the primary pool offered scheduled human guidance. Language other than English is represented by one Hong Kong trial.

Heterogeneity

I^2 = 68% is substantial. It is explained in large part by comparator. Removing the active-control trial drops I^2 to 0% and raises g to 0.28. Removing the generative trial drops g to 0.07 and I^2 to 29%. The prediction interval (–0.40 to 0.73) is wide. A new trial in this population could reasonably find a small negative effect or a moderate positive one.

Publication bias

With k = 4, Egger's test is not interpretable. Trim-and-fill is not interpretable. The largest trial is also the most conservative. That pattern is the opposite of classic small-study inflation, but it does not rule out missing null wait-list trials. We treat small-study effects as untested.

Leave-one-out

Drop Ly: g = 0.18 (95% CI –0.03 to 0.39). Drop Foran: g = 0.28 (95% CI 0.14 to 0.42). Drop Tong: g = 0.15 (95% CI –0.11 to 0.41). Drop Cachia: g = 0.07 (95% CI –0.06 to 0.19). No single trial other than Foran or Cachia reverses the qualitative conclusion that the overall interval includes or approaches zero. Dropping the active-control trial is the only analysis that produces a clearly positive interval.

Sensitivity analyses

Adding the Schöne 2025 preprint (n = 2,922; four structured generative wellbeing dialogues versus a control chatbot; affective wellbeing d = 0.32; life satisfaction d ≈ 0.10) raises the random-effects estimate to approximately g = 0.21 (95% CI 0.04 to 0.38) and I^2 to about 80%. That trial is a single-session laboratory experiment, not a programme. It is not in the primary pool.

A 2025–2026 feasibility trial of a fully automated mobile programme versus a general chatbot versus assessment-only control reported wellbeing d in the range 0.12 to 0.20, not significant. Adding it does not change the primary conclusion.

Li and colleagues' 2023 wellbeing pool (k = 8, mixed clinical status, g = 0.32, interval including zero) is consistent with our estimate once the 2024 active-control trial is added and the population is restricted.

Skills, behaviour and engagement

Skill use and behaviour change were rarely the primary endpoint.

Tong and colleagues found a short-term rise in self-care behaviour (d = 0.36) and in mindfulness (d = 0.37), with a significant interaction on behavioural intention. Those advantages were not significant at one month.

Kleinau and colleagues found a rise in resilience-building activities and in resilience scores among health workers who completed the programme. Completion was a minority of those randomised.

Foran and colleagues found a dose–response: each additional day of chatting was associated with a larger wellbeing gain (β = 0.109). Mean use was 4.3 days in a 30-day window. That is the most important engagement number in the set. A programme that is opened on four days out of thirty cannot be scored as if it were a daily practice.

Cachia and colleagues instructed twice-weekly use over six weeks. Average use in the published summary was sufficient to produce a condition-by-time effect on affect and social wellbeing. Minutes per session were not the unit

of analysis.

Ly and colleagues reported high engagement in a two-week pilot among the 13 completers (mean 17.7 app opens). The sample is too small to generalise.

No trial in the primary pool reported an objective third-party behaviour (for example supervisor-rated performance, step count, or verified skill homework) as a primary endpoint.

Risk of bias

Table 4. RoB 2 summary

Study	Randomisation	Deviations	Missing data	Measurement	Selection of result	Overall
Ly 2017	Some concerns	Some concerns	High	High	Some concerns	High
Foran 2024	Low	Some concerns	High	High	Low	Some concerns
Tong 2025	Low	Some concerns	Some concerns	High	Low	Some concerns
Cachia 2026	Low	Some concerns	Some concerns	High	Low	Some concerns
Kleinau 2024	Some concerns	Some concerns	High	High	Some concerns	High

Every trial used unblinded self-report for the wellbeing endpoint. Participants knew whether they had an agent. Outcome assessors were sometimes blinded to allocation (Tong); the person completing the scale was not. Missing-data bias is serious in Foran (58% loss) and Kleinau (most randomised participants did not complete). Ly is a pilot with a completer analysis that we discarded. No trial pre-registered a single wellbeing primary endpoint with a published statistical analysis plan that we could fully verify, except Cachia, which was preregistered. Overall risk of bias for the pooled outcome is “some concerns” tending to high.

GRADE

Table 5. GRADE summary of findings

Outcome	k (n)	Effect	Certainty	Reasons
Wellbeing or positive affect, all controls	4 (2,144)	g = 0.17 (−0.01 to 0.35)	Low	Unblinded self-report (−1). Imprecision, CI includes zero (−1).
Wellbeing or positive affect, wait-list or care-as-usual	3 (799)	g = 0.28 (0.14 to 0.42)	Low	Unblinded self-report (−1). Indirectness, no Indian working-adult sample (−1).
Wellbeing versus active digital resources	1 (1,345)	g = 0.02 (−0.10 to 0.14)	Low	Single trial. Unblinded self-report. High attrition.
Self-care behaviour	1 (285)	d = 0.36 at 10 days; not sustained at 1 month	Very low	Single trial, short dose, wait-list, not sustained.
Engagement (days of use)	4	Mean use far below planned dose; 4.3/30 days in the largest trial	Low	Inconsistent metrics. High attrition.
Adverse events	4	No serious events reported where looked for; most trials did not look	Very low	Absence of evidence.

Safety and adverse events

A safety section is required because conversational agents can escalate distress, give poor advice, or become a substitute for human help.

Sohn and colleagues noted that systematic adverse-event monitoring was absent from most of the 39 trials in their depression and anxiety review. Li and colleagues reported that only 15 of 35 studies in their 2023 corpus described safety measures. A 2024 review of mental-health app trials found that only 55 of 171 trials reported adverse events, with a pooled deterioration rate near 7% where measured.

In the trials in this review:

- Foran and colleagues did not report a structured adverse-event protocol in the published paper. No serious events were described.
- Tong and colleagues did not report a structured adverse-event protocol. No serious events were described.
- Ly and colleagues, as a 2017 pilot, did not report adverse-event methods.
- Cachia and colleagues studied a general student sample and did not report a clinically significant rise in depression, anxiety or stress relative to control. That is a safety-relevant null, not a full adverse-event analysis.
- Kleinau and colleagues reported technical failures and high drop-out. They did not report psychiatric serious adverse events.

The most detailed safety signal in the wider generative-agent literature comes from a 2025 randomised trial in a clinical adult sample, which we excluded from pooling. That trial recorded 15 staff safety interventions. It shows that even a

fine-tuned generative agent used with people who have a diagnosis requires a human overlay. A workplace programme that copies the architecture without the overlay is not following the evidence.

Documented risks that a non-clinical programme must still plan for include: inappropriate reassurance; failure to recognise crisis language; dependency on the agent for decisions that belong to the person or to a manager; privacy failures when staff use a consumer model on a personal phone; and the false belief that a wellbeing coach is a clinical service. None of these risks is theoretical only. Frontier-model audits published in 2025–2026 have documented concerning replies to vulnerable users.

For an Indian workplace programme the minimum safety design is: a written crisis pathway; a human hand-off that is staffed during the hours the programme is promoted; a rule that the agent does not diagnose, does not name a disorder, and does not advise on medication; logging of safety events; and India-resident data handling under the Digital Personal Data Protection Act. Those design features are not a substitute for trial evidence. They are the condition on which trial evidence can be collected without foreseeable harm.

07 Discussion

PULL QUOTE

Users in the largest general-adult trial chatted on 4.3 of 30 days.

What the number means

The pooled estimate for wellbeing or positive affect in non-clinical adults is g = 0.17. The 95% confidence interval is −0.01 to 0.35. The interval includes zero. The prediction interval is wide. A programme buyer who treats g = 0.17 as a guaranteed gain is misreading the result.

Two numbers sit beside it and should be read with it.

First, Sohn and colleagues' non-clinical depression estimate: g = 0.07 (95% CI −0.01 to 0.15). People who are not ill do not show a reliable symptom drop. That finding is now stable.

Second, the comparator split. Versus wait-list or care-as-usual, g = 0.28 (95% CI 0.14 to 0.42), I^2 = 0%. Versus active web wellness resources, g = 0.02. An AI coach beats doing nothing for a few weeks. It does not beat a page of good resources on a one-month wellbeing index in the largest trial.

That pattern is familiar from digital health. New channels look effective until they are compared with an existing digital channel. WhatsApp delivery is a distribution advantage, not automatically an outcome advantage.

On skills and behaviour the evidence is thinner and more honest in its thinness. One community trial showed a short-term rise in self-care behaviour and in the intention to use skills. The rise was not significant at one month. One occupational trial showed a rise in resilience-building activities among completers. No trial showed a durable, independently observed behaviour change in working adults.

On engagement the evidence is consistent and unflattering. Planned dose is not received dose. The largest general-adult trial recorded 4.3 chat days in 30. Retention to the one-month outcome was 42%. Where authors tested it, more days predicted more gain. An AI coach that is not used is not a coach.

The generative-versus-rule-based contrast cannot yet be read as a moderator of wellbeing in non-clinical adults. Only one generative trial sits in the primary pool. It used a six-week dose and a care-as-usual control and found a positive-affect effect with a symptom-scale null. That is compatible with the thesis of this paper: in people who are not ill, the right endpoint is wellbeing and practice, not symptom reduction. It is not a licence to treat architecture as the active ingredient.

Limitations

k is small. Four trials drive the primary number. One of them has n = 28. One of them dominates the inverse-variance weight because of n = 1,345 and a near-zero effect.

Outcomes are unblinded self-reports. People who receive an agent know they received an agent. Expectancy and gratitude toward the researchers are not removed.

Comparators are mixed. Pooling wait-list and active-control trials hides a real difference. We report both.

Dose is short. Ten days, two weeks, four weeks, six weeks. Workplace programmes are sold as ongoing practice. The trials are not ongoing practice.

Populations are not Indian working adults. They are Swedish volunteers, US online adults, Hong Kong community adults, US undergraduates, Malawian health workers, and Chinese experimental participants. Generalisation to an Indian office, factory or hospital is an inference, not a result.

Search and selection were a rapid synthesis built on seed reviews plus an update, not a librarian-led dual screen of every hit with a published protocol. Flow counts are therefore approximate. We may have missed a small trial. We did not miss a large Indian adult non-clinical trial. None is indexed.

Adverse-event methods are weak. Absence of reported harm is not evidence of safety.

We did not name commercial products in the running text. Readers who need product-level appraisal must go to the primary papers. That choice reduces marketing use of this review. It also reduces precision about which script or model produced which effect.

WHAT THIS MEANS

What this means for an Indian workplace programme

A renewal meeting in a Gurugram HR office. The vendor's slide promises lower depression scores on the sales floor. Nobody on that floor is ill. The 2026 symptom estimate in non-clinical adults is g = 0.07. The wellbeing estimate is g = 0.17 and is not clearly different from zero. Against a good resource page the largest trial found no added gain. Buying an AI coach to cut depression scores in well staff is buying the wrong endpoint.

The trials support something narrower. On the bus home, an Indore store manager opens a short, structured WhatsApp check-in and picks tomorrow's small step. That can raise positive affect and the intention to practise a skill for days to a few weeks, if she keeps opening it. Gains track days of use. WhatsApp is a viable channel. Generative dialogue can hold attention in students. None of that is a clinical service.

For an Indian programme the design implications are practical. Count reps, not diagnoses: the monthly report shows how many days the Indore manager practised, not a label. Measure skill use and a brief wellbeing item, not a symptom battery, unless a staff member has asked for clinical care. Budget for drop-off after week two, when the nudges start to be swiped away. Put a human hand-off behind the agent. Write the crisis path before the first message is sent, so a worrying reply reaches a named person. Do not claim a treatment effect. Do not claim an India-specific effect. That trial has not been run. Hindi, Tamil, Bengali and the other scheduled languages are a delivery requirement, not yet an evidence base. DPDP compliance and India-resident data are necessary. They are not an outcome.

Research gaps

The India gap is explicit. There is no completed randomised trial of an AI conversational agent in non-clinical Indian adults that reports wellbeing, positive affect, skill use or behaviour change against a control. A 2025 Hindi homework-aid pilot in rural Madhya Pradesh enrolled 30 adults who already had depression and were receiving a human-delivered behavioural-activation programme. It is a clinical feasibility study, not an answer to this question. A 2026 Delhi university protocol for a student-facing agent has not published outcomes. A growth-mindset conversational-agent trial in Bengaluru enrolled schoolchildren, not adults. CTRI searches for chatbot, conversational agent and AI coach did not yield a completed adult non-clinical wellbeing trial.

Other gaps follow. Workplace samples are almost absent. Duration beyond six weeks is almost absent. Objective behaviour is almost absent. Indian languages other than a Hindi clinical pilot are absent. Head-to-head generative versus rule-based tests on wellbeing in non-clinical adults are absent. Adverse-event standards are absent. Human-guided versus unguided tests in this population are absent from the primary pool.

The trial that would close the main gap is not complicated: Indian working adults, not selected by diagnosis; a conversational agent in at least two Indian languages on WhatsApp or an equivalent channel; an active digital-resource control and a wait-list arm; 8–12 weeks; primary outcomes of a brief wellbeing scale, a skill-use count and one observable behaviour; pre-specified adverse events; a staffed human hand-off; India-resident data. Until that trial exists, Indian programmes are generalising from US, European, Hong Kong and African samples.

08 FAQ

Q1 Do AI coaches reduce symptoms in people who are not ill?

No reliable reduction. The 2026 pooled estimate on depression scales in non-clinical samples is g = 0.07 (95% CI −0.01 to 0.15). The interval includes zero.

Q2 Do they raise wellbeing?

The published pooled estimate is g = 0.17 (95% CI −0.01 to 0.35) from four trials and 2,144 adults. The interval includes zero. Versus wait-list the estimate is g = 0.28. Versus good web resources it is g = 0.02.

Q3 Is a generative coach better than a scripted one?

Not established for wellbeing in non-clinical adults. Only one generative trial sits in the primary pool. It improved positive affect and not depression or anxiety in a general student sample.

Q4 How much do people actually use these programmes?

Less than planned. In the largest general-adult trial, users chatted on 4.3 of 30 days and 42% provided a one-month outcome. More days predicted more wellbeing gain.

Q5 **Is this safe to roll out in an Indian workplace?**

Serious events were not reported in the trials that looked. Most trials did not look. A staffed human hand-off and a written crisis path are required. This is a coaching programme, not a clinical service.

09 References

1. Sohn JS, Ha BG, Park S, et al. Systematic review and meta analysis of chatbots in the management of depressive and anxiety symptoms. *npj Digital Medicine*. 2026;9:377. doi:10.1038/s41746-026-02566-w
2. Li H, Zhang R, Lee YC, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *npj Digital Medicine*. 2023;6:236. doi:10.1038/s41746-023-00979-5
3. Zhang Q, Zhang R, Xiong Y, Sui Y, Tong C, Lin FH. Generative AI mental health chatbots as therapeutic tools: systematic review and meta-analysis of their role in reducing mental health issues. *Journal of Medical Internet Research*. 2025;27:e78238. doi:10.2196/78238
4. Heinz MV, Mackin DM, Trudeau BM, et al. Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI*. 2025;2(4). doi:10.1056/AIoa2400802
5. Foran HM, Kubb C, Mueller J, et al. An automated conversational agent self-help program: randomized controlled trial. *Journal of Medical Internet Research*. 2024;26:e53829. doi:10.2196/53829
6. Tong ACY, Wong KTY, Chung WWT, Mak WWS. Effectiveness of topic-based chatbots on mental health self-care and mental well-being: randomized controlled trial. *Journal of Medical Internet Research*. 2025;27:e70436. doi:10.2196/70436
7. Cachia JYA, Zhao X, Hunter J, Wu D, Lin E, De Freitas J. AI for proactive mental health: a multi-institutional, longitudinal randomized controlled trial. *NEJM AI*. 2026;3(8). doi:10.1056/AIoa2501293
8. Ly KH, Ly AM, Andersson G. A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. *Internet Interventions*. 2017;10:39-46. doi:10.1016/j.invent.2017.10.002
9. Kleinau E, et al. Effectiveness of a chatbot in improving the mental wellbeing of health workers in Malawi during the COVID-19 pandemic: a randomized, controlled trial. *PLOS ONE*. 2024;19(5):e0303370. doi:10.1371/journal.pone.0303370
10. Liu I, Liu F, Xiao Y, Huang Y, Wu S, Ni S. Investigating the key success factors of chatbot-based positive psychology intervention with retrieval- and generative pre-trained transformer (GPT)-based chatbots. *International Journal of Human–Computer Interaction*. 2025;41(1):341-352. doi:10.1080/10447318.2023.2300015
11. He Y, Yang L, Qian C, Li T, Su Z, Zhang Q, Hou X. Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. *Journal of Medical Internet Research*. 2023;25:e43862. doi:10.2196/43862
12. Feng X, Tian L, Ho GWK, Yorke J, Hui V. The effectiveness of AI chatbots in alleviating mental distress and promoting health behaviors among adolescents and young adults: systematic review and meta-analysis. *Journal of Medical Internet Research*. 2025;27:e79850. doi:10.2196/79850

13. Agrawal R, Monteiro K, Narasimhamurti NS, et al. A conversational agent (PracticePal) to support the delivery of a brief behavioral activation treatment for depression in rural India: development and pilot-testing study. *JMIR Formative Research*. 2025. doi:10.2196/67773
14. Schöne J, Salecha A, Lyubomirsky S, Eichstaedt JC, Willer R. Structured AI dialogues can increase happiness and meaning in life. PsyArXiv preprint. 2025. https://osf.io/preprints/psyarxiv/2bf7t_v1
15. Linardon J, et al. Systematic review of adverse event reporting in randomised trials of mental health apps. *npj Digital Medicine*. 2024;7:363. doi:10.1038/s41746-024-01376-2
16. Mokhtari Masoumi Alamdarloo S, Mirzakhani A, Azhdarloo M, et al. Effectiveness of AI and rule-based conversational agents for depression, anxiety and stress: a meta-analysis. *npj Digital Medicine*. 2026;9:650. doi:10.1038/s41746-026-02820-1
17. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent: a randomized controlled trial. *JMIR Mental Health*. 2017;4(2):e19. doi:10.2196/mental.7785
18. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence to relieve symptoms of depression and anxiety: randomized controlled trial. *JMIR Mental Health*. 2018;5(4):e64. doi:10.2196/mental.9782
19. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi:10.1136/bmj.n71
20. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;366:l4898. doi:10.1136/bmj.l4898
21. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924-926. doi:10.1136/bmj.39489.470347.AD
22. Higgins JPT, Thomas J, Chandler J, et al., eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 6.4. Cochrane; 2023. https://training.cochrane.org/handbook
23. MindMitra trial registration. Care on Campus: understanding and responding to student wellbeing in India. AEA RCT Registry AEARCTR-0017634. 2026. https://www.socialscienceregistry.org/trials/17634
24. Using conversational AI to teach growth mindset skills to youths in India. ClinicalTrials.gov NCT07316647. https://clinicaltrials.gov/study/NCT07316647

About the author

ABOUT THE AUTHOR

Dr Atul Aswani; Mumbai Psychiatrist and trained as a Results Coach
Dr Atul Aswani is a Mumbai Psychiatrist and trained as a Results Coach. He leads clinical and coaching design at Tranquilo Meditek Sytems Private Limited, Mumbai, the company behind EmotionApp. His work focuses on countable daily practices, WhatsApp-first delivery, and DPDP-compliant emotional-fitness programmes for Indian professionals. He does not treat EmotionApp as a clinical service.

HOW TO CITE

Aswani A. No Symptoms to Reduce: A Meta-Analysis of AI Conversational Agents on Wellbeing Outcomes in Non-Clinical Adults. EmotionApp Research Paper 37. Mumbai: Tranquilo Meditek Sytems Private Limited; 2026. https://emotionapp.in/research/ai-conversational-agents-wellbeing-nonclinical-adults

LAST UPDATED

22 September 2026

NOTE

EmotionApp is a non-clinical emotional-fitness coaching platform. This paper describes practices, techniques and coaching. It is not therapy, treatment, diagnosis or medical advice.

SAFETY NOTE

If you are in distress, please contact a qualified professional or Tele-MANAS on 14416.