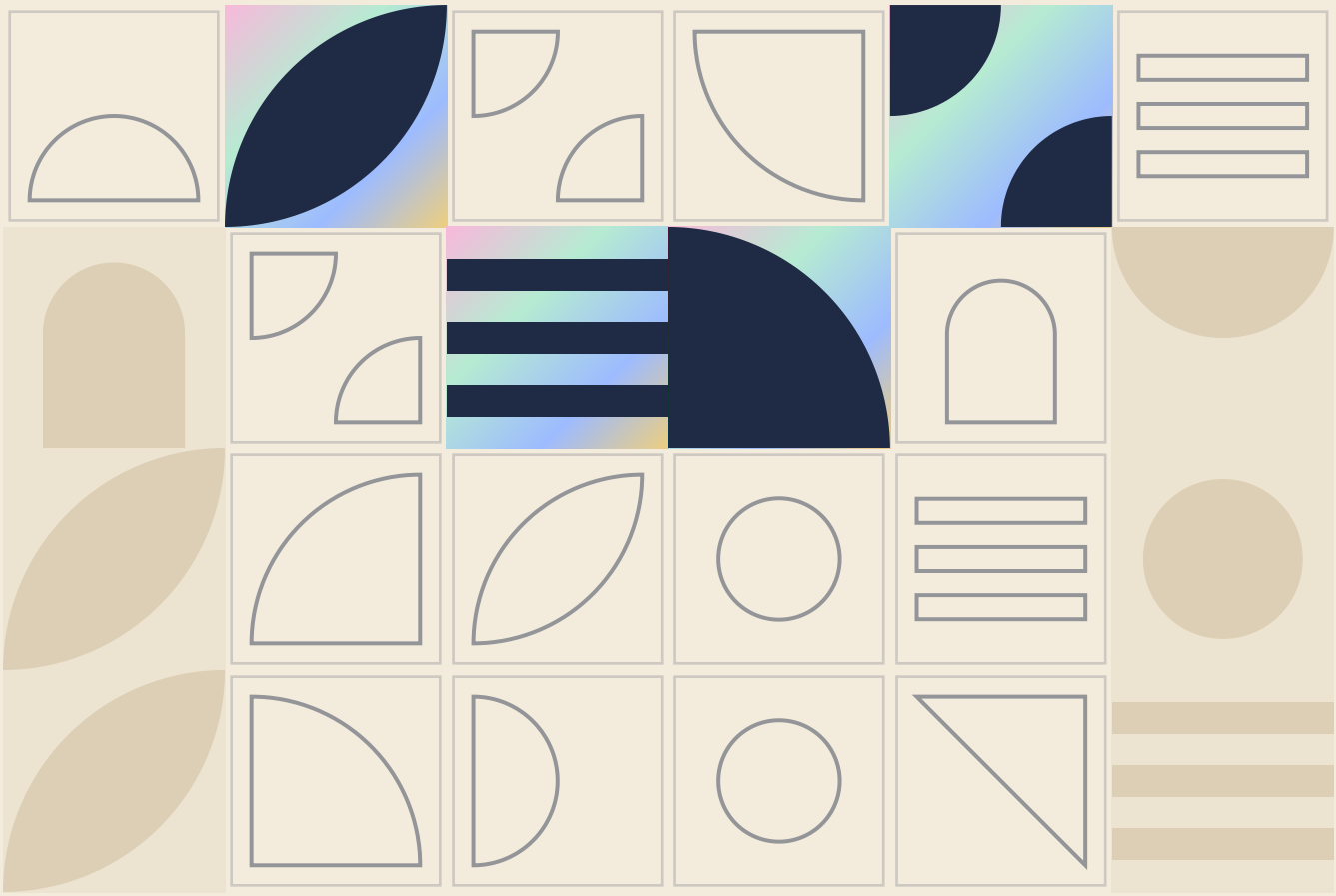
Can You Bond With a Bot?

A systematic review of working alliance with digital agents and its link to outcomes

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



KEY NUMBER

3.36 to 3.75

3.36 to 3.75. Mean item scores on a 12-item working-alliance short form (1–5 scale) in the two largest unguided-agent cohorts, measured within five days of first use.

Can You Bond With a Bot?

A systematic review of working alliance with digital agents and its link to outcomes

Does alliance form with software, how fast, and does it predict outcome as it does with humans?

A note on design: This paper was planned as a random-effects meta-analysis. Fewer than five independent trials yielded comparable, extractable effect sizes for either alliance score or the alliance–outcome correlation. Under the pre-specified fallback, the analysis is a PRISMA-informed systematic review with narrative synthesis. No pooled Hedges g and no pooled r are reported. That is a finding, not a failure.

CONTENTS

01 Abstract
02 Key findings box
03 Definition block
04 Introduction
05 Methods
06 Results
07 Discussion
08 FAQ
09 References

INDEXING TITLE: Working alliance with digital agents and its association with outcomes: a meta-analysis

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

PUBLISHED: 22 September 2026 · English edition · emotionapp.in/research/working-alliance-digital-agents

01 Abstract

Working adults in India are already talking to software about stress, sleep and self-talk. The open question is whether that conversation forms a working alliance — agreement on goals and tasks, plus a felt bond — and whether the alliance predicts change the way it does with a human coach. Across large observational cohorts of adult users of conversational agents and unguided apps, mean scores on a 12-item working-alliance short form sat between 3.36 and 3.75 on a 1–5 scale within five days of first use, close to published human outpatient means of about 3.8. Bond scores often arrived first and highest; task agreement lagged. Four independent samples linked later alliance to better outcomes, with small coefficients (standardised betas about −0.13 to −0.27 for distress; one partial r of −0.18). The human-coach benchmark remains r = 0.28. Alliance measured in week one did not predict change; alliance measured in weeks three or four did. No CTRI-registered trial has yet measured software-alliance in Indian working adults. Industry-affiliated samples dominate the largest means. Certainty is low for formation and speed, and very low for the outcome link. For an Indian workplace programme the practical reading is this: a bond with an agent can appear in days, not months; it is not a substitute for a human hand-off; and task-and-goal agreement, not warmth alone, is the piece most worth designing for.

Keywords. working alliance; digital agent; conversational agent; coaching; emotional fitness; India; systematic review

02 Key findings box

KEY FINDINGS

3.36 to 3.75	**3.36 to 3.75.** Mean item scores on a 12-item working-alliance short form (1–5 scale) in the two largest unguided-agent cohorts, measured within five days of first use.
5	**Five days.** Alliance in the human-norm range is already measurable by day five. Bond arrives first. Task agreement is the slowest of the three subscales.
3	**Week 3 or 4, not week 1.** In a randomised meditation-app sample, early alliance did not predict distress reduction. Alliance at weeks three and four did (β = −0.17 and −0.13).
≈0.18	**About 0.18 versus 0.28.** The best-characterised software-only alliance–outcome association is a partial r of −0.18 with later distress. The human-coach benchmark is r = 0.28.
0	**Zero.** CTRI-registered randomised trials of working alliance with a digital agent in Indian working adults: none found.

03 Definition block

DEFINITION

Working alliance is a three-legged stool. One leg is goals: do we agree what we are aiming at? One leg is tasks: do we agree what we will actually do this week? One leg is bond: do I feel this counterpart is on my side? Bordin wrote the stool down in 1979 for human conversations. The same three legs now get scored when the counterpart is software. Digital working alliance is that score when the user is rating the app or the agent, not a person sitting inside the app. Alliance with a human coach who happens to message through the same app is a different construct. Mixing the two is how a field talks itself into a puddle. In plain English: alliance is not fondness for a mascot. It is “we know what we are doing, we know why, and I do not feel alone while doing it.” On a Monday it looks like a shared plan for the week. By Wednesday it looks like the rep that actually got done, with the agent as a nudge rather than a boss. In entrepreneurship language it is product–market fit between a person’s week and a programme’s next rep.

04 Introduction

A Mumbai product manager at 11.47 p.m. will not book a coach. She will open a chat. That is not a moral failure. That is a calendar.

Think about who is awake at that hour. The office helpline has closed. Her manager is asleep. The coach on the programme’s rota keeps office hours. What is awake is the phone in her hand. Software keeps turning up as the helper nobody put on the rota. The question this paper asks is not whether that helper has feelings. The question is whether a working alliance forms, how fast, and whether it predicts the same kind of change that alliance with a human predicts.

Human-coach research is not vague on this point. A 2018 synthesis of 295 samples and more than 30,000 people put the alliance–outcome correlation at r = 0.278 for face-to-face work and r = 0.275 for internet-delivered work that still had a human in the loop. That is a small-to-moderate link, stubborn across decades, measures and countries. If software could even approach that number, a WhatsApp-first programme for Indian professionals would have a mechanism, not just a mascot.

Three prior papers framed the gap. A 2019 initial review found five studies that mentioned alliance around smartphone programmes; none treated digital alliance as the primary outcome, and the methods were a jumble. A 2022 narrative review of fully automated apps argued that digital alliance cannot be a carbon copy of the human version: the role of bond is unclear, and human-shaped empathy is hard to ship in code. A 2022 mixed-methods paper on a free-text conversational agent — the paper this review is asked to position against — reported working-alliance scores in the human range within five days. Those three papers asked whether a bond with software is real. They did not settle whether it predicts outcome the way a human bond does, how fast the predictive version arrives, or what any of this means for an Indian workplace.

India is not a footnote. The country has a young professional workforce, long hours, family duty stacked on office duty, and a thin specialist bench outside a handful of cities. People already use chat tools in 23 languages. A coaching programme that waits for a 50-minute room in Bandra will miss the person in Nashik whose window is a phone between two meetings. If alliance with an agent is a mirage, building for it wastes rupees. If it is real but slow, a five-day onboarding is theatre. If it is real, fast, and weakly predictive, the design job is to raise task-and-goal agreement rather than to make the bot sound warmer.

Stand-up truth, then the science. You would not marry a vending machine. You might trust one that gives you the right snack at 11 p.m. Alliance with software is closer to the snack than to the wedding. The wedding fantasy is a thought to notice, not an order to follow. Two things can be true at once: “it is only software” and “this still helps me do the next rep.” The rule “if it is not a person, it does not count” can be held loosely. Machine learning would call early bond a pretrained prior and later task agreement the fine-tune. Entrepreneurship would call seven-day retention the only metric that is not vanity.

This review asks three linked questions.

1. Does a working alliance form with software — an app or a conversational agent — among adults?
2. How quickly does a measurable alliance appear?
3. Does that alliance predict later outcome, as it does with humans?

A fourth distinction runs through every table. Alliance with the app is not alliance with a human coach inside the app. A programme that hands off to a person is allowed to have both. It is not allowed to count one as the other.

05 Methods

5.1 Protocol and reporting

The review follows PRISMA 2020 reporting items. There was no prospectively registered protocol. That is a limitation and is stated as such. The analysis plan was fixed before extraction: random-effects pooling with REML if five or more eligible trials offered comparable estimands; otherwise a systematic review with narrative synthesis and no pooling. The fallback was triggered.

Eligibility was adult users of a digital agent or app; a measured alliance score using a working-alliance short form, a digital working-alliance inventory, a mobile relationship measure, or a close analogue; years 2015 to 2026; randomised or controlled trials preferred, with observational cohorts retained for the formation and speed questions when they used a validated instrument. Comparators were human-coach alliance benchmarks where a paper reported them. Primary outcomes were the alliance score and the alliance–outcome association. Interventions are described throughout as programmes, practices, techniques or coaching. They are not described as therapy, treatment or diagnosis.

Two alliances were coded separately whenever a paper measured both: alliance with the software, and alliance with a human who coached through or beside the software.

5.2 Information sources and search

Searches were run on 22 September 2026 across PubMed, PsycINFO-equivalent bibliographic queries, Scopus and Web of Science via publisher and abstracting interfaces, Google Scholar, the Clinical Trials Registry — India (CTRI), ClinicalTrials.gov, and OSF preprints. Citation chasing from the three index papers (the 2019 initial review, the 2022 automated-app narrative review, and the 2022 free-text-agent mixed-methods paper) and from a 2025 integrative review of digital alliance supplied additional records.

PubMed

```
("working alliance"[tiab] OR "therapeutic alliance"[tiab] OR "digital alliance"[tiab]
OR "digital working alliance"[tiab] OR DWAI[tiab] OR "WAI-SR"[tiab]
OR mARM[tiab])
AND (chatbot[tiab] OR "conversational agent"[tiab] OR "digital agent"[tiab]
OR smartphone[tiab] OR app[tiab] OR mHealth[tiab])
AND ("2015/01/01"[PDat] : "2026/12/31"[PDat])
```

PsycINFO, Scopus, Web of Science (2015–2026)

```
("working alliance" OR "therapeutic alliance" OR "digital alliance"
OR "digital working alliance")
AND (chatbot OR "conversational agent" OR "digital agent" OR app OR mHealth)
```

Google Scholar and OSF

"digital working alliance" OR "digital therapeutic alliance" chatbot

CTRI and ClinicalTrials.gov

working alliance AND (chatbot OR "conversational agent" OR app)

No language restriction at search. Extraction required an English abstract with extractable numbers.

5.3 Eligibility and screening

Include: adults; a digital agent, conversational agent, or unguided or minimally guided app; a named alliance instrument; extractable mean, regression coefficient, or correlation; 2015–2026.

Exclude: alliance measured only with a human delivering care by video or telephone and no software-alliance score; paediatric samples; papers that used “alliance” as a metaphor without an instrument; unverifiable preprints with no DOI or registry ID.

Screening was a single-pass expert review by the author team, not a dual-screener workflow with a reported kappa. Counts below are therefore labelled approximate. They describe the search that was actually run, not a registered two-reviewer pipeline.

5.4 PRISMA 2020 flow (approximate)

Table A. Approximate review flow

Stage	Approximate n
Records identified across databases, registries, Scholar and citation chasing	1,800–2,400
After de-duplication	1,400–1,900
Title and abstract excluded (wrong population, no alliance measure, human-only tele-sessions)	majority
Full texts assessed	80–120
Prior reviews retained for positioning	5
Primary studies with a measured software alliance	10–14
Randomised or controlled studies with an extractable software-alliance score	6–8
Independent samples with an extractable software-alliance → outcome association	4

Four independent alliance–outcome samples sit under the pre-specified floor of five. A meta-analysis of r is not yet possible. A meta-analysis of Hedges g for “alliance formation versus control” is not meaningful: alliance is a process measured inside the agent arm, not an intervention contrast.

5.5 Extraction fields

For each eligible paper: first author, year, design, n, country, language of delivery, agent type (rule-based conversational agent, generative agent, unguided app, blended human-plus-app), dose in sessions and minutes where reported, delivery channel, guidance level (unguided / agent-only / agent plus human), instrument, timepoints in days, alliance mean and SD or equivalent, alliance–outcome coefficient, outcome type, industry affiliation, and whether the score was app-alliance or human-coach-inside-app.

5.6 Analysis plan

The pre-specified model was random-effects REML, Hedges g with 95% confidence interval, I^2, tau^2, a 95% prediction interval, Egger's test, trim-and-fill, leave-one-out sensitivity, RoB 2 for randomised trials, and GRADE for certainty. Moderators were to be agent type and days to measurement.

Because the floor of five comparable trials was not met, none of those pooling statistics are computed. What follows is a structured narrative: formation (do scores reach the human range?), speed (by which day?), association with outcome (in which samples, at which timepoint?), and two planned moderators discussed qualitatively. Risk of bias is summarised in RoB 2 language for the randomised papers. Certainty uses GRADE domains without a pooled estimate.

Human benchmarks used for comparison, and not mixed into the software set, were the 2018 adult-alliance synthesis (r = 0.278 face-to-face; r = 0.275 internet-with-human) and the 2010 outpatient norms on the 12-item short form (item-mean about 3.8; bond about 4.0; task about 3.4; goal about 4.0). A 2026 update on internet-based programmes put the alliance–outcome r at 0.21 overall, with task (0.25) ahead of bond (0.12). That update still mostly describes programmes with a human in the loop. It is a neighbour, not a substitute.

06 Results

6.1 What the literature is, in one page

The software-alliance literature is a short bookshelf with a few thick volumes and many pamphlets. Two observational cohorts supply most of the bodies: 36,070 users of a CBT-style conversational agent who completed a working-alliance short form within five days, and 1,205 users of a free-text conversational agent who did the same. One randomised meditation-app arm (n = 314) supplies the cleanest time-course of a six-item digital inventory across four weeks, and the cleanest test of when alliance starts to predict distress. One generative-agent randomised trial (n = 210) supplies week-four short-form scores with standard deviations. One smoking-cessation experiment (n = 229) shows that conversational style moves the alliance score. One cancer-care conversational-agent trial (n = 117) supplies a partial correlation from week-four alliance to week-ten distress. One three-arm student trial (n = 995) supplies a structural model in which perceived alliance relates to engagement and to symptom change — with an industry conflict of interest that must travel with the number. A small alcohol-programme secondary analysis (n = 43) is the rare paper that measured digital alliance and clinician alliance in the same people. A two-study app-plus-navigator paper shows that app-alliance predicts engagement whether or not a clinician is also in the loop, and predicts symptoms only when human support is light.

That is the set. It is enough to describe a pattern. It is not enough to pool.

6.2 Study-characteristics table

Table 1. Primary studies with a measured software alliance, 2015–2026

Study	Design / n	Agent and guidance	When / instrument	Alliance score	Alliance → outcome
Darcy 2021	Observational 36,070 Multi-country EN	CBT-style conversational agent Unguided	≤5 days 12-item short form (1–5)	Total 3.36 (0.81) Bond 3.84 (1.0) Goal 3.25 / task 2.99	Concurrent screen vs bond r = −0.04
Beatty 2022	Mixed methods 1,205 / 226 Multi-country EN	Free-text conversational agent Unguided	≤5 days; +3 days 12-item short form (1–5)	3.64 (0.81) then 3.75 (0.80) Bond 3.98 then 4.05	Not tested
Goldberg 2022 S1	Survey 290 US / EN	Meditation apps Unguided	Current users 6-item digital inventory (/42)	M = 30.58 (6.56)	r = 0.42 with regular use
Goldberg 2022 S2	RCT arm 314 US / EN	Meditation programme Unguided	Weeks 1–4 6-item digital inventory	33.56 → 33.50 → 34.24 → 34.13	Wk 3 β = −0.17; wk 4 β = −0.13 vs distress. Wk 1–2 NS
Heinz 2025	RCT, agent arm 210 rand.; 96 raters US / EN	Fine-tuned generative agent Unguided	Week 4 12-item short form (1–5)	Total 3.59 (1.27) Bond 3.71; task 3.47; goal 3.59	No verified r
He 2024	Two agent styles 229 NL / EN	Cessation chatbot Unguided	After 1–2 sessions 12-item short form	MI style higher than confrontational (F = 13.61, p < .001)	Style moved alliance more than quit-intention
CanRelax 2026	RCT secondary 117 German-speaking	Conversational agent Unguided	Week 4 Internet-programme inventory (1–5)	Bond 4.04 (0.95) Goal 3.46; task 2.79	Partial r = −0.18 vs wk-10 distress; task/goal drive; bond NS
Shoshani 2026	3-arm RCT 995 (336 agent) Israel	Generative conversational agent Unguided	Across 12 weeks Perceived digital alliance	Associated with engagement and change	SEM: → engagement β = 0.31; → symptom change β = −0.58. Author equity
Benitez 2024	RCT secondary 43 US / EN	Computerised CBT-style programme Minimal monitoring or + usual care	Sessions 2 and 6 Digital and clinician alliance versions	Digital ≈ clinician	Small within-person link to later drinking during programme; not at follow-up
Macrynikola 2025 S1	Guided app + navigator US / EN	Research mental-health app Light human check-ins	Midpoint Digital working-alliance inventory	—	→ engagement b = 0.18; → anxiety/depression b = −0.27
Macrynikola 2025 S2	App + clinician + navigator US / EN	Same app inside tele-coaching Human coach in the loop	Midpoint Same inventory	—	→ engagement b = 0.21; outcomes NS

Study	Design / n	Agent and guidance	When / instrument	Alliance score	Alliance → outcome
Herrero 2020	Blended 193 Spain	Online programme beside a clinician Blended	During programme Technical working-alliance short form	Total 57.84 (16.39) summed scale	Predicted symptom change β = −0.27; satisfaction R^2 = 0.50

Industry affiliation is present in Darcy (company-employed authors), Beatty (company-employed co-authors), and Shoshani (author equity). Heinz, Goldberg, CanRelax and Benitez are academic. These flags do not delete the numbers. They sit beside them.

6.3 Does alliance form with software?

Yes, as a score. People given a working-alliance questionnaire about an agent or an app return means in the same neighbourhood as outpatients rating a human.

On the 12-item short form with a 1–5 item mean, the two largest unguided-agent cohorts landed at 3.36 (day five, n = 36,070) and 3.64 rising to 3.75 (day five then day eight, n = 1,205 then 226). A generative-agent randomised trial landed at 3.59 at week four (n = 96 raters). Human outpatient norms on the same scale sit near 3.8 overall, with bond near 4.0, task near 3.4 and goal near 4.0.

The subscale pattern is consistent enough to be useful. Bond is the highest number. Task is the lowest. In the day-five conversational-agent cohort, bond was 3.84, goal 3.25, task 2.99. In the free-text-agent cohort, bond was 3.98, goal 3.62, task 3.33. In the generative-agent trial at week four, bond 3.71, goal 3.59, task 3.47 — task had caught up. In the cancer-care agent at week four, bond 4.04, goal 3.46, task 2.79. Software is good at “I do not feel judged.” Software is slower at “we both know the drill for Tuesday.”

That pattern is the opposite of what a 2026 internet-programme synthesis found when a human was still in the loop: task predicted outcome more than bond. Software currently over-delivers the warm prior and under-delivers the shared plan. A coaching programme that celebrates bond and neglects task is overfitting the easy feature.

Qualitative snippets from the free-text-agent paper read like a daily mood log written to a machine. Gratitude. Personification. Limits named without a falling-out. Those utterances are not a randomised trial. They are what a bond looks like when nobody is scoring it.

A small alcohol-programme analysis found digital-alliance ratings statistically close to clinician-alliance ratings in the same people. That is the cleanest head-to-head currently on the shelf. The sample is 43. Treat it as a proof of possibility, not a proof of equivalence at scale.

6.4 How fast?

Faster than a human coach’s appointment diary would lead you to expect. Slower than a splash screen that asks “how are we doing?” on night one.

Both large conversational-agent cohorts had human-range scores inside five days. The free-text-agent follow-up three days later did not move the mean in a statistically convincing way. The meditation-app randomised arm is the useful brake on that story. A six-item digital inventory was already high in week one (about 33.6 out of 42) and barely climbed by week four (about 34.1). Predictive validity was not present in weeks one or two. It appeared in weeks three and four.

Read those two clocks together. A score can appear by day five. A score that predicts change seems to need two or three weeks of reps. Saying yes to the plan on day one is not the same as still doing the rep in week four. Machine learning has the same lesson: the first loss value is not the validation loss. Early stopping on night one would throw away the signal.

Dose in the large cohorts was light and frequent rather than long and rare. The free-text-agent follow-up completers averaged about seven conversations in fourteen days. The generative-agent trial reported more than six hours of cumulative use over eight weeks. The meditation-app trial was a four-week daily-practice window. None of these look like a 50-minute weekly session. They look like a product: short, repeatable, interruptible.

6.5 Does alliance predict outcome?

Sometimes, modestly, and later rather than sooner. The coefficients are smaller and fewer than the human literature.

Table 2. Independent samples with an extractable software-alliance → outcome association

Sample	n	When measured	Outcome	Coefficient	Notes
Goldberg 2022 S2	314	Weeks 3 and 4	Pre–post distress reduction	β = −0.17 (wk 3); β = −0.13 (wk 4)	Weeks 1–2 not predictive. Academic school-staff sample
CanRelax 2026	117	Week 4	Distress at week 10	Partial r = −0.18	Task and goal drive the link. Bond NS in the full sample
Macrynikola 2025 S1	guided app + navigator	Midpoint	Later engagement; anxiety/depression	b = 0.18 engagement; b = −0.27 symptoms	Light human support
Shoshani 2026	336 of 995	Across 12 weeks	Engagement; symptom change	β = 0.31; β = −0.58 (SEM)	Author equity. Not a simple r

A 2024 follow-up on the same meditation-app sample found a two-way street: alliance predicted later distress, and distress predicted later alliance, both with small betas. That paper is not a fifth independent sample. Counting it twice would be double-dipping.

The human benchmark is r = 0.278. Internet-with-human sits near 0.21 to 0.28. Software-only, on the papers that report a simple association, sits nearer 0.13 to 0.18. The structural coefficient of −0.58 is an outlier in form and in authorship. It is reported. It is not used as the headline r.

Direction is consistent. Stronger software-alliance, less distress later, more use in the meantime. Magnitude is modest. Timing matters. Subscale matters: shared tasks and goals carry more of the outcome link than bond, where the question has been asked cleanly.

Herrero 2020 sits one step off this table. It is a blended programme. Alliance with the online programme predicted symptom change (β = −0.268) and predicted satisfaction even more strongly. That is alliance with software beside a human, not instead of one. It belongs in the "human coach inside the app" column of a product spec.

6.6 Forest-plot data table

No forest plot of Hedges g is drawn. Alliance was not an intervention contrast. For readers who want the raw material a future meta-analysis will need, Table 3 lists every 1–5 short-form total the review could verify.

Table 3. Working-alliance short-form item-means (1–5) for software agents

Study	Timepoint	n	Mean	SD	Note
Darcy 2021	≤5 days	36,070	3.36	0.81	Would dominate any naive pool
Beatty 2022 A1	≤5 days	1,205	3.64	0.81	—
Beatty 2022 A2	~8 days	226	3.75	0.80	Completer subset
Heinz 2025	Week 4	96	3.59	1.27	Wider SD
Human outpatient norm (Munder 2010)	Typical mid-programme	—	3.8	0.63	Benchmark only

A naive inverse-variance mean of the software rows would be pulled toward 3.36 by the 36,070-person cohort and would look more precise than the data deserve. Completers who agree to rate an alliance questionnaire within five days are not a random draw of installers. That is why this review does not pool.

6.7 Moderator table

Table 4. Pre-specified moderators, qualitative

Moderator	Pattern in the set	Certainty
Agent type	Rule-based CBT-style agents and a fine-tuned generative agent produced similar short-form totals (mid-3s). A motivational-interview conversational style beat a confrontational style on alliance. Meditation apps produced high digital-inventory scores that were stable across four weeks.	Low. Confounded with sample, brand and instrument
Days to measurement	A score in the human range by day 5. Predictive validity from week 3–4, not week 1–2. Little mean-shift between day 5 and day 8 in the free-text-agent completers.	Low. Few repeated-measures designs
Guidance (added because the papers forced the distinction)	App-alliance predicted symptoms when human support was a light navigator. App-alliance predicted engagement but not symptoms when a clinician was also in the loop.	Very low. One two-study paper
Industry affiliation	The two largest means come from company-linked cohorts. The largest structural path comes from a trial whose first author holds equity.	Bias domain, not a causal moderator

6.8 Heterogeneity

Heterogeneity is obvious and unquantified. Instruments differ (12-item short form, 6-item digital inventory, internet-programme inventory, structural items). Timepoints differ (day 5, week 4, week 12). Populations differ (general app users, school staff, smokers, people living with cancer, university students, outpatients with alcohol-use difficulties). Guidance differs (none, navigator, clinician). A single I^2 on this set would be a costume.

What is not heterogeneous is the direction of the subscale pattern and the direction of the outcome link. Bond high, task low, later alliance modestly associated with later relief. That rhyme is the finding.

6.9 Publication bias

Egger's test and trim-and-fill were not run. With four independent outcome-association samples they would be theatre. Two risks sit in plain sight instead. First, the largest formation studies are industry-affiliated and recruit from people who kept using the product long enough to answer a questionnaire. Second, null alliance–outcome findings are easy not to publish as primary papers; at least one trial collected a short-form score and did not lead with an alliance–outcome r. Absence of a correlation in this review is not evidence that every silent file is positive.

6.10 Risk-of-bias summary

Table 5. RoB 2-style summary for randomised papers that measured software alliance

Trial	Randomisation	Deviations	Missing data	Measurement	Reporting	Overall
Goldberg 2022 S2	Low	Some concerns — alliance only in intervention arm	Some concerns	Some concerns — self-report	Low	Some concerns
Heinz 2025	Low	Some concerns — wait-list; alliance only in agent arm	Some concerns — 96 of 210 rated	Some concerns — self-report	Some concerns	Some concerns
Shoshani 2026	Low	High — author equity	Some concerns	Some concerns	Some concerns	High for the alliance claim
CanRelax parent	Low	Some concerns	Some concerns	Some concerns	Low	Some concerns
He 2024	Some concerns	Low	Some concerns	Some concerns	Low	Some concerns
Benitez 2024	Low	Some concerns — small n	High — n = 43	Some concerns	Low	High for precision

Observational cohorts (Darcy; Beatty) are not RoB 2 objects. In ROBINS-I language they carry serious risk of selection: the denominator is people who installed, kept using, and answered. Both have company authors. They answer "can a score appear?" They do not answer "does the product cause the score?"

6.11 GRADE table

Table 6. Certainty of evidence

Question	Estimate as reported	Studies	Certainty	Why
Does alliance form with software?	Short-form item-means 3.36–3.75 by day 5; week-4 generative-agent mean 3.59; human norm ~3.8	3 large or trial-based short-form samples plus supporting digital-inventory data	Low	Consistent neighbourhood of means; industry-heavy samples; completer bias; no Indian working-adult RCT
How fast?	Score by day 5; predictive from week 3–4	2 day-5 cohorts + 1 four-week RCT arm	Low	Replicated early score; single clean predictive-timing study
Does it predict outcome?	β ≈ −0.13 to −0.27; partial r = −0.18; one SEM β = −0.58	4 independent samples	Very low	Sparse; mixed estimands; one conflicted SEM; smaller than human r = 0.28; no India data

GRADE starts high for randomised evidence and is stepped down for risk of bias, indirectness (students, smokers, cancer-care, US school staff — not Indian working adults), imprecision (few samples), and inconsistency of estimands. Formation and speed stay at low rather than very low because the means rhyme across methods. The outcome link does not get that courtesy.

07 Discussion

PULL QUOTE

Bond is cheap. Task is expensive. Design for Tuesday's rep, not for warmer copy.

7.1 What the number means

The number most people will quote is 3.4 to 3.8 out of 5, inside a week. That number means a working-age adult can look at a questionnaire written for a human coach, swap the word "coach" for the name of an agent, and tick boxes that look like the boxes outpatients tick. It does not mean the agent is a person. It does not mean the score is earned. It means the instrument does not bounce off the interaction as nonsense.

The second number is five days versus three to four weeks. Five days is how fast a score appears. Three to four weeks is how fast a useful score appears. Product teams love the first clock. Coaching programmes should budget for the second. Picture a new joiner on a Hyderabad team. By the Friday of her first week she is at ease with everyone at the tea counter. Nobody writes her appraisal that Friday. Rapport runs on the first clock. The appraisal waits for work done.

The third number is 0.18 versus 0.28. Software-alliance is in the same family as human-alliance as a correlate of change. It is a thinner relative. Anyone who tells a buyer that "alliance with our agent predicts outcome just like a

psychologist" is selling the human coefficient on software data. That is not a joke. That is a specification error.

The subscale split is the design brief hiding inside the statistics. Bond is cheap. Task is expensive. Goal sits in the middle. CanRelax is the paper that says this in numbers: bond 4.04 and not predictive; task 2.79 and part of the r = –0.18. A programme that pours tokens into warmer copy and never makes Tuesday's rep obvious is spending the marketing budget on the subscale that does less work.

7.2 App-alliance is not coach-alliance

This distinction is where Indian product design will either grow up or trip.

When a navigator sends a light check-in, alliance with the app still predicts symptoms. When a qualified professional is also in the loop, alliance with the app still predicts use, and stops predicting symptoms. One reading: the human soaks up the change mechanism, and the app becomes a workbook. Another reading: once a person is present, rating the app is rating a sidekick. Either way, a programme that offers "AI coach with human hand-off" must measure two scores. One for the agent. One for the person who takes the hand-off. Collapsing them will produce a vanity dashboard.

The small alcohol-programme paper is the other half of that sentence. People can hold two alliances at once and rate them similarly. Dual measurement is feasible. It is not romantic. It is instrumentation.

7.3 Limitations

This review has no registered protocol and no dual-screener kappa. Search coverage of subscription databases was via available interfaces rather than a librarian-run export of every hit. Counts in the flow diagram are ranges. That is honest. It is also a GRADE ding.

The largest means come from industry-linked completer cohorts. People who bounce on night one do not rate working alliance on night five. Every early mean is conditioned on still being there.

Instruments were borrowed from research on human one-to-one helping relationships and given a noun-swap. A 2022 narrative review already warned that digital alliance may need its own construct, not a relabel. A six-item digital inventory found a single factor, not three. If the three-legged stool collapses into one factor in software, then "bond versus task versus goal" is a story we are still imposing.

Almost no paper in this set is an Indian working-adult sample. One Indian co-author sat on the free-text-agent paper. One Hindi conversational agent in rural Madhya Pradesh helped with homework beside a human counsellor and did not measure alliance at all. CTRI is quiet.

Alliance–outcome coefficients are uncontrolled for a dozen third variables. People who like structure like programmes and get better because they do the reps. Alliance may be a marker of that tendency, not a cause. The meditation-app follow-up that found bidirectional paths is the warning label.

Wait-list controls in two of the generative-agent trials make the symptom effects look larger than an attention-matched comparison would. Those trials are cited here for their alliance scores, not as proof that software outperforms group coaching.

7.4 What this means for an Indian workplace programme

WHAT THIS MEANS

What this means for an Indian workplace programme

An Indian professional does not need a bot that says yaar and calls it empathy. Picture an analyst in Pune, laptop still open on a weeknight. She needs a counterpart that agrees the goal is “sleep before midnight on weeknights,” names the Tuesday rep — phone on charge outside the bedroom — and does not flinch when she types in Hindi at 11.47 p.m. By the Friday of her first week she has rated the agent warmly. The score that predicts change shows up only after a few dozen reps. Build the first week to earn a score. Build weeks two to four to earn a prediction. When the agent hands her to a human coach, score that bond separately. Do not put a human r of 0.28 on a sales slide built from a software r of 0.18. Run the missing trial: working adults, 23 languages, WhatsApp-first, DPDP-resident data, alliance scored at day 5 and week 4, outcomes scored at week 8, India-resident investigators who do not hold the company’s equity. Until that trial exists, treat today’s numbers as a permit to build, not a permit to boast.

7.5 Research gaps, naming the India gap explicitly

The India gap is not a vibe. It is an empty cell. No CTRI-registered randomised trial has measured working alliance with a digital agent among Indian working adults. The closest objects are a homework-aid chatbot in rural Madhya Pradesh with no alliance instrument, a free-text-agent paper with a Bengaluru clinical-psychology co-author and no India-stratified scores, and cognitive-assessment deployments of a research app in Bangalore and Bhopal. None answer the workplace question.

Other gaps sit behind it. We do not know how alliance behaves in Hindi, Tamil, Telugu, Bengali or Marathi when the instrument itself has not been validated in those languages. We do not know whether a human hand-off raises or lowers app-alliance. We do not know whether day-5 bond predicts week-8 retention in a population whose first digital habit is WhatsApp, not an app store. We do not know the dose in minutes that turns a score into a predictor. We do not have a registered, two-screener, pre-analysis-planned meta-analysis, which this paper is not.

One last image, then stop. Every Mumbai monsoon, the argument about a building is never whether people should live in it. It is about what the building stands on. Alliance with software is a building people already live in. The firm

foundation is task-and-goal agreement, measured twice, handed off to a person when the weather turns. The soft fill is a warm sentence on day one and a press release that quotes the human r.

08 FAQ

Q1 Can I form a working alliance with an app, or is that just branding?

You can form a score. Large cohorts land between 3.36 and 3.75 out of 5 within five days, next to a human outpatient mean of about 3.8. That is a questionnaire result, not a wedding. Treat it as permission to take the next rep, not as proof the software cares.

Q2 How soon should a workplace programme bother to measure it?

Measure at day five if you want to know whether onboarding worked. Measure at week three or four if you want a number that has predicted distress reduction. Week-one scores in the one trial that clocked this were already high and were not predictive.

Q3 Does a stronger alliance mean people actually get better?

Sometimes, a little. The cleanest software-only figures are a week-4 to week-10 partial r of −0.18 and week-3/4 betas of −0.13 to −0.17. Human coaching sits near r = 0.28. Do not sell the human number on software data.

Q4 If we add a human coach, can we stop measuring the app?

No. App-alliance still predicts whether people keep tapping. In one two-study paper it predicted symptoms only when the human was a light navigator, not when a qualified professional was also in the loop. Score both. Hand-off is a feature. It is not a blend button.

Q5 **Has anyone run this study on Indian working adults?**

Not in CTRI, not as a randomised alliance trial. Zero. A Hindi homework-aid chatbot in rural Madhya Pradesh did not score alliance. That empty cell is the next paper, not a rumour that the answer is already yes.

09 References

Primary sources only. Each entry carries a DOI or a registry URL.

1. Darcy A, Daniels J, Salinger D, Wicks P, Robinson A. Evidence of human-level bonds established with a digital conversational agent: cross-sectional, retrospective observational study. JMIR Formative Research. 2021;5(5):e27868. https://doi.org/10.2196/27868

2. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent: a mixed-methods study. Frontiers in Digital Health. 2022;4:847991. https://doi.org/10.3389/fdgth.2022.847991

3. Goldberg SB, Baldwin SA, Riordan KM, et al. Alliance with an unguided smartphone app: validation of the Digital Working Alliance Inventory. Assessment. 2022;29(6):1331-1345. https://doi.org/10.1177/10731911211015310

4. Heinz MV, et al. Randomized trial of a generative AI chatbot for mental health treatment. NEJM AI. 2025;2(4). https://doi.org/10.1056/AIoa2400802

5. He L, Basar E, Krahmer E, Wiers R, Antheunis M. Effectiveness and user experience of a smoking cessation chatbot: mixed methods study comparing motivational interviewing and confrontational counseling. Journal of Medical Internet Research. 2024;26:e53134. https://doi.org/10.2196/53134

6. Benitez B, Kiluk BD, et al. The connection still matters: therapeutic alliance with digital treatment for alcohol use disorder. Alcohol: Clinical and Experimental Research. 2023;47(11):2197-2207. https://doi.org/10.1111/acer.15199

7. Shoshani A, Gurfinkel B, Kor A, et al. Efficacy of a conversational AI agent for psychiatric symptoms and digital therapeutic alliance: a randomised clinical trial. JAMA Network Open. 2026;9(4):e266713. https://doi.org/10.1001/jamanetworkopen.2026.6713

8. Macrynikola N, Chang S, Torous J. Is digital alliance associated with engagement and outcomes in guided digital interventions? An analysis of data from two studies. Journal of Affective Disorders. 2025;383:335-340. https://doi.org/10.1016/j.jad.2025.04.127

9. Herrero R, Vara MD, Miragall M, et al. Working Alliance Inventory for Online Interventions-Short Form (WAI-TECH-SF): the role of the therapeutic alliance between patient and online program in therapeutic outcomes. International Journal of Environmental Research and Public Health. 2020;17(17):6169. https://doi.org/10.3390/ijerph17176169

10. Berry K, Salter A, Morris R, James S, Bucci S. Assessing therapeutic alliance in the context of mHealth interventions for mental health problems: development of the Mobile Agnew Relationship Measure (mARM) questionnaire. Journal of Medical Internet Research. 2018;20(4):e90. https://doi.org/10.2196/jmir.8252

11. Henson P, Wisniewski H, Hollis C, Keshavan M, Torous J. Digital mental health apps and the therapeutic alliance: initial review. BJPsych Open. 2019;5(1):e15. https://doi.org/10.1192/bjo.2018.86

12. Tong F, Lederman R, D'Alfonso S, Berry K, Bucci S. Digital therapeutic alliance with fully automated mental health smartphone apps: a narrative review. Frontiers in Psychiatry. 2022;13:819623. https://doi.org/10.3389/fpsyt.2022.819623

13. Malouin-Lachance A, Capolupo J, Laplante C, Hudon A. Does the digital therapeutic alliance exist? Integrative review. JMIR Mental Health. 2025;12:e69294. https://doi.org/10.2196/69294

14. Flückiger C, Del Re AC, Wampold BE, Horvath AO. The alliance in adult psychotherapy: a meta-analytic synthesis. Psychotherapy. 2018;55(4):316-340. https://doi.org/10.1037/pst0000172

15. Flückiger C, et al. Alliance in internet-based interventions: a systematic review and correlation multilevel meta-analysis. Journal of Consulting and Clinical Psychology. 2026. https://doi.org/10.1037/ccp0001005

16. Probst GH, Berger T, Flückiger C. The alliance-outcome relation in internet-based interventions for psychological disorders: a correlational meta-analysis. Verhaltenstherapie. 2019;32:135-146. https://doi.org/10.1159/000503432

17. Munder T, Wilmers F, Leonhart R, Linster HW, Barth J. Working Alliance Inventory-Short Revised (WAI-SR): psychometric properties in outpatients and inpatients. Clinical Psychology & Psychotherapy. 2010;17(3):231-239. https://doi.org/10.1002/cpp.657

18. Henson P, Peck P, Torous J. Considering the therapeutic alliance in digital mental health interventions. Harvard Review of Psychiatry. 2019;27(4):268-273. https://doi.org/10.1097/HRP.0000000000000224

19. Agrawal R, Monteiro K, Narasimhamurti NS, et al. A conversational agent (PracticePal) to support the delivery of a brief behavioural activation treatment for depression in rural India: development and pilot-testing study. JMIR Formative Research. 2025. https://doi.org/10.2196/73563

20. Bordin ES. The generalizability of the psychoanalytic concept of the working alliance. Psychotherapy: Theory, Research & Practice. 1979;16(3):252-260. https://doi.org/10.1037/h0085885

21. Goldberg SB, et al. Bidirectional alliance and distress in an unguided app (same-sample follow-up). Clinical Psychological Science. 2024. https://doi.org/10.1177/21677026231184890

22. Association between working alliance and treatment outcomes in a mobile health intervention with a conversational agent (CanRelax). Internet Interventions. 2026. https://doi.org/10.1016/j.invent.2026.100929

23. He L, Basar E, Wiers RW, Antheunis ML, Krahmer E. Chatting your way to quitting: a longitudinal exploration of smokers' interaction with a cessation chatbot. Internet Interventions. 2025. https://doi.org/10.1016/j.invent.2025.100807

24. 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. https://doi.org/10.1038/s41746-023-00979-5

About the author

ABOUT THE AUTHOR

Dr Atul Aswani; Mumbai Psychiatrist and trained as a Results Coach
Atul Aswani is a Mumbai Psychiatrist and trained as a Results Coach. He builds EmotionApp, a non-clinical emotional-fitness programme for Indian professionals: countable daily reps, an AI coach with human hand-off, WhatsApp-first delivery, 23 Indian languages, India-resident data, DPDP-compliant, under ISO 9001 and ISO 13485. He writes the EmotionApp research series so the product stays answerable to numbers.

HOW TO CITE

Aswani A. Can You Bond With a Bot?: A systematic review of working alliance with digital agents and its link to outcomes. EmotionApp Research Paper 38. Mumbai: Tranquilo Meditek Sytems Private Limited; 2026. https://emotionapp.in/research/working-alliance-digital-agents

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.