Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From "Crutch" to Coach: Patterns of Sustained Engagement and Deepening Support in AI Wellbeing Coaching

Max Major

max.major@unmind.com

Unmind Ltd

Viktoria M Ivan

Unmind Ltd

Katie M White

Unmind Ltd

Marcos Economides

Unmind Ltd

Article

Keywords:

Posted Date: March 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8949770/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. This study was funded in full by Unmind Ltd, the creator of the AI coach evaluated. Authors MM, VMI, and KMW are employed by and own share options at Unmind Ltd. ME was employed by Unmind Ltd as a paid contractor at the time the study was conducted and owns share options.

Abstract

Background: As use of on-demand, user-led AI chatbots for mental health support grows, a central question is whether these systems provide only short-term assistance or can foster sustained, coaching-style relationships designed to enable lasting change and wellbeing improvements. We examined this distinction in Nova, an AI wellbeing coach.

Methods: We analysed 14,293 sessions between January and August 2025. Session-to-session continuity was assessed with large language model (LLM) classification of whether returning users revisited prior issues or initiated new topics. Emotional engagement depth was measured using the Experiencing Scale (EXP), with ratings generated by an LLM calibrated against expert psychotherapists. Sessions were classified by topic using an LLM applied to a taxonomy developed by clinical psychologists.

Results: Returning users continued prior work in 70.8% of sessions, indicating sustained engagement rather than episodic support. Regular users (≥5 sessions) achieved deep emotional engagement (EXP ≥5) at more than twice the rate of single-session users (6.5% vs. 2.9%), though users with initially lower engagement scores were more likely to return. Continuity varied by topic: wellbeing- and work-related challenges were frequently revisited, whereas more complex concerns were typically addressed in single sessions, reflecting appropriate boundaries. Calibration of EXP ratings against expert raters yielded moderate to good reliability (ICC = 0.62-0.79).

Conclusions: Users not only seek but also develop sustained, coaching-style relationships with AI, within appropriate limits of scope. The convergence of topic continuity and deepening emotional engagement suggests that AI can support iterative coaching processes rather than providing only short-term relief or information. Our approach offers a framework for assessing relationship quality in AI tools for wellbeing.

Highlights

- Users showed 70% topic continuity across sessions, indicating sustained coaching relationships rather than episodic support.
- Rate of deep emotional engagement doubled for regular users (≥5 sessions) versus single sessions.
- Demonstrated scalable LLM methodology for measuring relationship quality in AI wellbeing coaching.
- AI wellbeing coach demonstrated capacity for facilitating emotional exploration within appropriate boundaries.

1. Introduction

The deployment of artificial intelligence (AI) applications in mental health and wellbeing has raised fundamental questions about the nature and quality of human-AI therapeutic relationships. Whilst AI chatbots are increasingly used to address demand for mental health support amid limited practitioner capacity [1–4], a distinction can be drawn between two models of AI-mediated support: episodic assistance versus sustained coaching relationships.

Episodic support typically involves brief, on-demand interactions that provide immediate assistance, information or emotional support for specific concerns. In contrast, coaching-style relationships are characterised by sustained engagement over multiple sessions, continuity of discussion topics and progressive deepening of emotional exploration The latter features traditionally associated with relationships that facilitate lasting behavioural and psychological change [5,6].

Recent evidence suggests that AI chatbots can deliver therapeutic communication of comparable quality to human therapists, with mental health professionals unable to reliably distinguish between human-AI and human-human therapeutic conversations [7]. However, whilst engagement with digital mental health interventions (DMHIs) has been studied [8,9], questions remain about whether users develop sustained, coaching-style relationships with AI systems. High attrition rates affect many DMHIs [9] and conversational agent-delivered interventions show variable retention patterns [10]. Understanding these patterns requires attention to multiple influences; for instance, theoretical frameworks like the Digital Cumulative Complexity Model highlight several factors influencing sustained engagement, including intervention design, user characteristics and contextual complexity [11].

Despite these theoretical advances, empirical evidence characterising the specific patterns of how users engage with coaching-oriented AI systems remains limited. Existing research has established that AI can deliver mental health support [4] and that some users engage positively with these tools. However, three gaps in our understanding of AI coaching relationships persist. First, whilst users sometimes return to AI systems, whether they return to continue working on the same concerns (reflective of coaching relationships) or primarily seek episodic help for new issues is unclear. Second, although therapeutic quality can be assessed in single interactions, evidence on whether emotional engagement depth increases over time with sustained use – a hallmark of deepening coaching and therapeutic relationships – is lacking. Third, no studies have combined session-to-session continuity tracking with progressive emotional depth measurement at scale in real-world settings.

The distinction between episodic support and sustained coaching relationships has implications for both the design and evaluation of AI wellbeing tools. Research on human-facilitated therapy indicates that deeper emotional processing and sustained engagement with personal concerns predict better outcomes [5]. Building on this foundation, longitudinal research has demonstrated that increased emotional processing over time predicts depression symptom reduction[12], supporting the relevance of tracking emotional engagement depth in therapeutic relationships. Recent evidence in adolescent therapy demonstrates that working Experiencing Scale (EXP) score (measured at session 6) accounted

for 11.6% of variance in psychological distress outcomes, whilst early EXP (session 1) was not predictive [13], suggesting that progression in emotional engagement depth over time may be particularly relevant to therapeutic benefit. The Experiencing Scale (EXP), a validated measure of emotional engagement depth originally developed for assessing human therapeutic conversations [14], provides a framework for quantifying this dimension of interaction quality. Higher EXP scores reflect richer self-reflection, emotional connection and meaning-making; qualities that distinguish surface-level information exchange from the kind of personal exploration associated with psychological growth and wellbeing improvement.

Drawing on this evidence, Nova was designed to support sustained, coaching-style engagement – including topic continuity across sessions and progressively deeper emotional exploration. The present study examined whether these design objectives are realised in practice, using real-world conversational data from workplace deployments. We investigated two complementary dimensions of coaching-style relationships: (1) session-to-session continuity, measured by whether returning users revisited previously discussed concerns or initiated entirely new topics; and (2) emotional engagement depth, assessed using LLM-generated EXP ratings calibrated against expert psychotherapists.

Our analysis addressed two primary research questions: First, do users demonstrate sustained engagement with specific concerns across multiple sessions or do they primarily seek episodic support for discrete issues? Second, does emotional engagement deepen with repeated use, suggesting progressive rather than static coaching relationships? We also explored the relationship between topic continuity and progression in emotional engagement depth and how engagement patterns varied across concern types (workplace, personal relationships, health and wellbeing).

Our methodological approach, combining LLM classification with expert clinical validation, offers a scalable framework for assessing relationship quality in AI wellbeing tools.

2. Methods

2.1 Study Design and Setting

This was a retrospective observational study of conversational data from Nova, an AI wellbeing coach deployed in workplace settings. The study period ran from 17th January 2025 to 8th August 2025.

2.2 Nova AI Wellbeing Coach

Nova is a Generative AI chatbot available to employees (aged 18 and over) of clients that provide access to Unmind products and services. Nova sits within Unmind’s broader workplace wellbeing platform, which includes self-guided content, structured programmes, and access to human therapy and coaching. Within this ecosystem, Nova serves as a first point of engagement, providing on-demand wellbeing coaching and connecting users to other resources and services where appropriate. Nova was powered by OpenAI's GPT-4o foundation model during the study period. Users initiate on-demand, self-paced

conversations via text interface to explore topics within the intended purpose for Nova, such as workplace challenges, personal wellbeing concerns and life stressors.

Upon first use of Nova, the user is presented with an Onboarding Conversation- an optional first-time conversation that helps Nova get to know the user with a few short questions about their life, work, and current wellbeing. That information is used to create a simple user summary and personalise Nova's future responses. Once enough information is gathered during this Onboarding Conversation, the flow ends and Nova switches into the normal coaching experience.

Consistent with its intended use, Nova is not a mental health treatment or therapeutic intervention. Strict technical and conversational guardrails maintain this scope by detecting and appropriately routing discussions that exceed wellbeing coaching boundaries, including crisis situations, acute mental health symptoms requiring clinical intervention and medical concerns. Nova maintains memory of previous conversations to support continuity across conversations and tailors responses based on user input within each session.

2.3 Data Source and Session Selection

A Nova session is defined as a conversation between the user and Nova, ending after an hour of inactivity. From 41,784 sessions representing a subset of conversations from client organisations during the study period, we excluded transactional sessions (n = 8,420; user requests for information, resources, or non-wellbeing queries, identified via automated session-type labels assigned by the Nova system) and sessions with insufficient conversational exchange for depth assessment (n = 19,071; defined as single user message or minimal exchanges), yielding a final analytical sample of 14,293 sessions.

2.4 Topic Continuity & Return Purpose Analysis

This set of analyses examined whether users revisited prior issues when returning to Nova, and, when they did, the nature of that return. From the 14,293 sessions in the analytical sample, there were 4,113 non-initial, non-trivial sessions eligible for topic-continuation and return-purpose analyses.

All classifications were generated using a large language model (LLM; OpenAI gpt-4o-mini; temperature = 0.3) with structured-output formatting. Transcripts were truncated at 3,750 tokens to remain within the model's context window whilst preserving sufficient conversational content. The model received the current transcript plus information from the previous session. A random subsample was reviewed by research staff and prompts were refined iteratively until classifications aligned with manual review.

2.4.1 Topic Continuity Classification

To assess whether users returned to previously discussed issues, the model was given the prior session summary and the current transcript and assigned one of four outcomes: Continuation, New topic,

Unclear, or Not applicable (if no previous session summary was available). Only sessions classified as continuations were included in return-purpose analysis.

2.4.2 Return Purpose Classification

For sessions identified as continuations, the model compared the full transcripts of the previous and current sessions and classified the purpose of the return as one of four categories:

- **Progress-Deepening:** explicit signs of moving forward or meaningful re-engagement.
- **Unresolved-Stuck:** limited progress or repeated difficulty with the same issue.
- **Mixed:** features of both movement and difficulty.
- **Unclear:** insufficient or ambiguous evidence.

2.5 Emotional Engagement Depth Measurement

Emotional engagement depth was assessed using the EXP Scale [14], a 7-point observational measure of moment-to-moment emotional involvement developed for psychotherapy research, ranging from impersonal discussion to deep exploration of felt experience. Full scoring criteria are available in [14] and at http://www.experiential-researchers.org. EXP ratings were generated by a LLM (OpenAI gpt-4o; temperature = 0.3), calibrated against expert psychotherapists as described below. We measured the peak EXP level achieved within each conversation rather than an average score across all exchanges per user, as peak experiencing is a stronger predictor of positive outcomes in human therapy research [6].

EXP Scoring Procedure:

EXP scores were developed using a calibration set of 50 transcripts drawn from Nova conversations between 22nd May and 7th August 2025. Eligible sessions were filtered using the same exclusion criteria as the analytical dataset. To ensure coverage across the full scale, transcripts were selected using stratified sampling targeting up to six examples per EXP score, supplemented with random sampling to reach the final set. Transcripts were presented in random order and raters were not informed of the stratification to prevent expectation and anchoring biases.

Two independent psychologists first scored the calibration set using detailed EXP guidance adapted from published manuals. Their ratings were used to iteratively refine the LLM scoring prompt until agreement metrics stabilised. A third psychologist then scored the same transcripts as a holdout set to check that prompt refinement had not overfitted to the initial raters.

Inter-rater agreement was assessed using intraclass correlation coefficients (ICCs), computed with the pingouin library for Python 3. Ninety-five percent confidence intervals were estimated for all ICCs and interpretation followed standard benchmarks. Pairwise ICCs between each human rater and the LLM were examined, along with overall agreement estimates with and without inclusion of the LLM.

2.6 Conversation Classification

An independent LLM-based taxonomy system (developed by Unmind's clinical team for classifying Nova conversations) classified conversations into high-level themes (Health & Wellbeing, Life Context, Work) and low-level themes (Worry-Related Concerns, Work Stress, Personal Relationships, Sleep, etc). The system identified up to five themes per conversation in order of prominence. For stratified analyses, the primary (most prominent) theme was used.

2.7 User Segmentation

Users were categorised by session frequency:

- **Single-session users:** 1 conversation
- **Occasional users:** 2-4 conversations
- **Regular users:** ≥5 conversations

For a sensitivity analysis, conversations were also categorised by type: Welcome Conversations (introductory onboarding sessions) versus Coaching Conversations (substantive wellbeing discussions). This distinction is determined by the user's entry point into the system – selecting a welcome interaction triggers a dedicated onboarding prompt, while all other sessions use the coaching prompt.

2.8 Statistical Analysis

Proportions of returning users who revisited previous topics versus initiated new topics were calculated and stratified by session number, first session type, conversation theme and initial EXP score. For EXP scores, we calculated median values, standard deviations, and proportions reaching two predetermined thresholds: EXP ≥ 3, termed moderate engagement (personally meaningful, emotionally connected content, representing the point at which users begin engaging with internal experience rather than remaining at surface-level interaction) and EXP ≥ 5, termed deep engagement (vivid, immediate experiencing, representing active exploration of felt experience). Group comparisons used non-parametric tests: Mann-Whitney U test for two-group comparisons and Kruskal-Wallis test for multiple-group comparisons. Statistical significance was set at $p < 0.05$.

3. Results

3.1 Sample Characteristics

The final sample consisted of users employed across a range of industries, including healthcare (24%), financial services (14%), education services (10%), software and technology (9%), and transportation (9%). Of the final sample, 4161 sessions (29.1%) involved users returning after at least one prior substantial session, making them eligible for topic continuity analysis. This included 1093 second sessions (7.6% of total; representing first return after an initial conversation) and 3068 sessions (21.5% of total) from users with two or more prior conversations (Table 1).

[Insert Table 1]

3.2 Topic Distribution

At the high-level theme classification, the majority of conversations addressed Health & Wellbeing concerns (encompassing worry-related concerns, low mood & motivation, stress and other wellbeing related requests; n = 9,482; 66.3%), followed by Life Context topics (addressing personal relationships, identity and life transitions; n = 2,121; 14.8%) and Work-related issues (including work stress, workload and professional relationships; n = 2,119; 14.8%).

Within these broad categories, conversations most commonly focused on worry-related concerns (n = 3,216), work stress or workload (n = 1,832), confidence or self-esteem (n = 1,448), sleep (n = 1,360), and lifestyle or personal growth topics (n = 1,330). Interpersonal relationships, low mood & motivation, and physical wellbeing were also frequently discussed, while more specialised topics (e.g. neurodiversity, parenting, bereavement, trauma) occurred less often, reflecting Nova's intended scope as a general wellbeing coach and its use of safeguards to direct users to appropriate alternative resources when topics exceed this remit.

3.3 Topic Continuity Across Sessions

Among non-initial, non-trivial sessions, 70.4% (n = 2,929 of 4,161) were classified as follow-up conversations, indicating that most returning users revisited previously discussed topics. Focusing specifically on second sessions, 78.3% (n = 755 of 953) were follow-ups to an earlier topic rather than a new issue. Return behaviour was significantly associated with whether the previous session was an onboarding interaction ($\chi^2 = 17.979$, $p < .001$), with users less likely to return to the same topic following onboarding sessions.

Topic continuity varied substantially by conversation theme. Work-related topics (Work Stress/Workload, Relationship with Manager, Fulfilment at Work) showed the highest continuation rates. Personal wellbeing topics (such as Worry, Confidence/Self-Esteem, Low Mood & Motivation, Personal Relationships) demonstrated moderate continuation rates (approximately 50-67%). Acute or bounded concerns (Bereavement/Serious Illness of a Loved One, Physical Health Conditions) showed the lowest continuation rates, typically appearing in single sessions only.

Although topic continuity was high across all engagement levels, sessions with higher initial emotional engagement (EXP 4-7) were significantly more likely to continue the same topic (72.3%) compared to sessions with lower engagement (EXP 1–3; 66.7%; $\chi^2 = 13.786$, $p < 0.001$).

3.4 Topic Return Purpose

Having established that most returning users revisited prior topics, we next examined the nature of these returns. Of the sessions that were deemed to be a follow-up on the previous topic,, 44.2% (n = 1,288) reflected progress or deepening of the original issue, while 30.8% (n = 897) showed a mixture of progress alongside unresolved elements. A smaller proportion (24.8%, n = 723) were characterised as unresolved or stuck, and very few sessions (0.2%, n = 6) could not be clearly classified. Overall, the

majority of follow-up conversations showed at least some evidence of continued reflection or movement on the original concern.

3.5 EXP Rating Reliability

Intraclass correlation coefficients (ICC; [15]) demonstrated good agreement between expert raters (ICC = 0.79) and moderate agreement between expert and LLM raters (ICC = 0.66 and 0.62). Combined three-rater ICC was 0.69 (moderate-to-good), representing 87% of expert-expert agreement. Holdout expert-LLM agreement was ICC = 0.63, confirming that the LLM prompt generalised beyond the calibration sample.

3.6 Emotional Engagement Depth: Overall Distribution

Across all 14,293 sessions, the distribution of peak EXP scores showed a median of 4.0 (SD = 1.01; IQR = 2). The distribution was concentrated around scores of 2-4, with approximately 92% of sessions falling within these three values. Scores of 1 (impersonal discussion) were rare, as were scores at the upper end of the scale (6-7).

Cumulative proportions indicated that 10,682 sessions (74.7%) reached EXP ≥ 3 and 236 sessions (1.7%) reached EXP ≥ 5. Mean EXP score was 4. See Table 2 for full breakdown of proportions and cumulative proportions.

[Insert Table 2]

3.7 Engagement Depth by Usage Frequency

Engagement depth varied by usage frequency, though all segments showed median EXP = 4.0. Standard deviations ranged from 0.97 (regular users) to 1.07 (occasional users), suggesting decreasing variability with sustained engagement. Single-session users (n = 5,550) showed 73.53% of sessions reaching EXP ≥ 3 and 1.05% reaching EXP ≥ 5. Occasional users (2-4 sessions; n = 4,185) demonstrated the lowest rate of moderate engagement (70.11% at EXP ≥ 3) but higher deep engagement (1.43% at EXP ≥ 5) than single-session users. Regular users (≥5 sessions; n = 4,558) exhibited the highest rates of both moderate engagement (80.45% at EXP ≥ 3) and deep engagement (2.58% at EXP ≥ 5; Table 3).

Kruskal-Wallis test revealed a statistically significant difference in EXP distributions across usage frequency groups ($H = 184.68$, $df = 2$, $p < 0.001$).

[Insert Table 3]

3.8 Robustness Check: Onboarding vs. Coaching Sessions

To confirm that findings were not biased by conversation type, we compared EXP distributions between onboarding sessions (n = 2,686) and substantive coaching sessions (n = 11,607). Both showed median EXP = 4.0 (SD = 0.95 and 1.02, respectively), with similar threshold attainment rates. Although

distributions differed statistically (U = 16,715,876.5, p < .001), the effect was minimal, indicating that emotional engagement patterns were consistent across session types.

3.9 Variation by Topic

Emotional engagement depth varied significantly by conversation topic. At the high-level theme classification (see Table 4), Life Context conversations achieved the highest moderate engagement (84.1% reaching EXP ≥ 3) and intermediate deep engagement (1.74% reaching EXP ≥ 5). Health & Wellbeing conversations showed the lowest moderate engagement (75.1% reaching EXP ≥ 3) but the highest deep engagement (1.97% reaching EXP ≥ 5). Work-related conversations demonstrated intermediate moderate engagement (81.1% at EXP ≥ 3) but the lowest deep engagement (0.566% at EXP ≥ 5). Kruskal-Wallis test confirmed that these differences in EXP distributions across high-level themes were statistically significant, with all pairwise comparisons yielding p < 0.001.

[Insert Table 4]

At the low-level theme classification (see Table 5), the same pattern persisted. Among the ten most frequent conversation topics, median EXP ranged from 2.0 (Sleep, Lifestyle, Habits & Growth) to 4.0 (most other topics), with Physical Wellbeing at 3.0. Standard deviations ranged from 0.67 for Low Mood & Motivation to 1.11 for Lifestyle, Habits & Growth, indicating substantial variation in emotional engagement depth. Topics involving personal growth and less specific concerns (Lifestyle, Habits & Growth, SD = 1.11; Other Psychological Concerns, SD = 1.06; Sleep, SD = 1.05) showed greater variability in engagement depth compared to more specific symptom-focused topics (Low Mood & Motivation, SD = 0.67; Bereavement, SD = 0.76; Worry-Related Concerns, SD = 0.71).

[Insert Table 5]

4. Discussion

This study examined patterns of sustained engagement and emotional depth in an AI wellbeing tool to address two primary research questions: 1) whether users demonstrate sustained versus episodic engagement, and 2) whether emotional depth increases with repeated use. We also explored how these patterns vary by topic domain. These findings can be interpreted through multiple lenses, each offering a distinct perspective on their implications for AI wellbeing coaching – as we now discuss.

Sustained Engagement and the Coaching Model

Returning users continued prior topics in 70% of sessions, with 75% of those returns showing evidence of progress or deepening. The 70% topic continuity rate suggests that a substantial proportion of users engage with Nova in ways that resemble coaching relationships rather than purely episodic help-seeking. However, this pattern alone does not establish that the system functions equivalently to human coaching. Research on human therapeutic relationships identifies continuity as necessary but not

sufficient on its own for effective coaching; and, perhaps unsurprisingly, the quality of engagement during continued interactions is important too [5].

The convergence of topic continuity with increased rates of deep engagement amongst regular users provides preliminary support for coaching-style relationships developing between users and Nova. Goldman et al. [6] found that deeper experiencing on core themes in later therapy sessions predicted better outcomes, suggesting that the progression observed here – from 2.9% to 6.5% achieving EXP ≥ 5 – may have functional significance. Examination of return purpose further supports this interpretation: amongst users who returned to the same topic, 75% showed at least some evidence of progress or deepening (44.2% clear progress, 30.8% mixed progress with unresolved elements), whilst only 24.8% appeared stuck or unresolved. This pattern suggests that topic continuity reflects iterative working-through processes rather than repetitive help-seeking for unresolved concerns. However, this interpretation requires caution: the observational design cannot determine whether sustained use causes deeper engagement or whether users predisposed to deeper processing are more likely to return.

Recent research shows that AI chatbots can evoke processes associated with therapeutic engagement once thought to require human interaction. For example, studies of systems such as Woebot have demonstrated that users can form meaningful therapeutic alliances with AI, predicting engagement and self-reported improvement [16]. Our findings complement this work by highlighting a parallel intrapersonal process: for a subset of users, engagement with Nova involved returning to prior topics and reaching deeper levels of emotional experiencing over time. Together, these patterns suggest that AI coaches can support iterative, relationship-like processes that foster reflection, perspective-taking, and behavioural experimentation – mechanisms known to underpin change in human coaching.

Emotional Depth and Therapeutic Process

Most sessions achieved moderate engagement (median EXP = 3.0; 61% reaching EXP ≥ 3), whilst deep engagement (EXP ≥ 5) was observed in 4.4% of sessions overall and 6.5% amongst regular users. The overall median EXP finding indicates that most conversations achieved personally meaningful engagement beyond surface-level information exchange. This finding aligns with Klein et al.'s [14] original conceptualisation of EXP level 3 as the point where individuals begin connecting with internal experience. However, the concentration of scores at levels 2-3 (approximately 85% of sessions) and the low rate of deep engagement (4.4% at EXP ≥ 5) suggests that most interactions remained at moderate depth.

It is important to note that Nova was designed as a wellbeing coach rather than a therapeutic intervention, with guardrails in place to signpost users to appropriate professional resources for out of scope requests. EXP levels 6-7, which involve actively working through emotional conflicts in the moment, are characteristic of formal psychotherapy and may be outside the appropriate scope for a coaching tool. From this perspective, the concentration of scores at EXP 3-5 may reflect appropriate

boundary maintenance rather than a limitation, suggesting users engage at personally meaningful depths while remaining within the tool's intended remit.

Even within these boundaries, however, the depth of engagement observed may carry functional significance. Meta-analytic evidence links higher experiencing with better outcomes, though with small-to-medium effect sizes (r = −.19 to −.25; [5]), and increased emotional processing over time predicts symptom reduction [12]. Notably, recent work shows that working EXP (engagement depth in later sessions) predicts outcomes in young people's therapy, whilst early EXP does not [13].

The present study measured the same underlying construct—depth of emotional experiencing—in the context of AI wellbeing coaching. The twofold increase in deep engagement amongst regular Nova users (from 2.9% to 6.5% reaching EXP ≥ 5) mirrors the pattern observed in human therapy research: progression toward deeper emotional processing over time. This parallel suggests that users may develop their capacity for meaningful self-exploration through sustained AI coaching interactions. However, the 6.5% rate amongst regular users remains substantially lower than rates typically observed in effective human therapy [6], indicating room for improvement..

Domain-Specific Patterns and Appropriate Boundaries

Work-related topics showed the highest continuity rates but the shallowest emotional depth (1.5% reaching EXP ≥ 5), whilst acute concerns such as bereavement showed the inverse pattern. This variation warrants careful interpretation. High continuity for work-related topics coupled with shallow emotional depth could reflect several mechanisms: users may seek practical problem-solving rather than emotional exploration for workplace concerns; organisational contexts may constrain willingness to process work-related emotions deeply; or the system's conversational strategies may be less effective at facilitating depth in professional domains. Research on workplace coaching suggests that effective professional development does involve emotional processing alongside skill-building [17], suggesting the shallow engagement observed here may represent a limitation rather than an appropriate match to user needs.

Conversely, the low continuity for bereavement and physical health concerns, combined with moderate-to-high depth when these topics were addressed, suggests appropriate boundary recognition. These patterns align with professional consensus that AI tools should support wellbeing and development rather than providing treatment for clinical conditions or acute crises [3]. However, whether this boundary recognition reflects user judgement, system design features, or both remains unclear from the present data.

4.1 Limitations

Several limitations impact interpretation of these findings. First, the observational design precludes causal inference. Users who developed sustained relationships may have differed systematically from

those who did not in ways unrelated to the AI system itself, such as motivation, problem severity, or access to alternative support resources.

Second, whilst LLM-based EXP scoring achieved moderate-to-good reliability (ICC = 0.69), it did not reach the good reliability of expert–expert agreement (ICC = 0.79). This 0.10-point gap represents measurement error that may attenuate observed effects and reduce sensitivity to change. Additionally, although the use of LLMs for classification enabled scoring of large datasets that would be impractical with human raters alone, it also introduces inherent non-determinism and limited interpretability, precluding full transparency about classification decisions. The clinical significance of small differences in EXP scores (e.g., 2.9% vs. 6.5% at EXP ≥ 5) remains uncertain.

Third, the study sample consisted of adult, employed individuals accessing Nova through workplace programmes, limiting generalisability to clinical populations, individuals in crisis, or those without organisational access. Engagement patterns may differ substantially across demographic groups and clinical presentations.

Fourth, this study measured only peak EXP within sessions rather than tracking experiencing trajectories within or across interactions. Research suggests that patterns of change may be more informative than single-point assessments [6,12], and evidence from adolescent therapy indicates that working EXP (later sessions) predicts outcomes whilst early EXP does not [13]. Additionally, topic continuity was assessed via LLM classification of system-generated summaries rather than user-reported goals, which may not fully capture user intent.

Finally, the absence of outcome measures currently prevents assessment of whether observed engagement patterns translate into wellbeing improvements. Whilst emotional depth correlates with outcomes in human therapy research, whether this relationship holds for AI interactions requires empirical verification through longitudinal designs with validated outcome measures.

4.2 Future Directions

Several questions merit investigation. First, prospective studies linking engagement patterns to standardised wellbeing and symptom measures would clarify whether topic continuity and deepening emotional engagement predict meaningful outcomes, and with what magnitude. Second, experimental designs manipulating specific conversational features could establish causal relationships between system design choices and engagement patterns. Third, examination of experiencing trajectories within and across sessions, rather than peak scores alone, may provide more sensitive indicators of relationship development. Fourth, replication in diverse populations – including clinical samples, adolescents, and individuals from varied cultural contexts – would clarify the generalisability of these patterns.

4.3 Implications

These findings suggest that users engage with AI wellbeing tools in sustained, iterative ways, with modest increases in emotional depth amongst regular users. This pattern indicates potential for AI systems to support coaching-style relationships within appropriate boundaries. However, the modest absolute rates of deep engagement (4.4% overall; 6.5% amongst regular users) and concentration of interactions at moderate depth suggest substantial opportunity for enhancement.

For system design, these findings support investment in continuity features (session summaries, follow-up prompts, longitudinal tracking) and conversational strategies that facilitate progressive elaboration. The particularly shallow engagement in high-volume work-related conversations represents a clear opportunity for system designers seeking to deepen coaching engagement beyond practical problem-solving. For evaluation, these findings demonstrate that metrics beyond usage volume – specifically, topic continuity rates and emotional depth distributions – can provide insight into engagement quality. Establishing baselines and tracking changes following feature releases could support evidence-based development.

4.4 Conclusion

This study provides evidence that sustained use of an AI wellbeing tool is associated with deeper emotional experiencing, with regular users consistently reaching deeper levels than those in single sessions. Coupled with high topic continuity amongst returning users, these patterns suggest that AI systems can support iterative, coaching-style self-exploration within appropriate boundaries. The magnitude of deepening observed was modest and questions of clinical significance remain, but the demonstrated feasibility of assessing relationship quality at scale using LLM-based methods calibrated against expert judgement provides a foundation for the research needed to connect engagement patterns to wellbeing outcomes.

Declarations

Author contributions:

MM, VMI, and ME conceived and designed the study, and wrote and refined the LLM prompts used for classification and scoring. MM wrote the initial draft and oversaw integration of co-author input into the final manuscript. VMI built the data pipelines and LLM classification systems and performed the statistical analyses with input from ME (who also made significant contributions to the main text). MM served as an expert rater for Experiencing Scale calibration and provided clinical oversight of Nova throughout the study period. KMW contributed to interpretation of results and the final manuscript, and prepared the manuscript for publication. All authors reviewed and approved the final version.

Competing interests:

Authors MM, VMI, and KMW are employed by and own share options at Unmind Ltd, the creator of the AI coach evaluated. ME was employed by Unmind Ltd as a paid contractor at the time the study was

conducted and owns share options.

Funding:

This study was funded in full by Unmind Ltd, the creator of the AI coach evaluated. No specific grant number is applicable for this internal funding.

Author Affiliations:

[1] Science & Data Department, Unmind Ltd, London, UK.

Corresponding Author:

Max Major, max.major@unmind.com

Data availability:

The datasets analysed during the current study are not publicly available due to privacy and confidentiality obligations but an aggregated view is available from the corresponding author on reasonable request.

Acknowledgements:

The authors would like to acknowledge Dr. Heather Bolton and Becky Minton for their role as expert raters in the calibration of the Experiencing Scale.

References

1. Lopes, R. M., Silva, A. F., Rodrigues, A. C. A. & Melo, V. Chatbots for Well-Being: Exploring the Impact of Artificial Intelligence on Mood Enhancement and Mental Health. *Eur. Psychiatry* **67**, S550–S551 (2024).
2. Gamble, A. Artificial Intelligence And Mobile Apps For Mental Healthcare: A Social Informatics Perspective. in *Proceedings of the International Conferences on Internet Technologies & Society (ITS 2020) and Sustainability, Technology and Education (STE 2020)* 101–104 (IADIS Press, 2020). doi:10.33965/its_ste2020_202001D014.
3. Balcombe, L. AI Chatbots in Digital Mental Health. *Informatics* **10**, 82 (2023).
4. Moylan, K. & Doherty, K. Expert and Interdisciplinary Analysis of AI-Driven Chatbots for Mental Health Support: Mixed Methods Study. *J. Med. Internet Res.* **27**, e67114 (2025).
5. Pascual-Leone, A. & Yeryomenko, N. The client “experiencing” scale as a predictor of treatment outcomes: A meta-analysis on psychotherapy process. *Psychother. Res.* **27**, 653–665 (2017).
6. Goldman, R. N., Greenberg, L. S. & Pos, A. E. Depth of emotional experience and outcome. *Psychother. Res.* **15**, 248–260 (2005).

7. Scholich, T., Barr, M., Wiltsey Stirman, S. & Raj, S. A Comparison of Responses from Human Therapists and Large Language Model–Based Chatbots to Assess Therapeutic Communication: Mixed Methods Study. *JMIR Ment. Health* **12**, e69709 (2025).
8. Boucher, E. M. & Raiker, J. S. Engagement and retention in digital mental health interventions: a narrative review. *BMC Digit. Health* **2**, 52 (2024).
9. Smith, K. A. *et al.* Engagement and attrition in digital mental health: current challenges and potential solutions. *Npj Digit. Med.* **8**, 398 (2025).
10. Jabir, A. I. *et al.* Attrition in Conversational Agent–Delivered Mental Health Interventions: Systematic Review and Meta-Analysis. *J. Med. Internet Res.* **26**, e48168 (2024).
11. Cross, S. P. & Alvarez-Jimenez, M. The digital cumulative complexity model: a framework for improving engagement in digital mental health interventions. *Front. Psychiatry* **15**, 1382726 (2024).
12. Pinheiro, P., Gonçalves, M. M., Sousa, I. & Salgado, J. What is the effect of emotional processing on depression? A longitudinal study. *Psychother. Res.* **31**, 507–519 (2021).
13. Geyer, D., Lam, V., Gilbert, H. & Cooper, M. Depth of emotional experiencing and outcome in therapy with young people. *Psychol. Psychother. Theory Res. Pract.* **98**, 308–321 (2025).
14. Klein, M. H., Mathieu-Coughlan, P., Gendlin, E. T., & Kiesler, D. J. *The Experiencing Scale: A Research and Training Manual.* (Wisconsin Psychiatric Institute., 1969).
15. Koo, T. K. & Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **15**, 155–163 (2016).
16. 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 Form. Res.* **5**, e27868 (2021).
17. Grant, A. M. An integrated model of goal-focused coaching: An evidence-based framework for teaching and practice. *Int. Coach. Psychol. Rev.* **7**, 146–165 (2012).

Tables

Table 1. Distribution of User Sessions for Topic Continuity Analysis.

Session category	n	% of total sample
Total analytical sample	14,293	100
First sessions (no prior substantial session)	10,132	70.9
Returning sessions (≥1 prior substantial session)	4161	29.1
Second sessions	1093	7.6
Third+ sessions	3068	21.5

Table 2. Distribution of Peak Experiencing (EXP) Scores Across Sessions.

EXP score	n	%	Cumulative %
1	928	6.49	6.49
2	2683	18.77	25.26
3	1993	13.94	39.2
4	8453	59.14	98.35
5	184	1.29	99.64
6	52	0.36	100
7	0	0	100
Total	**14,293**	**100**	—

Table 3. Engagement Depth (Peak EXP) by User Frequency Segment.

User segment	n	Median EXP (SD)	% EXP ≥ 3	% EXP ≥ 5
Single-session	5,550	4 (0.99)	73.53	1.05
Occasional (2-4 sessions)	4,185	4 (1.07)	70.11	1.43
Regular (≥5 sessions)	4,558	4 (0.97)	80.45	2.58

Table 4. Emotional Engagement Depth by High-Level Conversation Themes

Theme	% EXP ≥ 3	% EXP ≥ 5
Life Context	67.7	6.8
Health & Wellbeing	63.7	4.8
Work-related	61.3	1.5

Table 5. Engagement Depth and Variability by Low-Level Conversation Themes.

Theme	n	Median EXP	SD
Worry-Related Concerns	3,216	4.0	0.71
Work Stress / Workload	1,832	4.0	0.86
Confidence / Self-esteem	1,448	4.0	0.84
Sleep	1,360	2.0	1.05
Lifestyle, Habits & Growth	1,330	2.0	1.11
Personal Relationships	1,282	4.0	0.82
Low Mood & Motivation	991	4.0	0.67
Physical Wellbeing	462	3.0	1.03
Other Psychological Concerns	270	4.0	1.06
Bereavement / Serious Illness of a Loved One	263	4.0	0.76