Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Julie Y. A. Cachia, Xuan Zhao, John Hunter et al. AI for Proactive Mental Health: A Multi-Institutional, Longitudinal Randomized Controlled Trial. NEJM AI. DOI: 10.1056/AIoa2501293.

Supplemental Materials for

AI for Proactive Mental Health: A Multi-Institutional, Longitudinal, Randomized Controlled Trial

Table of Contents

Supplementary Methods: The Flourish App	**3**
Section 1: App Engagement	**10**
Section 2: Diversity and Representativeness of the Study Population	**11**
Section 3: Sensitivity Analyses Applying Additional Exclusion Criteria	**13**
Section 4: Exploratory Factor Analysis (EFA) on Outcomes	**15**
Section 5: Study Measures	**17**
Section 6: All Detailed Results	**19**
Section 7: Sensitivity Analyses Controlling for Demographic Covariates	**27**
Section 8: Exploratory Moderation Analyses	**28**
Section 9: Detailed Results for Subcomponents of Positive Affect	**31**
Section 10: Open-Ended Response Coding Rubric and Results	**33**
References	**39**

Supplementary Methods: The Flourish App

The Flourish app (https://www.myflourish.ai/) is a mobile well-being intervention app launched in 2024 and developed for iOS and Android using Flutter, with a cloud-based backend hosted on Amazon Web Services (AWS). It integrates generative AI with decades of research in well-being science to deliver personalized, gamified, strengths-based mental health support. At the center of the app is Sunnie, an AI-powered well-being chatbot that engages users in brief, emotionally intelligent conversations and recommends relevant knowledge, insights, and science-based coping strategies. Unlike static tools, Sunnie uses a sophisticated workflow based on a variety of large language models (LLMs) to adapt in real time to users' mood, goals, and behavior, adjusting its tone, content, and suggestions to meet users where they are.

The Flourish app is designed to promote well-being through an intuitive behavior loop (see Figure 1 in the manuscript; Figure S1 for a summary of the core generative AI components; and Figure S2 for a walkthrough of a typical daily user experience). The loop begins with a brief emotion check-in, followed by real-time support and feedback through multi-turn conversations with Sunnie that focus on building emotional awareness and learning relevant psychological knowledge. Next, Sunnie provides science-based activity recommendations—such as journaling for emotion regulation[1–3], positive psychology exercises (e.g., three good things[4]), social connection activities (e.g., gifting a compliment[5,6], meaningful conversations[7,8]), and mindfulness techniques (e.g., nature walks[9–11]). Then, each interaction is positively reinforced with rewards (e.g., AI-generated badges, streaks), and together, they contribute to an AI-generated weekly report, which synthesizes mood trends to help users reflect on their progress. In doing so, Flourish creates a closed, personalized feedback loop—moving from receiving support, to taking action, to gaining insight, and then back to renewed support interactions. This

adaptive loop is informed by what we call the STAR framework for beneficial AI use—Science-based, Timely, Action-oriented (activating positive behavior change, rather than only talking about thoughts and feelings), and Real-life-focused (encouraging connections with other people and communities, rather than only with AI).

AI Architecture. As an AI-native mobile application, the Flourish app's modular architecture integrates several interconnected AI services powered by dozens of LLM prompts and workflows.

The Conversational Agent manages user-Sunnie dialogues within a chatbot interface and incorporates an advanced long-term memory system, multimodal interaction capabilities (including text, voice, and image), in-chat resource recommendation, and tool-calling functions (such as surfacing personalized insight summaries or launching relevant well-being activities directly within the chat). It is also connected to an automatic notecard generator that summarizes key takeaways, actionable tips, and notable insights after each conversation. The Personalized Recommendation Engine leverages LLMs to provide science-based well-being activity suggestions tailored to each user's moods, contexts, and engagement history. The Activity Library hosts a collection of pre-developed well-being activities offered either as AI-guided chat experiences such as "Three Good Things," or as non-AI interactive activities such as "Breathing with Sunnie," which features calming music, animation, haptic feedback, and a timer. A Habit-Building and Gamification System rewards users with AI-classified badges based on the PERMA model of well-being[12] for completing chats, activities, or journal entries, and tracks engagement through a continuously increasing streak counter. An opt-in Push Notification Service delivers pre-drafted or AI-generated daily reminders, positive affirmations, and check-in messages to encourage consistent engagement. A dedicated Crisis Service continuously scans for

high-risk language and activates escalation protocols during the conversation. The Analytics and Insights module aggregates user data to produce weekly summaries. Finally, authentication and user management are handled through Firebase, and the infrastructure is orchestrated with Terraform, supported by continuous integration and deployment (CI/CD) pipelines powered by GitLab. At the time of this research, all conversational and recommendation prompts were powered by the OpenAI API, using either GPT-4o or GPT-4o-mini.

Conversation Design and Validation. Sunnie's chat behavior is designed to balance listening, reflection, and gentle guidance, leveraging methods commonly found across psychological, affective, and behavioral sciences. For example, when users experience positive events, Sunnie guides them to savor and deepen their positive emotions by noticing, amplifying, sharing the experience with others, or expressing appreciation to others. When users face challenges or distressing experiences, Sunnie draws on evidence-based strategies to foster resilience and self-compassion, which may include techniques such as building emotional awareness, practicing cognitive reappraisal, engaging in expressive writing, recognizing and reframing unhelpful thoughts, engaging in mindfulness and relaxation exercises, affirming personal values, and taking actions aligned with personal goals. When appropriate, Sunnie also promotes social connection and help-seeking, guiding users toward constructive communication, conflict resolution, and professional support when needed. All prompts were developed by psychologists and iteratively refined through extensive testing, red-teaming, and collaboration with mental health researchers and clinicians with diverse training backgrounds prior to field deployment. This process involved systematic evaluation of tone calibration, guardrail performance, and adherence to evidence-based content, ensuring that responses were natural, insightful, empathic, safe, and trustworthy.

Monitoring and Quality Assurance. The AI-first, human-in-the-loop continuous monitoring pipeline tracks multiple layers of performance and safety data, including conversation-level metrics (e.g., user-rated conversation quality, AI-generated mental health risk assessments), individual-level metrics (e.g., mental health profiles and ratings), and system-level metrics (e.g., latency, response compliance, and escalation triggers). The conversation-level feedback is fed back to Sunnie to refine and personalize its conversational style for each user, while mental health assessments and trigger data are used for quality monitoring and assurance by the human review team. Reviewers provide targeted feedback to continuously improve the AI system, ensuring accuracy, clinical safety, and emotional resonance. Importantly, no user data are ever used to train the underlying Large Language Models (LLMs).

Safety, Privacy, and Security. In addition to the OpenAI API's built-in guardrails, our purpose-built crisis identification and safety protocol operates automatically in real time. When potential risk statements are detected, the system triggers a just-in-time safety check to assess user safety and determine the appropriate course of action. When warranted, it displays relevant national resources (e.g., the 988 Suicide and Crisis Lifeline in the U.S.) or local community resources retrieved from our database using Retrieval-Augmented Generation (RAG), and generates a deidentified flag for human review.

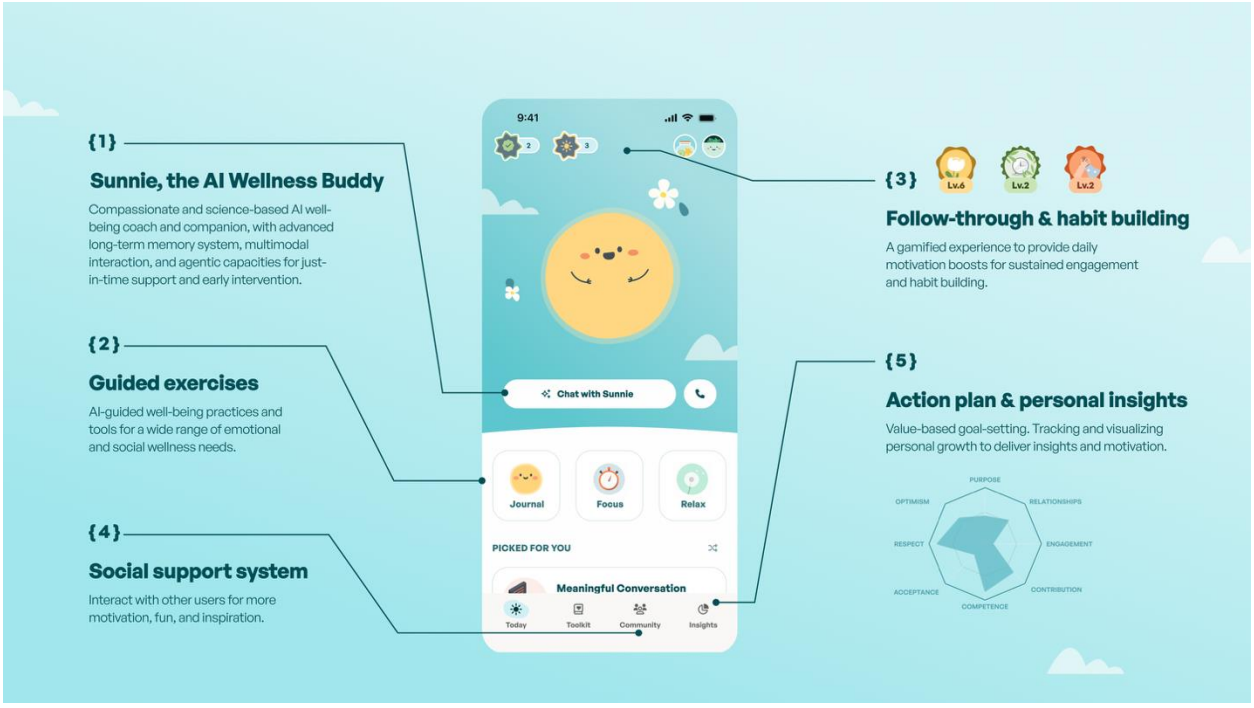
Flourish is a confidential well-being resource. No personally identifiable information is shared with partner institutions, except in cases of imminent risk. All data are encrypted in transit and at rest using industry-standard protocols and stored securely in compliance with HIPAA (Health Insurance Portability and Accountability Act), GDPR (General Data Protection Regulation), and CCPA (California Consumer Privacy Act) requirements. Data are never sold or shared with

external commercial entities. Access to information is restricted to authorized professionals working under strict data protection and ethical oversight, using secure, VPN-protected systems.

No medical advice or diagnosis was provided. No personally identifying information was stored for research purposes; all conversations were deidentified prior to human review, which is conducted solely to ensure AI safety, quality assurance, and program evaluation. Flagged conversations were reviewed by trained staff under the supervision of a clinical psychologist, following a predefined safety protocol. No adverse events were observed during the study, and no participants were contacted or provided with additional support.

Figure S1

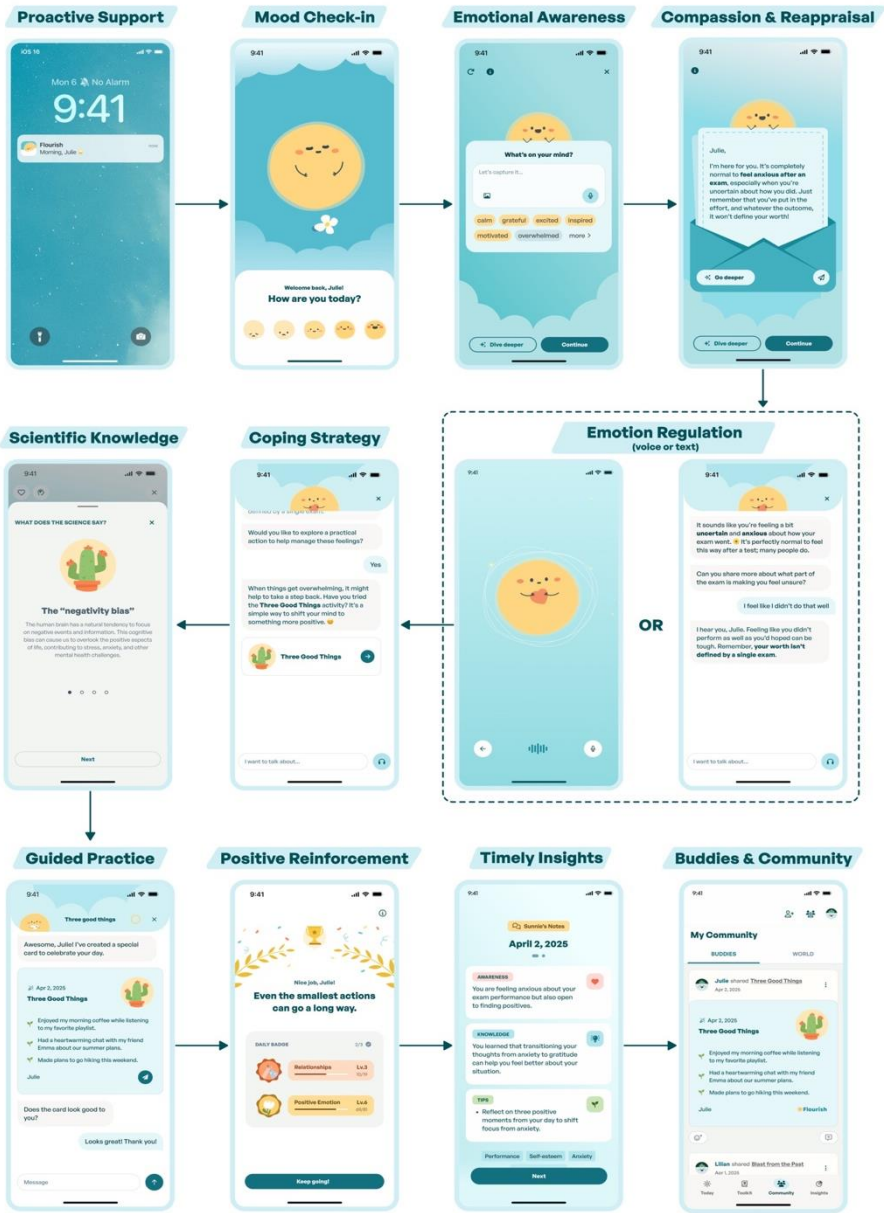
Overview of Core Generative-AI Components in the Flourish App



Note. The system integrates (1) an AI wellness companion with long-term memory and just-in-time support, (2) guided science-based well-being practices, (3) gamified engagement for positive habit building, (4) social support system, and (5) value-aligned action planning and personal well-being insights.

Source: Authors' original schematic of the Flourish app architecture.

Figure S2
Daily User Interaction Flow in the Flourish App



Note. The daily user flow includes 11 steps: (1) a proactive personalized prompt invites users to check in; they (2) select a mood and (3) choose keywords to identify their emotional state and its source. Based on this input, (4) Sunnie provides a supportive message with affirmation and insight. Users may then (5) engage in a deeper multi-turn reflection. Sunnie (6) recommends a tailored coping strategy and (7) briefly explains the science behind it. Users are (8) guided through the practice step-by-step and (9) receive positive reinforcement. The flow concludes

with (10) a summary of insights and (11) optional connection with peers through buddy or community features.

Source: Authors' original schematic of the Flourish app user flow.

Section 1: App Engagement

As a manipulation check, we examined behavioral engagement with the Flourish app among participants assigned to the treatment condition. Engagement was measured through participant transcription of app-tracked metrics displayed on their in-app Insights page. Based on these participant-transcribed app metrics, students averaged 3.49 days per week (*SD* = 3.46) and 3.39 completed activities per week (*SD* = 2.52).

Engagement was sustained across the intervention period. Students used the app for an average of 5.85 days (*SD* = 4.23) during Weeks 0–2, 4.58 days (*SD* = 4.31) during Weeks 2–4, and 5.46 days (*SD* = 4.92) during Weeks 4–6. Most students engaged at least once in each two-week interval (Weeks 0–2: 96%; Weeks 2–4: 78%; Weeks 4–6: 76%), and 72% engaged at least once during every interval. Overall, 43% of students met the recommended minimum of at least 2 active days per week across all periods. Only 1.74% of treatment participants (*n* = 3) had zero cumulative active days recorded by the end of study.

Section 2: Diversity and Representativeness of the Study Population

This section summarizes the representativeness of the study sample and key limits to generalizability.

Table S1
Diversity and Representativeness of the Study Population

Category	Description
Disease, problem, or condition under investigation	Emotional, social, and overall mental health and well-being.
Special considerations related to	
Sex and gender	Women are approximately twice as likely as men to be diagnosed with depression and approximately 1.7 times more likely to meet criteria for an anxiety disorder, whereas men are significantly less likely to use mental health services. Transgender, nonbinary, and gender diverse individuals face a 2- to 4-fold elevated risk of depression and suicidality relative to cisgender peers, yet are frequently underrepresented in digital mental health intervention research.
Age	Young adults aged 18–25 have the highest prevalence of any mental illness of any adult age group, with 36.2% meeting criteria compared to 29.4% of adults aged 26–49 and 13.9% of those aged 50 and older. Globally, approximately 62.5% of lifetime mental disorders first emerge before age 25, with a median age of onset of 18 years. The college years represent a particularly high-risk window, marked by major life transitions.
Race or ethnicity group	While the prevalence of mental illness is broadly similar across racial and ethnic groups, racial and ethnic minority individuals are substantially less likely to receive mental health treatment, with treatment rates approximately one-third lower among Black and Hispanic adults and less than half as high among Asian Americans compared to White adults. When mental illness goes untreated, it tends to be more persistent and disabling.
Geography	Mental health burden varies substantially across countries and regions. In global burden estimates, depressive disorder burden is especially high in sub-Saharan Africa and North Africa/Middle East, whereas eating disorder burden is higher in high-income regions. Digital mental health interventions have been tested largely in Western, high-income contexts, so findings from U.S.-based studies may have limited generalizability to other national and cultural settings.
Other considerations	Socioeconomic status (SES) is among the strongest predictors of mental health burden: individuals with low SES have approximately 80% higher odds of depression than those with high SES, with a dose-response relationship observed for both income and education level. Financial hardship, housing instability, and unemployment contribute independently to poor mental health while simultaneously limiting access to care. In college settings, first-generation and low-income students face compounded stressors related to financial insecurity, employment demands, and caregiving responsibilities, yet are underrepresented in mental health intervention research.
Overall representativeness of this trial	The sample of 486 undergraduate students was recruited without symptom-based screening from three U.S. institutions: two four-year universities (Chapman University and the University of Washington) and one community college (Foothill College). The sample skewed female across all sites (Chapman: 80% female, 20% male; University of Washington: 79% female, 19% male, 2% not specified; Foothill College: 73% female, 27% male), consistent with recruitment through undergraduate psychology research pools and courses. Mean ages were 19.75 years (SD = 1.99) at Chapman, 19.74 years (SD = 2.25) at the University of Washington, and 24.20 years (SD = 8.71) at Foothill

	College, reflecting the older age profile often observed in community college samples. Racial and ethnic composition varied by site: Chapman was predominantly White/European American (53%), with mixed-race (20%), Hispanic/Latinx (10%), Asian/Asian American (10%), and Black/African American (2%) students; the University of Washington included White/European American (38%), Asian/Asian American (29%), mixed-race (16%), Black/African American (6%), and Hispanic/Latinx (5%) students; and Foothill College was the most diverse, with White/European American (24%), Hispanic/Latinx (23%), Asian/Asian American (18%), mixed-race (15%), and Black/African American (9%) students. International students constituted 2.0%, 9.9%, and 16.7% of enrolled participants at Chapman, the University of Washington, and Foothill College, respectively. The study was conducted entirely in the United States and in English, limiting cross-national and cross-linguistic generalizability. The inclusion of a community college strengthened institutional diversity relative to single-site studies or studies conducted only at four-year universities. Nonetheless, the nonclinical, undergraduate sample limits generalizability to clinical populations, older adults, nonstudents, and more globally diverse settings.

Section 3: Sensitivity Analyses Applying Additional Exclusion Criteria

Sensitivity analyses applying the preregistered exclusion criteria yielded a substantively similar pattern of results. These criteria excluded participants with fewer than two completed timepoints ($n = 75$), treatment participants with zero cumulative active days by study end ($n = 2$), and control participants who reported exposure to or use of the Flourish app during the study period ($n = 7$).

For the 4-wave outcomes, the Condition × Time interactions for positive affect and loneliness remained significant and were similar in magnitude to the primary analyses, while null effects for negative affect, depression, anxiety, perceived stress, and vigour were unchanged. For the outcomes assessed only at baseline and Week 6, interaction effect sizes for belonging, resilience, mindfulness, and flourishing remained nearly identical in magnitude and direction, although p-values became more marginal in the reduced sample, consistent with reduced statistical power following exclusions (Table S5).

Table S5

Condition × Time Interaction Results Using Additional Exclusion Criteria

Outcome Variable	Condition × Time (*b*)	*SE*	*df*	*t*	*p*-value	*β*
Positive Affect	0.21	0.10	1086.84	2.15	.032	0.03
Calm	0.09	0.04	1092.45	2.17	.030	0.04
Well-Being	0.08	0.04	1089.61	2.17	.030	0.04
Vigour	0.04	0.04	1090.66	1.12	.262	–
Negative Affect	0.00	0.10	1086.02	-0.04	.964	–
Loneliness	-0.05	0.02	1084.42	-2.21	.027	-0.03
Belonging	0.05	0.03	350.51	1.77	.078	0.03
Closeness to Academic Institution	0.04	0.02	359.90	1.79	.074	0.04
Resilience	0.11	0.06	353.73	1.83	.068	0.04
Mindfulness	0.18	0.10	360.71	1.87	.063	0.05

Flourishing	0.18	0.09	350.92	1.95	.053	0.04
Depression	-0.03	0.02	1088.06	-1.26	.208	–
Anxiety	-0.01	0.03	1086.97	-0.52	.602	–
Perceived Stress	-0.02	0.04	1091.01	-0.47	.641	–

Note. Coefficients (*b*) reflect the Condition × Time interaction term from linear mixed-effects models.
Source: Authors' analyses of data from the present study.

Section 4: Exploratory Factor Analysis (EFA) on Outcomes

An exploratory factor analysis (EFA) was conducted on baseline outcome measures (n = 483) using maximum-likelihood extraction with varimax rotation in R using the psych package. Parallel analysis supported a three-factor solution accounting for ~51% of the variance (see Table S2). As expected, emotional well-being indicators (calm, well-being, vigor) loaded strongly on one factor; social connection indicators (belonging and closeness to one's academic institution) loaded clearly on a second factor; and overall well-being indicators (resilience and mindfulness) loaded on a third factor. Several individual indicators showed partial cross-loadings (e.g., flourishing across emotional and overall well-being; loneliness loading more strongly with overall well-being; and anger not cleanly loading on the emotional factor). Given their theoretical roles and interpretation in prior work, we retained these outcomes in their predefined conceptual domains. Overall, the EFA supported a multidimensional well-being structure with modest, expected overlap among domains, consistent with contemporary models of well-being.

Table S2

Exploratory Factor Loadings for Well-Being Outcomes (Varimax Rotation)

Outcome	Factor 1 (Emotional)	Factor 2 (Social)	Factor 3 (Overall)	Uniqueness
Positive Affect: Calm	**0.60**	0.06	-0.31	0.54
Positive Affect: Well-Being	**0.80**	0.22	-0.30	0.22
Positive Affect: Vigour	**0.82**	0.23	-0.13	0.26
Negative Affect: Depression	**-0.40**	-0.15	**0.70**	0.32
Negative Affect: Anxiety	-0.30	-0.09	**0.64**	0.49
Negative Affect: Anger	0.02	0.00	**0.60**	0.64

Loneliness	-0.34	-0.28	**0.47**	0.58
Belonging	0.23	**0.81**	-0.07	0.29
Closeness to Academic	0.08	**0.59**	-0.06	0.64
Institution				
Resilience	0.30	0.10	**-0.41**	0.73
Mindfulness	-0.25	0.04	**-0.55**	0.63
Flourishing	**0.49**	0.35	**-0.41**	0.47

Note. Loadings ≥ |0.40| in bold. Maximum-likelihood extraction, varimax rotation.
Source: Authors' analyses of baseline data from the present study.

Section 5: Study Measures

Table S3 summarizes all study measures, including their source scales, item content, response ranges, scoring, and timepoints of administration. Social fit was assessed but excluded from analyses because the two-item shortened measure showed low inter-item correlation (r = .20–.23).

Table S3

Summary of Measures

Measure	Description	Items / Subscales	Scale Range	Timepoints Administered
Positive and Negative Affect	Subcomponents of Affect Scale (SAS; Jenkins et al., 2023); 18 adjectives (9 positive, 9 negative)	Positive: Calm (calm, at ease, relaxed), Well-being (happy, cheerful, pleased), Vigour (full of pep, lively, energetic); Negative: Depression (sad, unhappy, depressed), Anxiety (on edge, tense, nervous), Anger (hostile, angry, resentful)	0 (not at all accurate) to 4 (extremely accurate); Composite scores: 0–36 (positive/negative)	All four timepoints (α = .74 –.92)
Loneliness	UCLA 3-item Loneliness Scale (Russell et al., 1980)	"Lack companionship", "Feel left out", "Feel isolated"	1 (hardly ever) to 3 (often); Total: 3–9	All four timepoints (α = .74 –.81)
Belonging	Single-item from Perceived Cohesion Scale (Bollen & Hoyle, 1990)	"I see myself as a part of the campus community"	0 (strongly disagree) to 10 (strongly agree)	Beginning and end of study (single-item)
Closeness to Academic Institution	Single-item Inclusion of Other in the Self (IOS) Scale (Aron et al., 1992)	Visual selection of 1–7 overlapping circles	1 (least overlap) to 7 (most overlap)	Beginning and end of study (single-item)
Resilience	6-item Brief Resilience Scale (Smith et al., 2008)	E.g., "I tend to bounce back quickly after hard times", "I have a hard time making it through stressful events" (reverse-coded)	1 (strongly disagree) to 5 (strongly agree); Total: 6–30	Beginning and end of study (α = .85 –.86)
Mindfulness	5-item state version of Mindful Attention Awareness Scale (MAAS; Brown & Ryan, 2003)	E.g., "I was finding it difficult to stay focused", "I was doing things automatically"	0 (not at all) to 6 (very much); Total: 0–30 (reverse coded so higher = more mindfulness)	Beginning and end of study (α = .84 –.86)
Flourishing	8-item Flourishing Scale (Diener et al., 2010)	E.g., "I lead a purposeful and meaningful life", "My social relationships are supportive and rewarding"	1 (strongly disagree) to 7 (strongly agree); Total: 8–56	Beginning and end of study (α = .87 –.89)

Measure	Description	Items / Subscales	Scale Range	Timepoints Administered
Depression	2-item Patient Health Questionnaire-2 (PHQ-2; Löwe et al., 2005)	E.g., "Little interest or pleasure in doing things"	0 (not at all) to 3 (nearly every day); Total: 0-6	All four timepoints (r = .64 - .68)
Anxiety	2-item Generalized Anxiety Disorder Questionnaire (GAD-2; Spitzer et al., 2006)	E.g., "Feeling nervous, anxious or on edge"	0 (not at all) to 3 (nearly every day); Total: 0-6	All four timepoints (r = .73 - .75)
Stress	4-item Perceived Stress Scale (PSS-4; Cohen et al., 1983)	E.g., "In the last two weeks, how often have you felt that you were unable to control the important things in your life?"	0 (never) to 4 (very often); Total: 0-16	All four timepoints (α = .66 –.73)
Academic Self-Efficacy	5-item General Academic Self-Efficacy Scale (GASE; van Zyl et al., 2022)	E.g., "I generally manage to solve difficult academic problems if I try hard enough"	1 (strongly disagree) to 5 (strongly agree); Total: 5-25	Beginning and end of study (α = .72 - .74)
Social Fit	2-item shortened version of the Sense of Social Fit Scale (Walton & Cohen, 2007)	E.g., "I fit in well at my school"	1 (strongly disagree) to 5 (strongly agree); Total: 2-10	Beginning and end of study (r = .20 - .23)

Note. α values represent Cronbach's alpha for multi-item measures, r values represent inter-item correlations for 2-item measures. The PSS-4 originally assesses stress over the past month; however, we modified the time frame to the past two weeks to align with our biweekly study design.
Source: Authors' compilation based on study instruments and cited source measures.

Section 6: All Detailed Results

In this section, we report the full detailed results for all primary study outcomes (for means and standard deviations for each outcome by condition and timepoint, see Table S4).

Positive and Negative Affect

For positive affect, there was a marginal effect of Condition (b = 0.54, SE = 0.27, t(473.18) = 1.97, p = .050, β = 0.17), indicating that those in the treatment condition reported greater positive affect. In contrast, there was no main effect of Time (b = -0.02, SE = 0.10, t(1148.56) = -0.25, p = .799). However, there was a significant Condition × Time interaction (b = 0.22, SE = 0.10, t(1153.08) = 2.33, p = .020, β = 0.04, 95% CI [0.01, 0.07]).

Post-hoc analyses revealed no significant condition differences at Time 1 (b = 0.22, SE = 0.31, t(482) = 0.73, p = .464), but the treatment group reported marginally higher positive affect at Time 2 (b = 0.58, SE = 0.34, t(390) = 1.70, p = .089, d = 0.17, 95% CI [-0.03, 0.37]) and significantly higher positive affect at both Time 3 (b = 0.84, SE = 0.37, t(353) = 2.25, p = .025, d = 0.24, 95% CI [0.03, 0.45]) and Time 4 (b = 1.08, SE = 0.35, t(347) = 3.07, p = .002, d = 0.33, 95% CI [0.12, 0.54]). Within-condition analyses indicated that while positive affect remained stable in the treatment condition (b = 0.20, SE = 0.14, t(572.82) = 1.44, p = .150), it showed a marginal decline in the control group (b = -0.25, SE = 0.13, t(573.71) = -1.89, p = .059, β = -0.04). These findings suggest that the treatment buffered against the decline in positive affect observed in the control condition.

When we examined the subcomponents of positive affect (i.e., Calm, Well-Being, Vigour), we found that the treatment showed benefits for the Calm and Well-Being subcomponents, in

contrast to no significant effect on Vigour, perhaps suggesting that the treatment is especially beneficial for mid- to lower-arousal positive emotions (see Supplement Section 9 for detailed results).

In contrast, for negative affect, there was no main effect of Condition (b = -0.24, SE = 0.27, t(469.49) = -0.88, p = .381), nor a main effect of Time (b = -0.13, SE = 0.10, t(1151.68) = -1.29, p = .198). Additionally, the Condition × Time interaction was not significant (b = 0.01, SE = 0.10, t(1151.68) = 0.08, p = .933). Similarly, none of the subcomponents of negative affect (i.e., Depression, Anxiety, Anger) showed any significant Condition x Time interaction effects (all ps ≥ .516).

Loneliness

There was no main effect of Condition (b = -0.09, SE = 0.07, t(485.32) = -1.29, p = .198), but there was a main effect of Time (b = -0.12, SE = 0.02, t(1162.49) = -5.25, p < .001, β = -0.08, 95% CI [-0.11, -0.05]), indicating that loneliness decreased over time. This effect was qualified by a significant Condition × Time interaction (b = -0.05, SE = 0.02, t(1162.49) = -2.06, p = .040, β = -0.03, 95% CI [-0.06, 0.00]).

Post-hoc analyses revealed no significant between-group comparisons for loneliness at Times 1-3 (all ps > .218), but the treatment group reported marginally lower loneliness at Time 4 (b = -0.15, SE = 0.09, t(348) = -1.69, p = .092, d = 0.18). Within-condition analyses indicated that the decline in loneliness was even steeper among those in the treatment condition (b = -0.17, SE = 0.03, t(577.31) = -5.23, p < .001, β = -0.11, 95% CI [-0.16, -0.07]) compared to those in the control condition (b = -0.07, SE = 0.03, t(585.22) = -2.23, p = .026, β = -0.05, 95% CI [-0.09, -0.01]). These results suggest that the intervention accelerated the reduction of loneliness.

Belonging

There was no main effect of Condition (b = 0.11, SE = 0.09, t(487.33) = 1.13, p = .258), but there was a main effect of Time (b = 0.09, SE = 0.03, t(369.36) = 3.23, p = .001, β = 0.06, 95% CI [0.02, 0.09]), indicating that belonging increased over time. This effect was qualified by a significant Condition × Time interaction (b = 0.06, SE = 0.03, t(371.49) = 2.15, p = .032, β = 0.04, 95% CI [0.00, 0.07]).

Post-hoc analyses revealed that while there was no significant difference in belonging between conditions at baseline (as expected with random assignment; b = 0.02, SE = 0.10, t(482) = 0.17, p = .866), the treatment group exhibited marginally greater belonging compared to the control group at Week 6 (b = 0.21, SE = 0.12, t(348) = 1.85, p = .066, d = 0.20). Within-condition analyses indicated that belonging significantly increased over time in the treatment condition (b = 0.14, SE = 0.04, t(185.33) = 3.66, p < .001, β = 0.10, 95% CI [0.04, 0.15]), whereas no change was observed in the control condition (b = 0.03, SE = 0.04, t(183.93) = 0.79, p = .430). These findings suggest that the treatment enhanced belonging over time relative to the control condition.

Closeness to Academic Institution

There was no main effect of Condition (b = 0.07, SE = 0.06, t(490.08) = 1.14, p = .257), but there was a significant main effect of Time (b = 0.05, SE = 0.02, t(389.19) = 2.15, p = .032, β = 0.05, 95% CI [0.00, 0.09]), indicating that people felt closer to their academic institutions over time. This effect was qualified by a marginally significant Condition × Time interaction (b = 0.04, SE = 0.02, t(392.08) = 1.80, p = .072, β = 0.04).

Post-hoc analyses revealed that there were no significant differences between the conditions at baseline (b = 0.006, SE = 0.06, t(482) = 0.10, p = .924) nor at Week 6 (b = 0.11, SE = 0.08, t(348) = 1.40, p = .162). Within-condition analyses indicated that closeness to one's academic institution significantly increased over time in the treatment condition (b = 0.09, SE = 0.03, t(195.79) = 2.78, p = .006, β = 0.09, 95% CI [0.03, 0.15]), whereas no change was observed in the control condition (b = 0.006, SE = 0.03, t(195.71) = 0.20, p = .845). These findings suggest that the treatment enhanced institutional closeness over time relative to the control condition.

Resilience

There was no main effect of Condition (b = -0.01, SE = 0.19, t(491.88) = -0.07, p = .945), and no main effect of Time (b = 0.09, SE = 0.06, t(374.67) = 1.57, p = .118). However, there was a significant Condition × Time interaction (b = 0.12, SE = 0.06, t(374.67) = 2.11, p = .036, β = 0.04, 95% CI [0.00, 0.08]).

Post-hoc analyses revealed that there were no significant differences between the conditions at baseline (b = -0.19, SE = 0.21, t(481) = -0.89, p = .372) nor at Week 6 (b = 0.29, SE = 0.23, t(348) = 1.24, p = .216). Within-condition analyses indicated that resilience significantly increased over time in the treatment condition (b = 0.21, SE = 0.08, t(188.80) = 2.56, p = .011, β = 0.07, 95% CI [0.02, 0.12]), whereas no change was observed in the control condition (b = -0.03, SE = 0.08, t(185.65) = -0.36, p = .718). These findings suggest that the treatment enhanced resilience over time, relative to the control condition.

Mindfulness

There was no main effect of Condition (b = 0.21, SE = 0.25, t(487.02) = 0.83, p = .405), but there was a significant main effect of Time (b = -0.35, SE = 0.09, t(387.48) = -3.76, p < .001, β = -

0.09, 95% CI [-0.13, -0.04]), indicating that mindfulness decreased over time. This effect was qualified by a significant Condition × Time interaction (b = 0.20, SE = 0.09, t(387.48) = 2.10, p = .037, β = 0.05, 95% CI [0.00, 0.09]).

Post-hoc analyses revealed that while there was no significant difference in mindfulness between the conditions at baseline (as expected with random assignment; b = -0.09, SE = 0.27, t(482) = -0.32, p = .747), at Week 6, the treatment group showed greater mindfulness (b = 0.66, SE = 0.33, t(348) = 2.04, p = .042, d = 0.22, 95% CI [0.01, 0.43]). Within-condition analyses revealed that mindfulness remained stable across time in the treatment condition (b = -0.15, SE = 0.14, t(192.82) = -1.05, p = .297), whereas it declined in the control group (b = -0.55, SE = 0.12, t(193.13) = -4.50, p < .001, β = -0.14, 95% CI [-0.19, -0.08]). These findings suggest that the treatment buffered against the decline in mindfulness observed in the control condition.

Flourishing

There was no main effect of Condition (b = 0.13, SE = 0.28, t(479.97) = 0.47, p = .640), nor a main effect of Time (b = -0.03, SE = 0.09, t(366.08) = -0.39, p = .696). However, there was a significant Condition × Time interaction (b = 0.19, SE = 0.09, t(368.66) = 2.18, p = .030, β = 0.04, 95% CI [0.00, 0.08]).

Post-hoc analyses revealed that there were no significant differences between the conditions at baseline (b = -0.19, SE = 0.30, t(482) = -0.64, p = .522) nor at Week 6 (b = 0.29, SE = 0.35, t(347) = 0.83, p = .410). Within-condition analyses indicated that flourishing scores remained stable in the treatment condition (b = 0.16, SE = 0.11, t(184.95) = 1.39, p = .166), whereas the control condition showed a marginal decline (b = -0.23, SE = 0.13, t(181.40) = -1.72, p = .087, β

= -0.05). These findings suggest that the treatment buffered against the decline in flourishing observed in the control condition.

Depression

There were no main effects of Condition (b = 0.04, SE = 0.07, t(1085.99) = 0.52, p = .604), nor main effect of Time (b = 0.02, SE = 0.02, t(1150.81) = 0.76, p = .445). There was also no Condition × Time interaction (b = -0.03, SE = 0.02, t(1155.62) = -1.59, p = .113).

Anxiety

There was a marginal effect of Condition (b = -0.12, SE = 0.07, t(478.63) = -1.81, p = .071, β = 0.10), indicating that overall, those in the treatment group reported lower anxiety than those in the control group. There was no main effect of Time (b = -0.05, SE = 0.02, t(1163.46) = -1.92, p = .056). There was also no Condition × Time interaction (b = -0.02, SE = 0.02, t(1163.46) = -0.73, p = .468).

Stress

There was no main effect of Condition (b = -0.08, SE = 0.11, t(469.98) = -0.73, p = .464), but there was a marginal effect of Time (b = -0.08, SE = 0.04, t(1163.19) = -1.88, p = .060, β = -0.03), indicating that stress marginally decreased over time. There was no Condition × Time interaction (b = -0.03, SE = 0.04, t(1163.19) = -0.73, p = .464).

Academic Self-Efficacy

There was no main effect of Condition (b = -0.002, SE = 0.17, t(484.89) = -0.01, p = .989), but there was a significant main effect of Time (b = 0.15, SE = 0.07, t(387.82) = 2.33, p = .020), indicating that academic self-efficacy increased over time. There was no Condition × Time interaction (b = 0.05, SE = 0.07, t(387.82) = 0.77, p = .443).

Table S4

Means and Standard Deviations of Study Variables by Time and Condition

Measure	Condition	T1 (Baseline), M (SD)	T2 (Week 2), M (SD)	T3 (Week 4), M (SD)	T4 (Week 6), M (SD)
Emotional Well-Being					
Positive Affect	Control	18.48 (6.58)	17.35 (6.85)	17.43 (7.30)	17.21 (6.98)
	Treatment	18.93 (6.90)	18.51 (6.65)	19.11 (6.76)	19.38 (6.21)
Calm	Control	5.54 (2.44)	5.36 (2.57)	5.47 (2.58)	5.44 (2.58)
	Treatment	5.72 (2.69)	5.79 (2.59)	6.05 (2.66)	6.28 (2.41)
Well-being	Control	7.03 (2.46)	6.70 (2.47)	6.57 (2.73)	6.51 (2.56)
	Treatment	7.24 (2.53)	6.85 (2.35)	7.16 (2.43)	7.25 (2.34)
Vigour	Control	5.91 (2.76)	5.30 (2.69)	5.39 (2.92)	5.26 (2.88)
	Treatment	5.97 (2.67)	5.88 (2.61)	5.92 (2.61)	5.88 (2.59)
Negative Affect	Control	13.35 (6.72)	12.48 (6.54)	12.16 (6.65)	12.85 (7.05)
	Treatment	12.84 (6.83)	12.13 (6.61)	11.84 (7.18)	12.29 (6.62)
Depression	Control	4.37 (2.99)	4.07 (2.89)	3.98 (2.96)	4.11 (3.01)
	Treatment	3.99 (3.05)	3.75 (2.79)	3.66 (2.99)	3.82 (2.65)
Anxiety	Control	6.48 (3.00)	6.06 (2.86)	5.85 (2.79)	6.22 (2.90)
	Treatment	6.21 (2.89)	5.78 (2.78)	5.49 (2.67)	5.80 (2.86)
Anger	Control	2.50 (2.29)	2.35 (2.24)	2.33 (2.36)	2.52 (2.58)
	Treatment	2.64 (2.56)	2.60 (2.64)	2.69 (2.85)	2.70 (2.61)
Social Well-Being					
Loneliness	Control	5.60 (1.64)	5.28 (1.60)	5.16 (1.70)	5.40 (1.71)
	Treatment	5.58 (1.65)	5.08 (1.58)	4.98 (1.66)	5.10 (1.67)
Belonging	Control	5.61 (2.07)	—	—	5.68 (2.18)
	Treatment	5.64 (2.22)	—	—	6.11 (2.14)
Closeness to Academic Institution	Control	3.24 (1.40)	—	—	3.31 (1.42)
	Treatment	3.25 (1.39)	—	—	3.52 (1.44)
Overall Well-Being					
Resilience	Control	19.04 (4.61)	—	—	18.73 (4.56)
	Treatment	18.67 (4.48)	—	—	19.31 (4.22)
Mindfulness	Control	15.41 (5.82)	—	—	13.67 (6.05)
	Treatment	15.23 (6.12)	—	—	14.99 (6.11)
Flourishing	Control	44.97 (6.58)	—	—	44.34 (6.62)
	Treatment	44.59 (6.45)	—	—	44.92 (6.59)

Note. Dashes indicate measures not administered at that timepoint.
Source: Authors' analyses of data from the present study.

Section 7: Sensitivity Analyses Controlling for Demographic Covariates

Although participants were randomly assigned to condition and groups did not differ significantly on demographic characteristics at baseline, we reran all Condition × Time interaction models with age, sex, and ethnicity included as additional fixed effects. Results were strikingly robust to the inclusion of these covariates: effect sizes, standard errors, and p-values were nearly identical to the unadjusted models across all eleven outcomes, with no change in the direction or significance of any interaction term. Covariate-adjusted estimates are reported in Table S6.

Table S6

Condition × Time Interaction Results Controlling for Demographic Covariates

Outcome Variable	Condition × Time (*b*)	*SE*	*df*	*t*	*p*-value	*β*
Positive Affect	0.22	0.10	1149.56	2.30	.022	0.04
Calm	0.10	0.04	1170.62	2.49	.013	0.04
Well-Being	0.08	0.04	1158.29	2.19	.028	0.04
Vigour	0.05	0.04	1152.40	1.25	.212	–
Negative Affect	0.01	0.10	1147.88	0.14	.886	–
Loneliness	-0.05	0.02	1156.98	-2.05	.041	-0.03
Belonging	0.06	0.03	370.45	2.20	.029	0.04
Closeness to Academic Institution	0.04	0.02	391.52	1.86	.063	0.04
Resilience	0.12	0.06	376.67	2.15	.034	0.04
Mindfulness	0.20	0.09	387.95	2.11	.036	0.05
Flourishing	0.19	0.09	367.40	2.19	.029	0.04

Note. All models include age, sex, and ethnicity as additional fixed effects. Coefficients (*b*) reflect the Condition × Time interaction term from linear mixed-effects models.

Source: Authors' analyses of data from the present study.

Section 8: Exploratory Moderation Analyses

To explore whether intervention effects were moderated by demographic characteristics or baseline mental health status, and whether participant-transcribed app engagement metrics were associated with outcomes among intervention participants, we conducted three sets of exploratory analyses. For demographic and baseline mental health moderators, we fit linear mixed-effects models using the lmer() function. These models included fixed effects for the Condition × Time × Moderator interaction and random intercepts for participant and institution. Seven dependent variables were tested: positive affect, loneliness, belonging, institutional closeness, resilience, mindfulness, and flourishing.

Demographic Moderators

We examined five demographic moderators: age, sex, international student status, self-reported socioeconomic status (SES), and ethnicity. Each moderator was tested in a separate model as a three-way interaction with condition and time (Condition × Time × Moderator), predicting each of the seven primary outcome variables.

Across these 35 models, only one three-way interaction was significant at the uncorrected $p < .05$ level: age moderated the effect of the intervention on positive affect, such that younger students showed greater gains in positive affect over time ($b = –0.05$, $SE = 0.03$, $t(1158.99) = –2.08$, $p = .038$) —possibly because they were newer to the college environment and more in need of support. All other interaction terms had $p \geq .069$. No effects survived Bonferroni correction for multiple comparisons (adjusted $\alpha = .0014$).

Mental Health Moderators

We also tested whether baseline levels of depression, anxiety, loneliness, or perceived stress moderated intervention effects. Each moderator was tested in a separate model as a three-way interaction with condition and time (Condition × Time × Moderator), predicting each of the seven outcome variables, resulting in 28 total models.

Two effects were significant at the uncorrected $p < .05$ level. On one hand, baseline anxiety moderated changes in belonging: participants with higher initial anxiety experienced greater increases in belonging over time ($b = 0.03$, $SE = 0.02$, $t(373.29) = 2.20$, $p = .028$). These findings align with prior work suggesting that generative AI and positive psychology interventions yield greater benefits for those facing greater initial challenges[13–15]. At the same time, baseline depression moderated changes in flourishing: participants with lower initial depression scores showed greater improvements in flourishing ($b = –0.15$, $SE = 0.06$, $t(391.42) = –2.39$, $p = .018$), consistent with prior evidence that those already functioning relatively well may experience amplifying effects[16,17]. All other interaction terms were non-significant ($p \geq .153$). These patterns suggest strengths-based tools may have both compensatory and enhancing effects. However, given that no effects survived Bonferroni correction for multiple comparisons (adjusted $\alpha = .0018$), these results should be interpreted as exploratory, as with the demographic moderators.

App Engagement

We also conducted exploratory analyses examining whether participant-transcribed app engagement metrics were associated with outcomes among participants in the Flourish condition. Because engagement was only meaningful for those assigned to the intervention, these analyses were restricted to the treatment group. Unlike the moderator analyses above, these models did

not use a Condition × Time × Moderator mixed-effects framework. Instead, engagement was indexed by cumulative active days of app use, and each outcome was tested in a separate linear regression model predicting Week 6 outcomes from baseline levels of the same outcome and cumulative active days, while controlling for institution, resulting in seven total models.

One effect was significant at the uncorrected $p < .05$ level: greater cumulative active days was associated with lower flourishing at Week 6 ($b = –0.036$, $SE = 0.016$, $t = –2.28$, $p = .024$). Positive affect showed a marginal association in the opposite direction, such that greater cumulative active days was associated with higher Week 6 positive affect ($b = 0.034$, $SE = 0.018$, $t = 1.92$, $p = .056$). All other associations were non-significant ($p \geq .282$). However, given that no effects survived Bonferroni correction for multiple comparisons (adjusted $\alpha = .0071$), these results should be interpreted as exploratory.

Section 9: Detailed Results for Subcomponents of Positive Affect

Calm Subcomponent

For the calm subcomponent of positive affect, there was a significant effect of Condition (b = 0.22, SE = 0.10, t(472.38) = 2.26, p = .025, β = 0.19, 95% CI [0.07, 0.31]), indicating that those in the treatment condition reported greater calm. In addition, there was a significant main effect of Time (b = 0.11, SE = 0.04, t(1173.98) = 2.72, p = .007, β = 0.05, 95% CI [0.01, 0.08]), suggesting that calm affect increased over time. Finally, there was also a significant Condition × Time interaction (b = 0.10, SE = 0.04, t(1173.98) = 2.49, p = .013, β = 0.04, 95% CI [0.01, 0.08]).

Post-hoc analyses revealed no significant condition differences at Time 1 (b = 0.09, SE = 0.12, t(482) = 0.75, p = .457), but the treatment group reported marginally greater calm at Time 2 (b = 0.22, SE = 0.13, t(390) = 1.65, p = .099, d = 0.17), and significantly greater calm at Time 3 (b = 0.29, SE = 0.14, t(354) = 2.10, p = .036, d = 0.22, 95% CI [0.01, 0.43]) and Time 4 (b = 0.42, SE = 0.13, t(348) = 3.14, p = .002, d = 0.34, 95% CI [0.12, 0.55]). Within-condition analyses indicated that calm significantly increased over time in the treatment condition (b = 0.21, SE = 0.06, t(580.03) = 3.63, p < .001, β = 0.09, 95% CI [0.04, 0.14]), whereas it showed no significant change in the control group (b = 0.009, SE = 0.06, t(593.86) = 0.17, p = .866). These findings suggest that the treatment increased calm over time relative to the control condition.

Well-Being Subcomponent

For the well-being subcomponent of positive affect, there was a marginally significant effect of Condition (b = 0.19, SE = 0.10, t(470.54) = 1.92, p = .056, β = 0.17), indicating that those in the treatment condition reported greater well-being. There was no main effect of Time (b = -0.06, SE = 0.04, t(1157.49) = -1.55, p = .121). However, there was a significant Condition × Time interaction (b = 0.08, SE = 0.04, t(1162.58) = 2.25, p = .025, β = 0.04, 95% CI [0.00, 0.07]).

Post-hoc analyses revealed no significant condition differences at Time 1 ($b = 0.11$, $SE = 0.11$, $t(482) = 0.93$, $p = .352$), nor at Time 2 ($b = 0.08$, $SE = 0.12$, $t(390) = 0.62$, $p = .535$), but the treatment group reported significantly greater well-being at both Time 3 ($b = 0.30$, $SE = 0.14$, $t(353) = 2.16$, $p = .032$, $d = 0.23$, 95% CI [0.02, 0.44]) and Time 4 ($b = 0.37$, $SE = 0.13$, $t(348) = 2.80$, $p = .005$, $d = 0.30$, 95% CI [0.09, 0.51]). Within-condition analyses revealed that well-being remained stable across time in the treatment condition ($b = 0.02$, $SE = 0.05$, $t(575.72) = 0.47$, $p = .641$), whereas it declined in the control group ($b = -0.14$, $SE = 0.05$, $t(581.16) = -2.72$, $p = .007$, $\beta = -0.06$, 95% CI [-0.11, -0.02]). These findings suggest that the treatment buffered against the decline in well-being observed in the control condition.

Vigour Subcomponent

There was no main effect of Condition ($b = 0.14$, $SE = 0.11$, $t(471.73) = 1.32$, $p = .187$), but there was a marginal main effect of Time ($b = -0.07$, $SE = 0.04$, $t(1152.06) = -1.89$, $p = .059$, $\beta = -0.03$), indicating that vigour decreased over time. There was, however, no Condition × Time interaction ($b = 0.05$, $SE = 0.04$, $t(1156.75) = 1.25$, $p = .212$).

Section 10: Open-Ended Response Coding Rubric and Results

In this section, we first summarize the overall pattern of findings from participants' open-ended responses and then describe the coding rubric used to analyze those responses, along with the distribution of codes across response categories (see Table S7 for representative responses).

Two independent coders conducted a content analysis of participants' open-ended responses after Week 6. Responses to the question, "What was it like using the Flourish app?", suggested meaningful engagement with and perceived benefit from the intervention. Many participants spontaneously described improvements in overall well-being (58%; κ = .63, agreement = 88%), emotional well-being (24%; κ = .82, agreement = 93%), and social well-being (10%; κ = .79, agreement = 95%). Some participants also attributed benefits to specific app features, including the chatbot (22%; κ = .87, agreement = 95%), activities (8%; κ = .78, agreement = 96%), and weekly insights (1%; κ = .66, agreement = 99%), while others provided broader positive feedback (e.g., "Using the Flourish app has helped me to gain clarity and move forward on a path to recovery and evolution of myself with ease"). These results indicate meaningful engagement with and benefit from the app.

Note: All codes are binary (1 = present, 0 = absent).

Sheet 1: experience_flourish

Responses to the question: "What was it like using the Flourish app?"

Evaluation of Experience

Column	Evaluation	Examples	Percentage
Positive Experience	Overall judgment is clearly favorable	"interesting"	83% (κ = .79, agreement = 97%)
Mixed Experience	Response contains *both* pros & cons, or unsure	"*X* was good but I didn't like *Y*", "it was okay"	9% (κ = .79, agreement = 94%)

App Components

Column	Component referenced	Examples	Percentage
good UI	General interface qualities (e.g., simplicity, navigation, aesthetics, music)	"clean look", "easy to navigate"	6% (κ = .60[1], agreement = 94%)
1_chatbot_support	Chatbot/Sunnie or AI buddy conversations	"interacting with Sunnie was fun"	22% (κ = .87,

[1]Due to low inter-rater agreement of $\kappa < .60$, these codes were retained in the supplement for transparency but are not discussed in the main manuscript and should be interpreted with caution.

Column	Component referenced	Examples	Percentage
			agreement = 95%)
2_activities	Specific activities (e.g., breathing, Pomodoro, mood check-in)	"breathing exercise relaxed me"	8% (κ = .78, agreement = 96%)
3_weekly_insights	Weekly insights	"weekly insights were eye-opening"	1% (κ = .66, agreement = 99%)

Explicitness rule: Mark 1 only when the component is *named or unmistakably described*; inferring is not enough (e.g., "insightful patterns" ≠ weekly insights).

Benefits

Column	Benefit domain	Examples	Percentage
emotional	Emotional benefits, such as downregulating negative or upregulating positive states (includes relaxation, de-stressing)	"helped me chill", "reduced stress"	24% (κ = .82, agreement = 93%)
social	Social benefits, such as greater connection to others, sense of community, feeling less lonely	"felt less alone", "shared with friends"	10% (κ = .79, agreement = 95%)
overall well-being	Broader gains: focus, productivity, self-insight, therapeutic value	"improved focus", "felt like having a 24/7 therapist"	58% (κ = .63, agreement = 88%)

Additional rule: benefits are mutually exclusive. In other words, if emotional or social benefits are cited, and no broader benefit is cited, overall well-being is 0.

Sheet 2: recommend_flourish

Responses to the question: "Do you think this would be a useful thing that your university should offer as a wellness tool for students? Why or why not? Please explain."

Adoption Stance

Column	Adoption Stance	Examples	Percentage
Yes	Clear endorsement of university adoption	"should definitely use it", "I recommend it"	77% (κ = .76, agreement = 91%)

Column	Adoption Stance	Examples	Percentage
No	Clear rejection of university adoption	"not worth it", "shouldn't use"	2% (κ = .76, agreement = 99%)
Conditional/ Qualified	Adoption contingent on caveats or improvement, or unsure	"only if it's free", "depends on student needs", "maybe"	20% (κ = .76, agreement = 91%)

Only one of the three columns should be 1; choose the clearest stance. If there is an empty response (e.g., "na"), all columns are 0.

Sheet 3: feedback_flourish

Responses to the question: "What would you change about the app, if anything?"

Suggestion Presence

Column	Suggestion Presence	Examples	Percentage
No Suggestions/ Positive	No suggestions for improvement, or only praise	"none", "it's perfect", "n/a"	52% (κ = .97, agreement = 98%)

If No Suggestions/Positive = 1, leave all other suggestion-type columns 0.

Type of Suggestion

Column	Type of Suggestion	Examples	Percentage
UI	Feedback related to interface/layout (e.g., navigation, colors, music)	"make layout simpler", "add dark mode"	7% (κ = .57[1], agreement = 92%)
chatbot	Feedback related to chatbot	"Sunnie responses were slow"	17% (κ = .78, agreement = 93%)
activities	Feedback on Flourish activities	"add more breathing exercises"	16% (κ = .92, agreement = 98%)
other	Feedback that does not fit UI / chatbot / activities	"make privacy policy more visible"	6% (κ = .41[1], agreement = 88%)

Table S7
Example Open-Ended Participant Responses

Code Category	Example Responses
Evaluation of Experience	
Positive (83%)	"Using the Flourish app was a calming and positive experience. It felt like a simple way to stay mindful and check in with myself daily. The reminders were helpful without feeling overwhelming, and it became a nice part of my routine." "It was really interesting learning about ways I could help myself. The app helped comfort me multiple times."
Mixed (9%)	"Sometimes I was glad that, I had it, other times it felt a bit like a chore." "The app was helpful when I needed to be grounded, however sometimes I found it not as helpful for other aspects."
App Components	
Chatbot (22%)	"It was very helpful, and it was nice that Sunnie would remember and mention our past conversations." "I was very impressed at how well [Sunnie] was able comprehend and give real answers and terms a therapist would use when asked deep questions. It would call back and ask follow-up questions about things I told it 2 weeks prior."
Activities (8%)	"I only used the flourish app for the daily mood check in, as well as the breathing exercise. I also used it for staying focused. I really liked that feature because it made me focus on my work." "It was nice to have an app to have guided activities for self care"
Weekly Insights (1%)	"I found it easy to use, fascinating, and useful for tracking my wellness behaviors. It offered a novel approach to visualizing health trends."
Benefits	

Emotional (24%)	"I really enjoyed using the Flourish app! I think that it works well and helps me organize my thoughts and emotions and gives accurate advice to recenter." "It was beneficial to help track mood and encourage improvements."
Social (10%)	"It was fun and made me connect a lot with myself and with other people that I don't necessarily speak to on that level all of the time." "I also like how the chatbot is available as a source of companionship and also advice."
Overall Well-Being (58%)	"The app was able to empathize through text and quickly bring up solutions to any situation I posed - emotional or non-emotional." "Comforting knowing I had something to build a routine off of and almost a relationship with. It helped motivate me to engage every day."
Adoption Stance	
Yes (77%)	"Yes, I learned so much from the app; it does a great job of listening. Also, it's a great way to get support between therapy sessions." "Yes, I think wellness tools such as the flourish app can be a great mental health resource for students that does not involve a larger effort to reach out to - rather than workshops or in-person appointments"
No (2%)	"No" "Not particularly, I felt that a lot of the advice I got I already knew about."
Conditional / Qualified (20%)	"I think it could be useful, but I don't know that many students would actually make use of it as a wellness tool" "Yes, although I think it depends on the student. Some might not find it helpful, but I could see some students using it as wellness tool."
Suggestion Presence	

No Suggestions / Positive (52%)	"Nothing, I like it." "I don't think there was anything that needed to be changed."
Type of Suggestion	
Chatbot (17%)	"It would help to have some open-ended prompt examples for the questions asked when chatting with Sunnie." "I would make Sunnie more willing to challenge us like a normal therapist would and actual say what the problem might be."
Activities (16%)	"I think more options for different mindfulness activities would be useful and make the app more engaging." "I would add some other tools under "less stress" in the toolkit section. I think that that was the most helpful section for me, and therefore I would utilize more tools."

Note. Responses were coded using a structured binary rubric (1 = present, 0 = absent). Interrater reliability across codes was strong ($\kappa = .63–.97$). Codes not meeting $\kappa \geq .60$ are not included in this table but are documented in the coding rubric for transparency.
Source: Authors' coding and selection of deidentified participant responses from the present study.

References

1. Lieberman, M. D. *et al.* Affect labeling disrupts amygdala activity in response to affective stimuli. *Psychol. Sci.* **18**, 421–428 (2007).
2. Millgram, Y., Nock, M. K., Bailey, D. D. & Goldenberg, A. Knowledge about the source of emotion predicts emotion-regulation attempts, strategies, and perceived emotion-regulation success. *Psychol. Sci.* **34**, 1244–1255 (2023).
3. Torre, J. B. & Lieberman, M. D. Putting feelings into words: Affect labeling as implicit emotion regulation. *Emot. Rev.* **10**, 116–124 (2018).
4. Rippstein-Leuenberger, K., Mauthner, O., Sexton, J. B. & Schwendimann, R. A qualitative analysis of the Three Good Things intervention in healthcare workers. *BMJ Open* **7**, e015826 (2017).
5. Zhao, X. & Epley, N. Insufficiently complimentary?: Underestimating the positive impact of compliments creates a barrier to expressing them. *J. Pers. Soc. Psychol.* **121**, 239–254 (2021).
6. Zhao, X. & Epley, N. Kind words do not become tired words: Undervaluing the positive impact of frequent compliments. *Self Identity* **20**, 25–46 (2021).
7. Aron, A., Melinat, E., Aron, E. N., Vallone, R. D. & Bator, R. J. The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Pers. Soc. Psychol. Bull.* **23**, 363–377 (1997).
8. Epley, N., Kardas, M., Zhao, X., Atir, S. & Schroeder, J. Undersociality: Miscalibrated social cognition can inhibit social connection. *Trends Cogn. Sci.* **26**, 406–418 (2022).
9. Lee, I. M. & Buchner, D. M. The importance of walking to public health. *Med. Sci. Sports Exerc.* **40**, S512–S518 (2008).

10. Ma, J., Lin, P. & Williams, J. Effectiveness of nature-based walking interventions in improving mental health in adults: a systematic review. *Curr. Psychol.* **43**, 9521–9539 (2024).

11. Nisbet, E. K., Zelenski, J. M. & Grandpierre, Z. Mindfulness in nature enhances connectedness and mood. *Ecopsychology* **11**, 81–91 (2019).

12. Seligman, M. E. P. *Flourish: A Visionary New Understanding of Happiness and Well-Being*. (Simon and Schuster, 2011).

13. Brynjolfsson, E., Li, D. & Raymond, L. Generative AI at work. *Q. J. Econ.* **qjae044**, (2025).

14. Noy, S. & Zhang, W. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* **381**, 187–192 (2023).

15. Peng, S., Kalliamvakou, E., Cihon, P. & Demirer, M. The impact of AI on developer productivity: Evidence from github copilot. Preprint at https://arxiv.org/abs/2302.06590 (2023).

16. Lyubomirsky, S. & Layous, K. How do simple positive activities increase well-being? *Curr. Dir. Psychol. Sci.* **22**, 57–62 (2013).

17. Sin, N. L., Della Porta, M. D. & Lyubomirsky, S. Tailoring positive psychology interventions to treat depressed individuals. in *Applied positive psychology: Improving everyday life, health, schools, work, and society* (eds Donaldson, S. I., Csikszentmihalyi, M. & Nakamura, J.) 79–96 (Routledge, New York, NY, 2011).