Original research article

DIGITAL HEALTH

Generative AI versus human conversational agent for reducing procrastination: A randomized pilot trial

DIGITAL HEALTH
Volume 12: 1–16

DOI: 10.1177/20552076261450336
journals.sagepub.com/home/dhj
Sage

Severin Hennemann[1,2], Julia M. Fähnrich[1], Cornelius Tietze[1], Michael Witthöft[3], and Stefanie M. Jungmann[4]

Abstract

Objectives: The use of generative artificial intelligence (AI) based chatbots in mental health interventions is rapidly evolving, but evidence of their efficacy and safety is scarce. This randomized pilot trial investigated the feasibility, efficacy, mechanisms, and unwanted effects of a generative AI chatbot (ChatGPT-4) vs. human conversational agent as a cognitive behavioral micro-intervention for procrastination. **Methods:** 62 university students (M_{age} 23.60 years, 93.5% female) were analyzed and assigned (concealed) to a three-session chat-based intervention with either the chatbot or human conversional agent. Outcomes were assessed using questionnaires at baseline, post-intervention, and 6-month follow-up. The primary outcome was self-reported procrastination. Secondary outcomes included depression, anxiety, worry, and rumination. Additionally, unwanted negative effects and common factors of therapeutic change were assessed. **Results:** Intention-to-treat analyses revealed a non-significant group × time interaction effect ($p = .117$, $\eta^2_p = 0.05$). Unwanted negative effects at post-assessment were reported by 43.1% of participants, with no significant group difference. Common factors were rated similarly across groups and did not mediate change in procrastination. **Conclusions:** Our findings do not indicate superiority of either conversational agent and both interventions had comparable, rather high rates of unwanted effects, indicating that more research in larger, clinically meaningful samples is needed to confirm the efficacy and safety of AI-based (micro-) interventions.

Keywords

artificial intelligence, chatbot, large language model, chatGPT, procrastination

Received: 1 January 2026; Revised: 17 April 2026; Accepted: 25 April 2026

Introduction

The vision of a fully automated, artificial intelligence (AI) based psychotherapist has long intrigued and perplexed researchers, clinicians, and the public. Automated conversational agents (CAs) are particularly promising for mental health care, where the need for scalable, accessible, and preventative treatment solutions remains critical, given the global burden of mental disorders[1,2] and their ongoing unmet treatment need.[3,4] Earlier rule-based CAs relied on predefined scripts, limiting

[1]Department of Clinical Psychology, Psychotherapy and Experimental Psychopathology, University of Mainz, Mainz, Germany
[2]Department of Clinical Psychology and Psychotherapy, Charlotte Fresenius Hochschule, Wiesbaden, Germany
[3]Department of Clinical Psychology and Psychotherapy, Ruhr-University of Bochum, Bochum, Germany
[4]Department of Clinical Psychology and Psychotherapy of Childhood and Adolescence, University of Mainz, Mainz, Germany

Corresponding author:
Severin Hennemann, Department of Clinical Psychology and Psychotherapy, Charlotte Fresenius Hochschule, Moritzstr. 17a, 65185 Wiesbaden , Germany.
Email: severin.hennemann@charlotte-fresenius-uni.de

their ability to respond flexibly to users.[5] In contrast, recent generative AI-based CAs powered by large language models (LLMs) enable more context-sensitive, adaptive, and personalized interactions. The growing body of research on chatbots for mental distress primarily focuses on rule-based CAs in reducing negative affect (e.g., depressive and anxious symptoms) in both clinical and non-clinical populations, with intervention duration ranging from minutes to months.[6,7] A recent meta-analysis of 15 trials found a moderate to large average effect (hedge's $g = 0.7$) for AI-based CAs in reducing psychological distress compared to various control conditions.[7] Notably, generative AI-based CAs demonstrated larger effects than rule-based CAs, suggesting that the increased flexibility and contextual adaptability of LLM-powered systems may enhance symptom reduction and user engagement. However, the overall quality of existing research remains low to moderate, with few randomized controlled trials (RCTs) and even fewer studies specifically investigating generative AI-based CAs.[8] Moreover, there is a significant gap in research on the safety and potential adverse effects of chatbots in mental health applications.

While most research on AI-driven interventions has focused on alleviating symptoms of depression and anxiety, there is growing interest in their potential for addressing behavioral and self-regulation difficulties. One such issue, which is both widespread and highly relevant in academic and professional settings, is procrastination. Procrastination is a work-related mental problem characterized by postponing or delaying tasks despite potentially negative consequences. Extensive or chronic procrastination has been linked to worse mental health (i.e., higher stress levels, depressive and anxiety symptoms), somatic symptom reporting and unhealthy lifestyle behaviors, delays in medical and mental health treatment, loneliness, economic difficulties such as lower income or higher unemployment rate, and reduced performance.[9–11] Around one in five adults in Western countries demonstrate problematic procrastination,[12] which is particularly prevalent in adolescents and emerging adults (i.e., age range around 14–29 years).[13] Around half of university students engage in problematic procrastination by delaying exam preparation or putting off writing tasks.[9,14,15] Motivational, learning, and cognitive theories of procrastination emphasize task-related (e.g., task aversiveness) and person-related factors (e.g., low motivation, impaired self-control, low self-efficacy, dysfunctional emotion regulation, self-handicapping, and fear of failure) as key determinants of procrastination.[16,17] The academic environment can be considered another important factor, given the high degree of autonomy and reduced external structure (as compared to childhood and adolescence), requiring strong self-regulation skills to manage academic responsibilities.[18] Negative reinforcement, in the form of short-term relief from negative affect by postponing an unfavorable course of action is considered a central mechanism sustaining procrastination processes.[16]

The high prevalence of procrastination and its far-reaching consequences underscore the need for effective interventions. A large meta-analysis of 24 RCTs in non-clinical samples by van Eerde and Klingsieck[19] found evidence for the general positive effects of psychological interventions. Among these, cognitive-behavioral therapy (CBT) has been identified as a particularly effective approach for treating severe procrastination.[16,19] CBT interventions for procrastination typically focus on improving time management, setting realistic goals, reducing avoidance behaviors and distractors, as well as fear of failure and perfectionist demands.[20] Notably, research indicates that short-term interventions can be as effective as longer treatments,[19] making them a valuable option for reaching a wide range of individuals in need of timely support. Digital interventions may be particularly well suited in this context, offering low-threshold, accessible, and scalable solutions. For example, Åsberg et al.[21] investigated a single-session intervention ("Focus") targeting procrastination in university students, consisting of tailored behavioral advice based on self-report assessment, which however, did not lead to change in procrastination compared to a control group without feedback. Amarnathet al.[22] investigated a guided CBT-based online-intervention ("GetStarted") including five main modules (i.e., psychoeducation, self-observation, negative thoughts, cognitive restructuring, relapse prevention) that yielded small effects after four weeks compared to a waitlist control. Rozental found no superiority of guided over unguided self-help in reducing procrastionation.[23] However, most digital interventions are typically highly standardized self-help approaches and thus rely heavily on users' self-regulatory capacities and adherence – both of which are core deficits in procrastination.[17] In contrast, generative AI-based CAs may provide activating, emotionally engaging support that directly targets self-regulation. Their dialogic nature can foster a dialectical perspective (e.g., short-term relief vs. long-term costs) and deliver personalized, actionable strategies with continuous guidance, potentially enhancing adherence and behavior change. For instance, Mutter et al.[24] conducted an RCT comparing a five-week CBT-based internet intervention for university students with more chronic and severe procrastination (as determined by a cut-off in the Irrational Procrastination Scale; IPS)[25] with two forms of coaching: an automated, rule-based chatbot and a human coach. Both conditions led to significant reductions in procrastination at post-assessment, with no significant differences between groups. These findings indicate that accompanying deterministic CAs can be as effective as human CAs in reducing procrastination in the context of brief interventions, at least in the short term. However, it remains unclear whether generative CAs, which can offer more adaptive and personalized support, might yield even greater benefits.

Research objectives

Therefore, this pilot study aimed to evaluate the feasibility and efficacy of a publicly available generative AI (genAI) CA, that is ChatGPT-4, in a cognitive-behavioral micro-intervention for adults with procrastination as compared to a human CA. As there is little and heterogeneous evidence on the difference in efficacy between AI-based and human-led CAs,[7] this study was explorative in nature and did not test directional or non-inferiority hypotheses. Firstly, we tested differences in efficacy with regard to reducing self-reported procrastination from baseline to post-assessment (primary endpoint) between ChatGPT-4 vs. human CA. Secondly, we examined differences in improvements in secondary outcomes (depression, anxiety, worrying, and rumination) and thelong-term effects from baseline to the 6-month follow-up in primary and secondary outcomes. Thirdly, we examined differences in common impact factors of psychotherapy (i.e., working alliance, problem actualization, and problem mastery) between both groups and if the treatment effect on the primary endpoint (post-assessment) is mediated by these factors. Fourthly, we explored the safety of both CAs by evaluating the type and frequency of any participant-reported unwanted negative intervention effects and investigated intervention usage and the effectiveness of blinding of participants.

Methods

Design and recruitment

The study was designed as a prospective, two-armed, single-blinded, randomized pilot trial with an active control group. The study protocol was approved by the local ethics committee of the Department of Psychology, University of Mainz, Germany (reference number 2023-JGU-psychEK-026, July 2023) and was pre-registered via the platform aspredicted.org (reference number #150,559; https://aspredicted.org/w5qc4.pdf, date: 2023/11/10). The study was not registered in a conventional clinical trial registry, as it was conducted as a pilot study in a non-clinical sample and did not involve a medical or clinical intervention. All procedures were conducted in accordance with the ethical standards of the Declaration of Helsinki. The Consolidated Standards of Reporting Trials (CONSORT) Checklist and the Chatbot Assessment Reporting Tool (CHART) are provided as supplementary files. Inclusion criteria were age above 18 years, sufficient knowledge of the German language, and informed consent. The exclusion criterion was ongoing or scheduled psychotherapy. The study was conducted between October 2023 and July 2024 and the recruitment period was three months. All study procedures were conducted online and there was no physical contact between the participants and the study team. Participants were recruited via various university e-mail distribution lists (e.g., of the Institute of Psychology of the University of Mainz), social media, and locally with flyers at the student counseling of the University of Mainz. Participants were reimbursed with a raffle of 5 x 15 EUR or course credit. An apriori power calculation for at least small between-group effects ($f \geq .15$) at $\beta = .80$ and $\alpha = .05$ in a two-factor repeated measures analysis of variance (time: pre vs. post; group: genAI vs. human CA) yielded a necessary sample size of 56 participants.

Participants

One hundred and seven participants started to fill out the online baseline assessment. Forty-three (40.1%) participants were excluded since they did not provide complete data at baseline, including ten participants (9.3%) who did not sign up for the first coaching session as mandatory part of the baseline questionnaire. Therefore, 64 participants were randomized. Since two participants did not provide a correct study code to identify longitudinal assessments, $n = 62$ participants were analyzed as a modified intention-to-treat (mITT) sample. The study flow can be found in Figure 1.

Procedures

Potential participants received a link to an online questionnaire (T0) via the survey platform https://www.soscisurvey.de/ with detailed information on the study and a check of eligibility criteria. Eligible participants gave written informed consent to participate and were asked to provide an email address and telephone number (as a security measure in case of suicidality or crisis intervention by a trained clinical psychologist of the study team), which were stored separately from the rest of the data. Other data were pseudonymized by using an individually generated code. After completion thereof, participants were randomly allocated to either receive the genAI-based CA or human CA. Randomization and allocation were performed by using a randomization list via https://www.randomizer.org/. The list was stored in a non-editable format and accessed only at the time of assigning participants after baseline completion. Investigators did not alter the sequence, and allocation occurred strictly in the order generated. The allocation ratio was 1:1. Participants were blinded to the study conditions. After allocation,

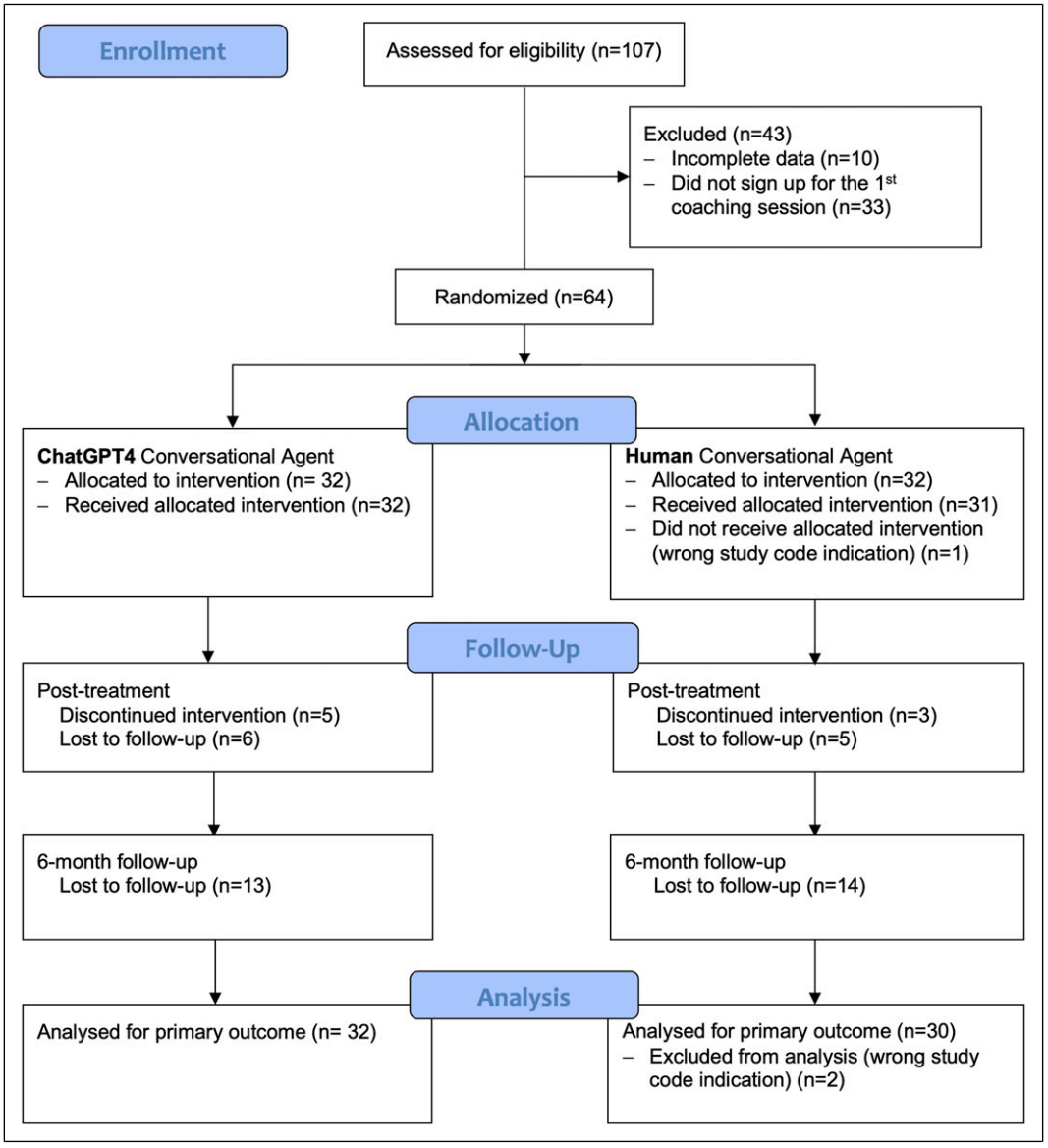

Figure 1. Participant flow.

participants initiated the three-session chat-session protocol. At the end of each session, participants were directed (a) to a calendar to select the date and time for the next intervention session and (b) to the online questionnaire.

Study arms

Both conditions included three approximately 20-minute sessions of chat sessions within the university-hosted platform BigBlueButton (BBB), version 2.7.2 (https://bbb.rlp.net/). In the human CA condition, study personnel (master students in clinical psychology) typed their answers directly into the chat, in the genAI condition, participants' input first was fed (anonymously) into the original ChatGPT platform (https://chatgpt.com/), and its answers were copied back into the ongoing chat in BBB to ensure blinding of participants regarding the condition. The rationale for the coaching was based on a cognitive-behavioral treatment rationale for severe procrastination by Höcker et al.[16] and is available in Supplements A and B. We selected three strategies in collaboration with experts on procrastination treatment in emerging adults, which are displayed in Table 1.

Each session included a standardized welcome and farewell message, which was typed in manually in both conditions. Each chat-coaching was supervised by the study team, to be able to intervene in a set of pre-defined cases: (a) invalidating

Table 1. Content of the micro-intervention.

Session	Topic	Content	Homework
1	Psychoeducation and behavioral analysis	Definition of procrastination (vs. postponing) Conditional model Behavioral analysis	Self-observation
2	Self-management strategies	Identification of disruptive factors using weekly planner Coping strategies (e.g., setting priorities, mitigating disruptions in a goal-oriented manner, efficient and realistic work schedules)	Implementation of self-management strategies
3	Cognitive modification and reward strategies	Identification of procrastination-promoting thoughts and attitudes Reality testing of dysfunctional cognitions, formulation of alternative motivating cognitions and affirmations Instructions for rewarding achievements of (partial) goals.	Provision of intervention summary

content production by ChatGPT-4 or participants, (b) answers that would have compromised blinding participants, and (c) acute crisis reaction of participants. In this case, participants would have been contacted by a licensed psychotherapist of the study team. Human corrections to the chat content in the genAI condition only occurred for reason (b), see Supplement C for details.

Human coaches

For organizational reasons, all participants were assigned one of two coaches, who were master students in clinical psychology. Both underwent a one-day training course on procrastination treatment beforehand and were instructed in the intervention rationale. Coaches had weekly to bi-weekly supervision with a licensed clinical psychologist of the study team. For the control group (human CA), we developed a coaching manual to ensure comparability to the ChatGPT prompts (see Supplement B). Coaches were instructed to adhere to the same predefined word limit (100 words) and timing (approximately 20 minutes).

Intervention group (genAI-based CA)

We developed prompts for each of the three coaching sessions with an iterative procedure,[26] including test trials with volunteers and content supervision by a clinical psychologist specializing in procrastination treatment. The intervention employed the paid model ChatGPT-4 (OpenAI, 2023, https://openai.com/de-DE/index/gpt-4-research/), without additional fine-tuning or model-specific adaptation. Prompts were developed specifically for this study based on CBT principles and the intervention manual by Höcker et al.[16] and were implemented using a mostly zero-shot prompting approach, i.e., without providing example dialogues or few-shot demonstrations. ChatGPT was prompted to adhere to a dialogic conversation, to write in an empathetic, valued, motivating way, and to take the role of a cognitive-behavioral-oriented coach. We set a word limit of 100 words for ChatGPT-4's output. Each session was structured into semi-detailed commands, for example, "Ask your conversation partner what he/she/they is procrastinating and what he/she/they thinks is the reason for this" followed by "Ask your conversation partner, how he/she/they would define procrastination, and based on the answer provide a definition yourself, focusing on the difference between postponing and procrastinating" (session 1, see also Table 1). Some prompts included "if-then" commands, that is, one-shot prompting, such as "If your conversation partner finds it difficult to develop helpful thoughts themselves, you can give them helpful examples. These can look like this…" (session 3, see also Table 1). The detailed prompts can be found in Supplement B. When approaching a session length of approximately 20 minutes, supervising coaches decided on their own to prompt the ending of the session.

Assessments and outcomes

All outcomes were self-assessed through online questionnaires via https://www.soscisurvey.de/. Primary and secondary outcomes were assessed at baseline (T0), post-intervention (T3), and six months after randomization (T4). After each coaching session (T1, T2, T3) participants filled out the Mainz Hourly Assessment Form (MSB).[27] For the current study, internal consistencies at baseline are reported using McDonald's Omega (ω).

Primary outcome

Procrastination was assessed with the German version of the General Procrastination Questionnaire (APROF).[16] The APROF consists of 18 items, that are rated on a seven-point rating scale from 1 = "never" to 7 = "always". The APROF has three subscales: procrastination tendency (e.g., "I delay completing certain important tasks."), task aversiveness (e.g., "I have to overcome my discomfort to start important tasks."), and alternative preference (e.g., "When I want to start an important task, other activities come to my mind."), which can be combined into a total mean score, ranging from 1 to 7. The internal consistency for each of the scales was high to excellent in student samples (procrastination tendency: $\alpha = .92$, task aversiveness: $\alpha = .89$, alternative preference: $\alpha = .85$).[28] In the current study, the internal consistency for the total score was $\omega = .97$.

Secondary outcomes

Depressive symptoms were assessed with the German version of the Patient Health Questionnaire-9 (PHQ-9),[29] which includes nine items that are rated on a 4-point scale from 0 = "not at all to" to 3 = "nearly every day". Total sum scores range between 0 and 27 points, with a proposed cut-off for screening for depression ≥10. The PHQ-9 has demonstrated high internal consistency ($\alpha = 0.93$) in a representative German sample. Internal consistency was $\omega = .92$ in our study. Anxiety was assessed with the German version of the Generalized Anxiety Disorder Questionnaire (GAD-7),[30] which includes seven items on symptoms of general anxiety within the last two weeks that are rated on a 4-point scale from 0 = "not at all" to 3 = "nearly every day". Total sum scores range from 0 to 21 points, with a proposed cut-off for screening for anxiety disorders ≥10. The German GAD-7 has demonstrated high internal consistency ($\alpha = .89$) in the general population.[31] Internal consistency was $\omega = .88$ in our study. Rumination was assessed with the German version of the subscale "Rumination" of the Rumination-Reflection-Questionnaire (RRQ),[32] including 12 items that are rated on a five-point answer scale from 1 = "strongly disagree" to 5 = "strongly agree". The mean scale score ranges from 1 to 5. Internal consistency of the subscale proved to be high in a previous study ($\alpha = .90$)[33] and was $\omega = .94$ in our study. Worrying was assessed with the German version of the Penn State Worry Questionnaires (PSWQ),[34] which includes 16 items that are rated on a five-point answer scale from 1 = "not at all typical" to 5 = "very typical of me". Total sum scores range from 16 to 80 and scores > 62 are considered indicative of a generalized anxiety disorder diagnosis. Internal consistency of the original German PSWQ proved to be high ($\alpha = .90$)[34] and was $\omega = .95$ in our study.

Additional measures

Demographic characteristics assessed at baseline included age, gender, the highest level of education, occupation, relationship status, living situation, mother tongue, and intake of psychopharmaceutics. Furthermore, participants were asked how often they have used ChatGPT so far and for what purpose. Other items assessed the general attitude towards AI technologies on a five-point scale from 1 = "very positive" to 5 = "very negative" and intervention expectancy on a 10-point answer scale from 0 = "no improvement" to 10 = "greatest conceivable improvement". At post-intervention, participants were asked to guess their allocated study condition (genAI vs. human CA).

Participant-rated negative effects that occurred as a consequence of the intervention were assessed at post-intervention and follow-up using the German version of the Negative Effects Questionnaire (NEQ).[35] The NEQ consists of 20 items covering symptoms, quality, dependency, stigma, and hopelessness. Participants first indicate the occurrence of a negative effect (yes/no), then rate the perceived burden on a five-point answer scale from 0 = "not at all" to 4 = "extremely" and then indicate whether they were caused by the treatment (yes/no). The original NEQ comprises six sub-scales with internal consistencies ranging from $\alpha = .72$ to .92. In our study, internal consistency was $\omega = .97$ (perceived burden rating for negative effects attributed to the intervention).

Common factors of therapeutic change were assessed with the MSB,[27] which includes 15 items that are answered on a seven-point answer scale from 1 = "not at all" to 7 = "strongly". The MSB has three sub-scales: therapy alliance, problem activation, and problem mastery, which form a general factor. The total sum score was calculated, ranging from 15 – 105. In psychotherapy outpatients, the internal consistencies for the sub-scales at various assessment points during treatment ranged from $\alpha = .80$ to .91 and from $\alpha = .91$ to $\alpha = .94$ for the total scale.[27] In our study, internal consistency ranged from $\omega = .89$ to $\omega = .93$ at T1.

Statistical analyses

Analyses were conducted with R (v. 4.5.0) and RStudio (v. 2024.12.1). Analyses of primary and secondary outcomes were conducted with the mITT sample. To detect small between-group effects ($f \geq .15$) at $\beta = .80$, $\alpha = .05$ in a two-factor repeated measures ANOVA (interaction effect; time: pre vs. post by group: ChatGPT-4 vs. human CA), a target sample size of at least $N = 56$ was calculated. However, given the pilot nature of this trial, analyses should be considered exploratory. To investigate

the intervention effect for the primary endpoint, a linear mixed model (LMM) was calculated with primary and secondary outcomes as dependent variables, fixed effects for condition (genAI vs. human CA), and measurement time (T0, T3) as well as random effects for the test subjects to account for within-subject dependencies. A *time* × *condition* interaction was included to assess whether the intervention effect differed across measurement time points. We chose LMMs instead of repeated measures ANOVA (as preregistered) because LMMs enable modeling of change trajectories using all available data, effectively handling missing values under the ITT principle.[36] To assess within-group effects for the condition variable, estimated marginal means were compared using Bonferroni-corrected pairwise tests. Other than pre-registered, we included an additional assessment point (6-months follow-up) assessment in our extended analyses, as its logistical feasibility was unclear during the thesis-based pre-registration phase. Thus, to investigate the long-term intervention effects, we repeated LMMs including all measurement points (T0, T3, T4) and within-group effects (T0 vs. T4). Model estimation was performed using the *lmerTest* package with restricted maximum likelihood. Predicted values were visualized using the *ggeffects* package, and descriptive statistics of observed values were reported for each time point and condition (see Table 3). Partial eta squared (η^2_p) was used to report the proportion of variance explained by the main and interaction effects in the model-based analysis. Within and between-group effect sizes were calculated as Cohen's *d* using observed data and including correlation between time points and standard deviation of the pretest for within-group effect sizes. To examine the mediating effect of common factors, a subgroup of participants who had taken part in at least one chat coaching session was selected ($n = 51$) and analyzed with a cross-lagged panel model. The model was estimated using full information maximum likelihood as well as robust standard errors and a scaled test statistic, as implemented in the *lavaan* package in R, to handle missing data (T2 was missing for $n = 3$ participants) and non-normality. The model included autoregressive effects for the APROF and the MSB, cross-lagged effects (common factors as a predictor for procrastination) as well as residual covariances between exogenous variables (condition, APROF at baseline). Finally, the type and frequency of self-reported side effects in the NEQ were compared between groups using χ^2 tests and independent *t*-tests for severity ratings.

Results

Participant characteristics

The demographic characteristics of the sample are listed in detail in Table 2. Participants were on average 23.60 years old ($SD = 3.84$), 93.5% ($n = 58$) were female, and all were university students, with 98.4% ($n = 61$) studying psychology. Around one-third (32.3%, $n = 20$) indicated using ChatGPT at least once a week, mainly for their studies (59.7%, $n = 37$). Half of the participants ($n = 37$, 59.7 %) indicated "positive" or "very positive" attitudes towards ChatGPT (highest response categories) and overall expected a moderate change in procrastination through the intervention ($M = 5.38$, $SD = 1.95$). Participants had a moderate average level of procrastination in the APROF at baseline and overall depression (PHQ-9) and anxiety (GAD-7) scores were below the clinically relevant cut-offs, see Table 3. Groups did not differ with regard to their baseline scores in primary and secondary outcomes, $ps > .05$. The loss-to-follow-up at post-assessment was 9.7% ($n = 6$), see also Figure 1. Across conditions, 12.9% ($n = 8$) of participants did not complete all three coachings (ChatGPT-4: 15.6 %, $n = 5$; human CA: 10%, $n = 3$), see also Figure 1.

Primary outcome

Table 3 presents the observed values of each outcome per time point, along with model-based interaction effects (pre-post) and corresponding effect sizes. mITT analyses of the APROF score in $n = 62$ participants using linear mixed models revealed a significant fixed effect of *time* from baseline to post-intervention, F (1, 50.76) = 13.07, $p < .001$, $\eta^2_p = 0.20$, but no significant effects of *group*, $F(1, 59.94) = 0.23$, $p = .631$, or the *group x time interaction*, $F(1, 50.76) = 2.55$, $p = .117$, $\eta^2_p = .05$. The observed between-group effect at post-assessment was small ($d = -0.12$). The estimated mean difference between measurement points was significant in the human CA condition (0.46, $SE = 0.13$, $p < .001$), corresponding to a large (observed) within-group effect ($d = -0.92$), but not in the genAI condition (0.18, $SE = 0.12$, $p = .156$, $d = -0.19$).

Next, we included the follow-up data in the models to investigate long-term intervention effects. Results showed a significant fixed effect of *time* from baseline to follow-up, $F(2, 85.42) = 7.0$ $p = .001$, $\eta^2_p = 0.14$, but no significant effects of *group* $F(1, 59.56) = 0.006$, $p = .937$ or (marginally non-significant) for the *group x time interaction*, $F(2, 85.42) = 2.92$, $p = .059$. The estimated mean difference between baseline and follow-up was significant in the human CA condition (0.44, $SE = 0.14$, $p = .006$), but not in the genAI condition (-0.005, $SE = 0.13$, $p = 1.00$). Figure 2 displays the change in model-estimated procrastination scores across all measurement points for the two conditions.

Table 2. Demographic characteristics.

Variables	Total $N = 62$	ChatGPT-4 ($n = 32$)	Human CA ($n = 30$)
Gender, *n* (%)[a]			
Female	58 (93.5%)	29 (90.6%)	29 (96.7%)
Male	4 (6.5%)	3 (9.4%)	1 (3.3%)
Age, *Mean* (*SD*, range)	23.60 (3.84; 18 – 36)	24.88 (4.08; 18 – 36)	22.23 (3.08; 19 – 31)
Highest educational qualification, *n* (%)[b]			
High	62 (100%)	32 (100%)	30 (100%)
Moderate	-	-	-
Low	-	-	-
Relationship status, *n* (%)			
In Partnership	36 (58.1%)	20 (62.5%)	16 (53.3%)
Not in partnership	26 (41.9%)	12 (37.5%)	14 (46.7%)
Living situation, *n* (%)			
Living alone	12 (19.4%)	9 (28.1%)	3 (10%)
Living with others	50 (80.6%)	23 (71.9%)	27 (90%)
Mother tongue			
German	59 (95.2 %)	30 (93.8 %)	29 (96.7 %)
Other	3 (4.8 %)	2 (6.3 %)	1 (3.3 %)
Intake of psychotropic medication (yes), *n* (%)	5 (8.1 %)	3 (9.4 %)	2 (6.7 %)
Expected change (range 0-5)[c], *Mean* (*SD*)	5.21 (1.95)	5.22 (2.07)	5.20 (1.85)

[a]*Note.* No non-binary gender indication.
[b]High = general or subject-restricted university entrance qualification, university degree; Moderate = secondary school certificate, graduation from a polytechnic secondary school, entrance qualification for universities of applied sciences; Low = no school-leaving certificate, lower secondary/elementary school-leaving certificate.
[c]0 = no improvement to 10 = greatest conceivable improvement.

Secondary outcomes

When analyzing change from baseline to post-assessment, LMMs indicated significant main effects of *time* for rumination ($p = .023$), but not for depression and anxiety, while the effect was marginally non-significant for worrying ($p = .053$). No main effects for *group* were observed for any of the secondary outcomes. For depression scores, a significant *group* × *time* interaction, $F\ (1, 51.52) = 4.23$, $p = .045$, $\eta^2_p = .08$ in favor of the human CA condition was observed (see also Table 3), whereas no interaction effects were found for other secondary outcomes. Observed between-group effects at post-assessment were small (*d*s = -0.14 to -0.33). Significant change in estimated means was only observed in the human CA condition for worrying (2.70, $SE = 1.30$, $p = .044$) and rumination (0.18, $SE = 0.09$, $p = .039$), corresponding to a small within-effects (*ds* = -0.31 and -0.31). LMMs including pre-, post, and follow-up scores revealed a significant effect of *time* only for worrying, $F\ (2, 84.16) = 3.13$, $p = .049$, $\eta^2_p = .07$, but no effects for *group* or the *time x group* interaction for all secondary outcomes.

Mediation through common factors

Observed values for the total and subscales of the MSB are provided in Supplementary Table D. LMMs including T1 (session one), T2 (session two), and post-assessment (session three) showed significant main effects of *time* for the subscales "problem activation" ($F\ (2, 102.92) = 5.98$, $p = .003$) and "problem mastery" ($F\ (2, 103.86) = 65.08$, $p < .001$) but not for "therapy alliance" ($F\ (2, 103.82) = 0.11$, $p = .898$). The effects for *group* and the *group x time* interactions were non-significant for all sub-scales, indicating no differences in the change of common factors between conditions. Mediation analysis showed autoregressive effects of the total score of the MSB and procrastination scores in the APROF. However, the total score did not predict procrastination scores at post-measurement either after the first ($\beta = 0.01$, $SE = 0.01$, $p = .167$) or after the second coaching session ($\beta = -0.01$, $SE = 0.01$, $p = .446$), see Figure 3.

Intervention usage

Fifty-nine (95.2 %) participants received at least one chat-coaching session and around nine in ten (87.1%, $n = 54$) completed all three sessions. The average duration of chat sessions was 29.38 minutes ($SD = 7.42$), which did not differ significantly between groups, $\chi^2\ (3) = 0.15$, $p = .985$. The average intervention duration was 17.21 days ($SD = 4.59$).

Table 3. Observed values per time point, model-based interaction effects (pre-post), and effect sizes for primary and secondary outcome measures.

Measure (total score range)	Baseline (T0)[a]	Post (T3)[b]	Follow-up (T4)[c]	Effect size[d] [95% CI]	
	M (SD)	*M (SD)*		Within (Pre/Post)[d]	Between (post)
APROF (1-7)	*Interaction effect: F* (1, 50.76) = 2.55, p = .117, η^2_p = 0.05				
ChatGPT-4	4.36 (1.12)	4.21 (0.96)	4.53 (0.98)	-0.19 [-0.84, 0.25]	-0.12 [-0.67, 0.43]
Human CA	4.61 (0.80)	4.10 (0.80)	4.22 (0.75)	-0.92 [-1.91, -0.75]	
PHQ-9 (0-27)	*Interaction effect: F* (1, 51.52) = 4.23, p = .045*, η^2_p = 0.08				
ChatGPT-4	6.41 (5.59)	7.46 (4.72)	7.53 (3.98)	0.22 [-0.28, 0.83]	-0.14 [-0.69, 0.41]
Human CA	8.03 (4.89)	6.8 (4.65)	7.19 (4.59)	-0.30 [-0.92, 0.2]	
GAD-7 (0-21)	*Interaction effect: F* (1, 50.44) = 2.16, p = .148, η^2_p = 0.04				
ChatGPT-4	7.72 (4.7)	7.73 (4.4)	7.63 (2.97)	0.003 [-0.540, 0.55]	-0.21 [-0.76, 0.34]
Human CA	7.53 (4.12)	6.8 (4.44)	6.12 (4.13)	-0.26 [-0.93, 0.19]	
RRQ, subscale rumination (1-5)	*Interaction effect: F* (1, 50.69) = 0.47, p = .495, η^2_p = 0.001				
ChatGPT-4	3.72 (0.71)	3.64 (0.74)	3.62 (.71)	-0.19 [-0.85, 0.24]	
Human CA	3.58 (0.76)	3.41 (0.64)	3.63 (.68)	-0.37 [-1.21, -0.09]	-0.33 [-0.89, 0.22]
PSWQ (16-80)	*Interaction effect: F* (1, 49.88) = 0.95, p = .335, η^2_p = 0.02				
ChatGPT-4	54.03 (11.04)	54.15 (11.3)	53.26 (9.82)	0.02 [-0.51, 0.58]	-0.31 [-0.87, 0.24]
Human CA	52.77 (12.75)	50.64 (11.15)	53.44 (9.77)	-0.30 [-1.10, 0.01]	

[a]*Note.* Sample size ChatGPT-4: n = 32, human CA: n = 30.
[b]ChatGPT-4: n = 26, human CA: n = 25.
[c]APROF: ChatGPT-4: n = 20, human CA: n = 17; other questionnaires: ChatGPT-4: n = 19, human CA: n = 16. Interaction term: group (ChatGPT-4 vs. human CA) x time (baseline, post).
[d]Correlation between time points: APROF = 0.76, PHQ-9 = 0.64, GAD-7 = 0.77, RRQ = 0.82, PSWQ = 0.84. APROF, General Procrastination Questionnaire; PHQ-9, Patient Health Questionnaire, 9-item version; GAD-7, Generalized Anxiety Disorder Questionnaire, RRQ, Rumination-Reflection-Questionnaire; PSWQ, Penn State Worry Questionnaires.

Out of n = 51 participants with valid answers, around two-thirds (68.6 %, n = 31) correctly guessed their assigned condition at post-assessment, this rate did not differ significantly between groups, χ^2 (1) = 0.157, p = 0.692 (ChatGPT-4: 19, 73.1%, human CA: 16, 64%). Participants' comments (see Supplementary Table E) suggest that condition attribution was primarily based on process-related features of the interaction rather than on the content itself. In particular, limited emotional responsiveness and highly polished or generic language were frequently cited as indicators of the (expected) ChatGPT-4

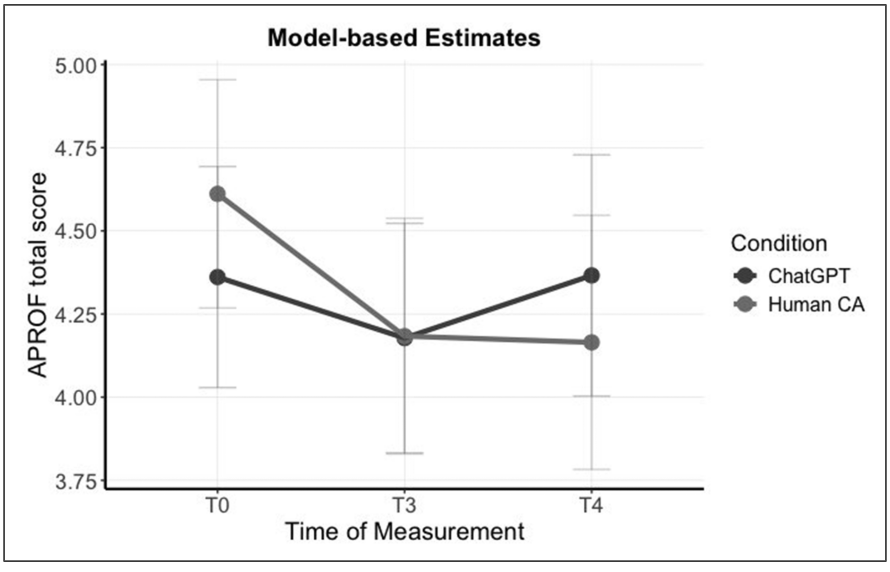

Figure 2. Change in procrastination scores as a function of intervention type.
Note. T3: Post-intervention. T4: 6 months-follow-up. Error bars represent 95% confidence intervals.

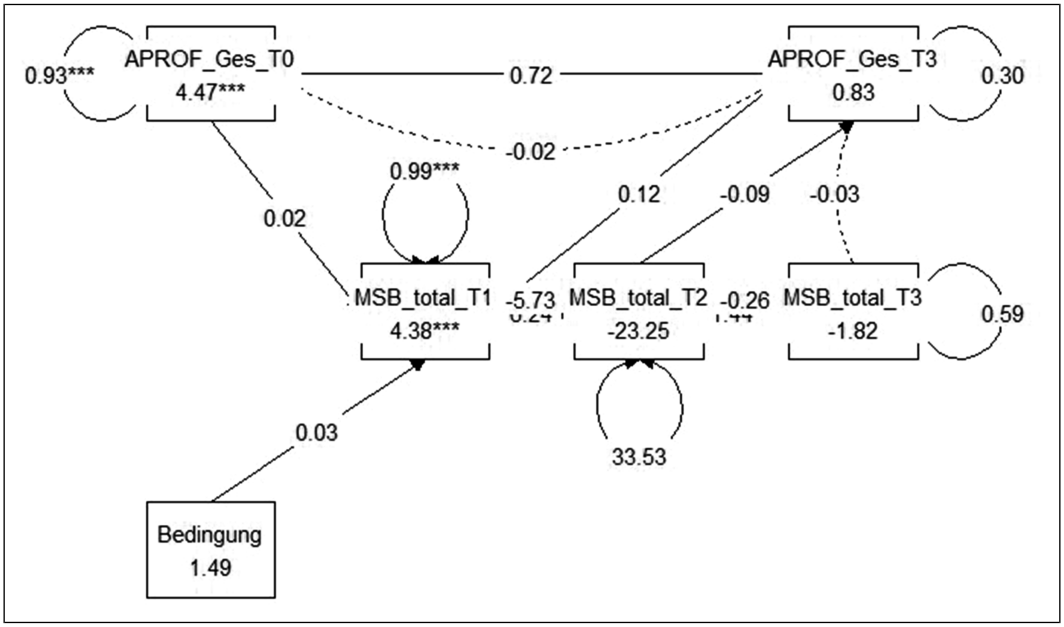

Figure 3. Cross-lagged-panel model illustrating relationship between condition, MSB total score and APROF total score at different assessment points.
Note: Values represent standardized path coefficients (significant paths indicated by asterisks: *$p < .05$, **$p < .01$, ***$p < .001$). Numbers within ellipses represent variance accounted for. Bedingung = condition, APROF_ges = APROF total score, T0 = Baseline, T1 = Post, T3 = Follow-Up.

based CA condition, whereas empathic responding and dialogical coherence were more often associated with (expected) human CA.

Negative effects

In the sub-group of participants who completed at least the first coaching session and filled out the NEQ at post-measurement ($n = 51$), around one-third (43.1%, $n = 22$) reported at least one unwanted negative effect (of any intensity) due to the intervention, which did not differ statistically between groups (genAI: 12/26, 46.2%, human CA: 10/25, 40.0%, $\chi^2(1) = 0.03$, $p = .872$). The most frequent negative effects in the genAI condition concerned perceived lack of benefit, difficulties with comprehension or engagement, and emotional distress. In the human CA condition, most frequent negative effects similarly clustered around perceived lack of benefit, emotional distress, and process-related difficulties (e.g., problems understanding the coach or the coaching process). Table 4 contains the list of negative effects reported by condition. Of all negative effects reported ($n = 63$), four (6.3%), were rated as "extremely" or "very". In both conditions, no serious adverse events (SAEs) such as hospitalization, suicidality, or mental breakdown were recorded.

Discussion

Principal findings

This exploratory, hypothesis-generating pilot trial investigated the efficacy, safety, and mechanisms of genAI based CA using ChatGPT-4 compared to human CA as a text-based micro-intervention for adults with procrastination. Our findings regarding insignificant group by time interactions did not indicate a clear superiority of either CA for procrastination or secondary outcomes. This finding aligns with a previous study by Mutter et al.,[24] which found no significant differences in procrastination reduction between an Internet-based intervention accompanied by an AI-based guidance (i.e., rule-based feedback) and one with a human guidance. One interpretation could be that within relatively structured, text-based intervention contexts, chatbots may effectively leverage their strengths up to those of written human guidance. In the context of guided internet-based interventions, these findings indicate that genAI-based chatbots may serve as a viable and cost-effective alternative to human guidance.[37] At the same time, it seems premature to equate the effectiveness of genAI-based CAs with that of human coaching by the mere absence of a significant difference between groups, particularly since evidence for non-inferiority requires larger sample sizes.[38] Moreover, the pattern of slightly stronger within-group improvements in

Table 4. Frequencies (*n, %*) of self-reported adverse effects attributed to coaching by participants ($n = 51$) at the primary endpoint (T3) in the Negative Effects Questionnaire (NEQ).

Items		ChatGPT-4 ($n = 26$)	Human CA ($n = 25$)
1	Increased sleep problems	0	0
2	Increased stress	2 (7.7%)	4 (16%)
3	Increased anxiety	0	0
4	Increased worry	0	1 (4%)
5	Increased hopelessness	1 (3.8%)	1 (4%)
6	Increased unpleasant feelings	0	1 (4%)
7	Symptoms got worse	1 (3.8%)	1 (4%)
8	Unpleasant memories resurfaced	1 (3.8%)	1 (4%)
9	Fear that others would find out about coaching	0	0
10	Suicidal thoughts	0	0
11	Feelings of shame because of coaching	0	0
12	Lost belief that things can get better	3 (11.5%)	3 (12%)
13	Thinking that symptoms could not get better	4 (15.4%)	2 (8%)
14	Feeling of dependency on coaching	1 (3.8%)	0
15	Did not always understand coaching	4 (15.4%)	4 (16%)
16	Did not always understand coach (therapist)	4 (15.4%)	3 (12%)
17	Did not have confidence in the coaching	4 (15.4%)	1 (4%)
18	Coaching did not produce any results	5 (19.2%)	2 (8%)
19	Expectations for the coach were not fulfilled	3 (11.5%)	3 (12%)
20	Feeling that coaching was not motivating	2 (7.7%)	1 (4%)

the human CA condition and a small to moderate estimated effect for the group by time interaction on the primary outcome could point to advantages of the human-delivered CA in reducing procrastination, although these differences remain tentative.

The brief and low-intensity nature of the intervention may have been sufficient to produce significant improvements over time, while limiting the ability to detect differential effects between conditions. Prior studies on digital interventions for procrastination vary widely in duration, from single-session[21] to multi-week interventions. For example, Rozental et al.[23] investigated a ten-week CBT-based Internet intervention (with and without human guidance) for procrastination and found moderate effects against a waitlist control but not differences between active interventions. Mutter et al.[24] studied an approximately five-week CBT-based Internet intervention which included a broader range of therapeutic components (i.e., psychoeducation, time management, motivation, self-regulation, and relapse prevention).

Furthermore, because the specific combination of treatment elements in our intervention was not evaluated a priori, our findings therefore cannot fully rule out the effectiveness of the micro-intervention approach for procrastination per se. Additionally, our sample exhibited only mild to moderate baseline procrastination levels, limiting sensitivity to change compared to clinical presentations.[39] This aligns with meta-analytic findings by Li et al.[7] which indicate stronger aggregated effects of digital interventions in clinical populations. At the same time, prior research suggests that individuals with more severe symptoms tend to prefer human over digital forms of guidance or counseling.[40,41] However, preferences and attitudes toward AI-based interventions are complex and evolving. Some studies suggest substantial preferences for AI-driven therapy, particularly due to advantages such as the ability to comfortably discuss embarrassing experiences, accessibility, and remote communication.[42] The current global fascination with the eloquence and seemingly convincing nature of genAI-based dialogues may contribute to the acceptance of AI-driven interventions. In contrast, other studies have indicated reservations and critique about using AI tools in psychotherapy,[43] including therapeutical, ethical, social or aspects of data security.

In terms of therapeutic mechanisms, this study also examined common factors across conditions, including problem activation, therapeutic alliance, and problem mastery. No significant group differences emerged, and these factors did not act as mediators of treatment outcomes. This finding is consistent with prior research suggesting that positive therapeutic relationships can emerge in Internet- or mobile-based interventions (IMIs), even in mostly self-guided (or even fully automated) formats.[44,45] This appears counter-intuitive to many practitioners, as IMIs are often perceived as lacking the critical therapeutic relationship factor, which is a common reason for skepticism toward both IMIs and AI-driven interventions.[43] In contrast to face-to-face treatment, therapeutic alliance in remote interventions seems to be less associated with treatment outcomes than with adherence.[45] The common factors in this study were relatively stable over time, which

could be explained by the low dose of the intervention. While for example, therapeutic alliance tends to stabilize early in treatment, our text-based coaching may have had too little variability and time to induce this process.[46]

This study is among the first to systematically assess the adverse effects of generative CAs. While no significant differences in the frequency of self-reported negative effects emerged between conditions, the overall rate of reported side effects (43.1% with at least one unwanted negative effect) was noteworthy, and previous studies on guided, CBT-based Internet interventions have reported much lower frequencies of user-rated side effects between 11 – 19%.[47–50] Unwanted effects were predominantly related to perceived lack of benefit, emotional strain, or process-related difficulties rather than harmful or deteriorating effects. Only a small portion of unwanted effects were rated as severe, with no reported SAEs. Our findings suggest that adverse experiences may be less attributable to the modality of delivery (genAI vs. human CA) and more related to unmet expectations and the limited impact of a micro-intervention. Similar patterns have been reported in user feedback from university students participating in a single-session digital intervention targeting procrastination.[21] At a more descriptive level, slightly higher rates of perceived lack of improvement and reduced confidence in the intervention were observed in the genAI condition, which may reflect differences in perceived credibility or engagement, although this remains speculative. With regard to change in self-reports, there was even a (descriptive) increase in depressive symptoms in the genAI condition. From an ethical perspective, such brief (AI-based) micro-interventions may create an expectation–outcome mismatch, as users may attribute therapeutic depth to a conversational agent that cannot be met within a short interaction, warranting not only dose-response-research but also transparent expectation management. Moreover, the high degree of standardization of most rule-based or even non-AI-based digital self-help interventions may limit responsiveness to individual needs and can be associated with higher attrition rates.[51] Furthermore, limited opportunities for adaptive monitoring or escalation in case of distress may leave negative experiences insufficiently addressed. In line with emerging frameworks for responsible AI in mental health (e.g., Readiness Evaluation for AI-Mental Health Deployment and Implementation, READI),[52] future work should therefore emphasize structured assessment of adverse effects, explicit expectation management, and appropriate safeguards.

A key strength of this study is the participant-blinded design. However, post-hoc participant feedback revealed that many were able to correctly identify their assigned condition. This could be explained by a subjective response latency (although session durations were comparable) and a distinct linguistic "AI" style of conversation by ChatGPT-4 (possibly even more pronounced in languages other than English) in as compared to the human CA. Such partial unblinding may have introduced expectancy effects and thereby influenced participants' perceptions and evaluations of the intervention, potentially limiting the internal validity of findings. Prior research suggests that beliefs about the nature of agency (i.e., human vs. AI) behind computerized interventions can influence user behavior and devaluing alliance or feeling heard.[53] For example, Gratch et al.[54] found reduced impression management when users were made to believe they were interacting with a fully automated avatar rather than when told they interacted with a human-operated avatar, even when this was the same for both conditions.

Limitations

Despite its strengths, this study has some limitations. First, given the pilot nature and small sample size of this study, the study was not sufficiently powered to detect smaller effects nor to formally establish non-inferiority and conclusions should remain tentative until further large-scale trials confirm these findings. Second, reliance on self-reported outcomes without clinical diagnostic confirmation may limit the clinical validity of findings. Third, the sample was restricted to (psychology) university students and was largely female, representing a typical "WEIRD" (Western, Educated, Industrialized, Rich, Democratic)[55] and highly homogeneous population, which may limit the generalizability of our findings. In addition, participants exhibited relatively high openness to AI-based systems, raising concerns about selection bias and limiting generalization to more diverse or less technology-affine populations as well as other forms of (non-academic) procrastination (e.g., workplace-related). At the same time, procrastination is particularly prevalent among the student population and is associated with significant psychological and physical impairments.[9] In addition, the human coaches received relatively brief training, which may have limited intervention fidelity. However, the intervention was guided by a detailed and structured manual and supported by supervision, which likely contributed to a certain degree of standardization. Although allocation concealment was limited, assignments followed the fixed pre-generated order, which minimizes potential selection bias. We acknowledge that the rather simple prompts used in our study do not fully exploit the potential of LLMs. Text-based intervention formats may inherently limit effectiveness, as prior research suggests that multi-modal or voice-based CAs could enhance engagement and therapeutic outcome.[7] On the other hand, our interest was particularly drawn to the near 'out-of-the-box' potential of publicly accessible chatbots such as ChatGPT, which are increasingly used for mental health issues,[56] even though not designed for this purpose. Here, our findings not only highlight its potential for low-threshold applications but also the caveat for misuse, as it now becomes feasible for virtually anyone to offer mental health interventions using genAI-based chatbots without sufficient safeguards.

Conclusion

This pilot randomized trial did not provide evidence for the superiority of either a ChatGPT-4-based or a human conversational agent in reducing procrastination within the scope of a three-session micro-intervention, neither in the short nor in the longer term. Notably, the ChatGPT-4-based intervention failed to produce significant change over time, raising important questions about its current effectiveness in this context. Our findings also underscore the need for research on the potential risks and limitations of AI-driven interventions. Future research should focus on developing structured dialogue frameworks and context-sensitive design principles to enhance the effectiveness, transparency, and clinical utility of AI-based mental health tools.

Acknowledgments

We like to thank Nina Lück for her supervision in developing the micro-intervention manual. This research received no external funding.

ORCID iD

Severin Hennemann https://orcid.org/0000-0001-8780-9985

Ethical considerations

The study protocol was approved by the local ethics committee of the Department of Psychology, University of Mainz, Germany (reference number 2023-JGU-psychEK-026, July 2023). All procedures were conducted in accordance with the ethical standards of the Declaration of Helsinki.

Consent to participate

Written informed consent was obtained from all participants prior to participation.

Author contributions

SH: Conceptualization, Project administration, Methodology, Formal analysis, Visualization, Writing - Original Draft, Writing - Review & Editing. JMF: Investigation, Data Curation, Formal analysis, Writing - Review & Editing. CT: Investigation, Data Curation, Formal analysis, Writing - Review & Editing. MW: Writing - Review & Editing. SMJ: Supervision, Conceptualization, Writing - Review & Editing.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

The datasets analyzed, analysis code, and additional materials supporting the findings of this study are publicly available on the Open Science Framework: https://osf.io/2prnb/overview.

Trial Registration

https://aspredicted.org/w5qc4.pdf (number: #150,559; date: 2023/11/10).

Supplemental material

Supplemental material for this article is available online.

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Psychiatry* 2022; 9: 137–150. https://doi.org/10.1016/S2215-0366(21)00395-3
2. McGrath JJ, Al-Hamzawi A, Alonso J, et al. Age of onset and cumulative risk of mental disorders: a cross-national analysis of population surveys from 29 countries. *Lancet Psychiatry* 2023; 10: 668–681. https://doi.org/10.1016/S2215-0366(23)00193-1

3. Ghafari M, Nadi T, Bahadivand-Chegini S, et al. Global prevalence of unmet need for mental health care among adolescents: A systematic review and meta-analysis. *Arch Psychiatr Nurs* 2022; 36: 1–6. https://doi.org/10.1016/j.apnu.2021.10.008
4. Han B, Compton WM, Blanco C, et al. Prevalence, Treatment, And Unmet Treatment Needs Of US Adults With Mental Health And Substance Use Disorders. *Health Aff (Millwood)* 2017; 36: 1739–1747. https://doi.org/10.1377/hlthaff.2017.0584
5. Bendig E, Erb B, Schulze-Thuesing L, et al. The next generation: chatbots in clinical psychology and psychotherapy to foster mental health–a scoping review. *Verhaltenstherapie* 2019; 32: 1–13. https://doi.org/10.1159/000501812
6. Abd-Alrazaq AA, Rababeh A, Alajlani M, et al. Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. *J Med Internet Res* 2020; 22: 16021. https://doi.org/10.2196/16021
7. Li H, Zhang R, Lee Y-C, et al. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *NPJ Digit Med* 2023; 6: 236. https://doi.org/10.1038/s41746-023-00979-5.
8. Heinz MV, Mackin DM, Trudeau BM, et al. Randomized Trial of a Generative AI Chatbot for Mental Health Treatment. *NEJM AI* 2025; 2. https://doi.org/10.1056/AIoa2400802.
9. Johansson F, Rozental A, Edlund K, et al. Associations Between Procrastination and Subsequent Health Outcomes Among University Students in Sweden. *JAMA Netw Open* 2023; 6: 2249346. https://doi.org/10.1001/jamanetworkopen.2022.49346
10. Nguyen B, Steel P and Ferrari JR. Procrastination's Impact in the Workplace and the Workplace's Impact on Procrastination. *Int J Selection Assessment* 2013; 21: 388–399. https://doi.org/10.1111/ijsa.12048
11. Sirois FM, Melia-Gordon ML and Pychyl TA. I'll look after my health, later: an investigation of procrastination and health. *Pers Individ Dif* 2003; 35: 1167–1184. https://doi.org/10.1016/s0191-8869(02)00326-4
12. Ferrari JR, O'Callaghan J and Newbegin I. Prevalence of Procrastination in the United States, United Kingdom, and Australia: Arousal and Avoidance Delays among Adults. *N Am J Psychol* 2005; 7: 1–6.
13. Beutel ME, Klein EM, Aufenanger S, et al. Procrastination, Distress and Life Satisfaction across the Age Range - A German Representative Community Study. *PLoS One* 2016; 11: 0148054. https://doi.org/10.1371/journal.pone.0148054
14. Day V, Mensink D and O'Sullivan M. Patterns of Academic Procrastination. *Journal of College Reading and Learning* 2000; 30: 120–134. https://doi.org/10.1080/10790195.2000.10850090
15. Onwuegbuzie AJ. Academic procrastination and statistics anxiety. *Assess Eval High Educ* 2004; 29: 3–19. https://doi.org/10.1080/0260293042000160384
16. Höcker A, Engberding M and Rist F. *Prokrastination: ein manual zur behandlung des pathologischen aufschiebens*. Hogrefe Verlag GmbH & Company KG, 2017.
17. Steel P. The nature of procrastination: a meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychol Bull* 2007; 133: 65–94. https://doi.org/10.1037/0033-2909.133.1.65
18. Svartdal F, Dahl TI, Gamst-Klaussen T, et al. How Study Environments Foster Academic Procrastination: Overview and Recommendations. *Front Psychol* 2020; 11: 540910. https://doi.org/10.3389/fpsyg.2020.540910
19. van Eerde W and Klingsieck KB. Overcoming procrastination? A meta-analysis of intervention studies. *Educ Res Rev* 2018; 25: 73–85. https://doi.org/10.1016/j.edurev.2018.09.002
20. Rozental A and Carlbring P. Understanding and Treating Procrastination: A Review of a Common Self-Regulatory Failure. *Psychology (Irvine)* 2014; 05: 1488–1502. https://doi.org/10.4236/psych.2014.513160
21. Åsberg K, Löf M and Bendtsen M. Effects of a single session low-threshold digital intervention for procrastination behaviors among university students (Focus): Findings from a randomized controlled trial. *Internet Interventions* 2024; 36: 100741. https://doi.org/10.1016/j.invent.2024.100741
22. Amarnath A, Ozmen S, Van Klaveren C, et al. Effectiveness of a guided internet-based intervention in reducing procrastination among university students – a randomized controlled trial. *Internet Interventions* 2025; 42: 100878. https://doi.org/10.1016/j.invent.2025.100878
23. Rozental A, Forsell E, Svensson A, et al. Internet-based cognitive-behavior therapy for procrastination: A randomized controlled trial. *J Consult Clin Psychol* 2015; 83: 808–824. https://doi.org/10.1037/ccp0000023
24. Mutter A, Küchler A-M, Idrees AR, et al. StudiCare procrastination - Randomized controlled non-inferiority trial of a persuasive design-optimized internet- and mobile-based intervention with digital coach targeting procrastination in college students. *BMC Psychol* 2023; 11: 273. https://doi.org/10.1186/s40359-023-01312-1
25. Steel P. Arousal, avoidant and decisional procrastinators: Do they exist?. *Pers Individ Dif* 2010; 48: 926–934. https://doi.org/10.1016/j.paid.2010.02.025
26. Young HP, Learning by trial and error. *Games and Economic Behavior* 2009; 65(2): 626–643. https://doi.org/10.1016/j.geb.2008.02.011
27. Bräscher A-K and Witthöft M. Die Erfassung allgemeiner Wirkfaktoren in der Psychotherapie. *Z Klin Psychol Psychother (Gott)* 2020; 49: 128–135. https://doi.org/10.1026/1616-3443/a000583

28. Tillmann A, Niemeyer J and Krömker D. Das schaue ich mir morgen an“ - Aufschiebeverhalten bei der Nutzung von eLectures: Eine Analyse. DeLFI 2016–Die 14. E-Learning Fachtagung Informatik; Potsdam: 2016; 47–57. https://www.uni-potsdam.de/fileadmin/projects/multimedia/delfi2016/DeLFI2016_Proceedings.pdf
29. Kroenke K and Spitzer RL. The PHQ-9: A new depression diagnostic and severity measure. *Psychiatr Ann* 2002; 32: 509–515. https://doi.org/10.3928/0048-5713-20020901-06
30. Spitzer RL, Kroenke K, Williams JBW, et al. A brief measure for assessing generalized anxiety disorder: The GAD-7. *Arch Intern Med* 2006; 166: 1092–1097. https://doi.org/10.1001/archinte.166.10.1092
31. Löwe B, Decker O, Müller S, et al. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. *Med Care* 2008; 46: 266–274. https://doi.org/10.1097/MLR.0b013e318160d093
32. Schütz A. A German Version of the Rumination-Reflection Questionnaire. *Chemnitz University of Technology* 2000; https://www.uni-bamberg.de/fileadmin/perspsych/Tests_und_Frageboegen/RQQ.pdf.
33. Trapnell PD and Campbell JD. Private self-consciousness and the five-factor model of personality: distinguishing rumination from reflection. *J Pers Soc Psychol* 1999; 76: 284–304. https://doi.org/10.1037//0022-3514.76.2.284
34. Glöckner-Rist A and Rist F. *Deutsche Version des Penn State Worry Questionnaire (PSWQ-d)*. ZIS - GESIS Leibniz Institute for the Social Sciences, 2006. https://doi.org/10.6102/zis219. Epub ahead of print 1 January 2006.
35. Rozental A, Kottorp A, Boettcher J, et al. Negative Effects of Psychological Treatments: An Exploratory Factor Analysis of the Negative Effects Questionnaire for Monitoring and Reporting Adverse and Unwanted Events. *PLoS One* 2016; 11: 0157503. https://doi.org/10.1371/journal.pone.0157503
36. Bell ML and Rabe BA. The mixed model for repeated measures for cluster randomized trials: a simulation study investigating bias and type I error with missing continuous data. *Trials* 2020; 21: 148. https://doi.org/10.1186/s13063-020-4114-9
37. Carlbring P, Hadjistavropoulos H, Kleiboer A, et al. A new era in Internet interventions: The advent of Chat-GPT and AI-assisted therapist guidance. *Internet Interv* 2023; 32: 100621. https://doi.org/10.1016/j.invent.2023.100621
38. Rief W and Hofmann SG. Some problems with non-inferiority tests in psychotherapy research: psychodynamic therapies as an example. *Psychol Med* 2018; 48: 1392–1394. https://doi.org/10.1017/S0033291718000247
39. Scholten W, Seldenrijk A, Hoogendoorn A, et al. Baseline Severity as a Moderator of the Waiting List-Controlled Association of Cognitive Behavioral Therapy With Symptom Change in Social Anxiety Disorder: A Systematic Review and Individual Patient Data Meta-analysis. *JAMA Psychiatry* 2023; 80: 822–831. https://doi.org/10.1001/jamapsychiatry.2023.1291
40. Berle D, Starcevic V, Milicevic D, et al. Do patients prefer face-to-face or internet-based therapy? *Psychother Psychosom* 2015; 84: 61–62. https://doi.org/10.1159/000367944
41. Renn BN, Hoeft TJ, Lee HS, et al. Preference for in-person psychotherapy versus digital psychotherapy options for depression: survey of adults in the U.S. *NPJ Digit Med* 2019; 2: 6. https://doi.org/10.1038/s41746-019-0077-1
42. Aktan ME, Turhan Z and Dolu İ. Attitudes and perspectives towards the preferences for artificial intelligence in psychotherapy. *Comput Human Behav* 2022; 133: 107273. https://doi.org/10.1016/j.chb.2022.107273
43. Prescott J and Hanley T. Therapists' attitudes towards the use of AI in therapeutic practice: considering the therapeutic alliance. *MHSI* 2023; 27: 177–185. https://doi.org/10.1108/mhsi-02-2023-0020
44. Hennemann S, Böhme K, Kleinstäuber M, et al. Is Therapist Support Needed? Comparing Therapist- and Self-Guided Internet-Based CBT for Somatic Symptom Distress (iSOMA) in Emerging Adults. *Behav Ther* 2022; 53: 1205–1218. https://doi.org/10.1016/j.beth.2022.06.006
45. Tremain H, McEnery C, Fletcher K, et al. The Therapeutic Alliance in Digital Mental Health Interventions for Serious Mental Illnesses: Narrative Review. *JMIR Ment Health* 2020; 7: 17204. https://doi.org/10.2196/17204
46. Wampold BE. *The Great Psychotherapy Debate*. Routledge, 2015. https://doi.org/10.4324/9780203582015. Epub ahead of print 1 January 2015).
47. Boettcher J, Rozental A, Andersson G, et al. Side effects in Internet-based interventions for Social Anxiety Disorder. *Internet Interv* 2014; 1: 3–11. https://doi.org/10.1016/j.invent.2014.02.002
48. Hedman E, Axelsson E, Andersson E, et al. Exposure-based cognitive–behavioural therapy via the internet and as bibliotherapy for somatic symptom disorder and illness anxiety disorder: Randomised controlled trial. *Br J Psychiatry* 2016; 209: 407–413. https://doi.org/10.1192/bjp.bp.116.181396
49. Hennemann S, Böhme K, Kleinstäuber M, et al. Internet-based CBT for somatic symptom distress (iSOMA) in emerging adults: A randomized controlled trial. *J Consult Clin Psychol* 2022; 90: 353–365. https://doi.org/10.1037/ccp0000707
50. Newby JM, Smith J, Uppal S, et al. Internet-based cognitive behavioral therapy versus psychoeducation control for illness anxiety disorder and somatic symptom disorder: A randomized controlled trial. *J Consult Clin Psychol* 2018; 86: 89–98. https://doi.org/10.1037/ccp0000248
51. Baumel A, Muench F, Edan S, et al. Objective User Engagement With Mental Health Apps: Systematic Search and Panel-Based Usage Analysis. *J Med Internet Res* 2019; 21: 14567. https://doi.org/10.2196/14567

52. Stade EC, Eichstaedt JC, Kim JP, et al. Readiness evaluation for artificial intelligence-mental health deployment and implementation (READI): A review and proposed framework. *Technology, Mind, and Behavior* 2025; 6: 111–122. https://doi.org/10.1037/tmb0000163
53. Yin Y, Jia N and Wakslak CJ. AI can help people feel heard, but an AI label diminishes this impact. *Proc Natl Acad Sci USA* 2024; 121: e2319112121. https://doi.org/10.1073/pnas.2319112121
54. Gratch J, Lucas GM, King AA, et al. It's only a computer: the impact of human-agent interaction in clinical interviews. In: Lomuscio A, Scerri P, Bazzan A, et al. (eds). *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems*, 2014, pp. 85–92.
55. Henrich J, Heine SJ and Norenzayan A. The weirdest people in the world? *Behav Brain Sci* 2010; 33: 61–135. https://doi.org/10.1017/S0140525X0999152X
56. Rousmaniere T, Zhang Y, Li X, et al. Large language models as mental health resources: Patterns of use in the United States. *Practice Innovations* 2025. https://doi.org/10.1037/pri0000292, Epub ahead of print 21.