nature portfolio

Corresponding author(s): Dr Max Rollwage

Last updated by author(s): Jan 15, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Initial screening survey data and weekly clinical outcome data for the real-world investigation subset were collected using the commercial web-based platform Typeform. The experimental therapy sessions were hosted via a web application developed using the open-source library Streamlit (v1.40.2), operating within a Python (v3.10.4) environment. Custom Python code was employed within this Streamlit application to manage the deployment of either the AI agents or the human agent for conducting the therapy sessions and capturing interaction data. Real-world data collection used the Limbic Care app and clinical outcomes from the UK's National Health Service were recorded using undisclosed patient management system(s).
Data analysis	The analysis used on Python (v3.10.4), utilizing the open-source libraries 'statsmodels' (v0.14.3) and 'scipy' (v1.14.1) for statistical computations. Evaluation of clinical performance was conducted using OpenAI's commercial GPT-4o large language model (version: gpt-4o-2024-08-06) via its API. Custom Python code was used to implement the analysis pipeline. All custom code required to reproduce the statistical analyses presented in this manuscript from the provided data is written in Python (v3.10.4) and is publicly available via the GitHub repository: https://doi.org/10.5281/zenodo.17176593.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data supporting the findings of this study are publicly available on GitHub at https://doi.org/10.5281/zenodo.17176593. This repository contains the de-identified, processed dataset necessary to replicate the statistical results reported in the paper. Raw data containing sensitive therapeutic transcripts are not publicly available to protect participant privacy.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Biological sex information was obtained based on participant self-report provided upon registration with the Prolific platform. Gender identity information was not collected for this study component.Biological sex was considered during the design of the experimental study. Recruitment via the Prolific platform aimed for approximate balance between male and female participants. Randomization to experimental conditions was stratified by biological sex (along with age group) to ensure comparable distributions across conditions. The experimental sample (N=227) consisted of 49.1% female participants. The de-identified dataset (see Data Availability statement) contains data on biological sex for reproducibility. No specific statistical analyses comparing outcomes between male and female participants were pre-planned or conducted as part of the primary analysis plan. The study was primarily powered to detect the main effects of the AI architecture across LLMs rather than sex-based subgroup differences. Stratification by sex was employed primarily as a design control measure. Participant sex and gender were not considered in the design or data collection for the real-world investigation component.This information was not routinely collected by the smartphone application used for the real-world deployment of the cognitive layer architecture. Consequently, data from the real-world investigation cannot be disaggregated or analyzed based on sex or gender. No sex- or gender-based analyses were possible for this study component due to the lack of collected data.
Reporting on race, ethnicity, or other socially relevant groupings	Information regarding participant race and ethnicity was collected for the experimental study component via the Prolific recruitment platform. Participants self-reported their race and/or ethnicity upon registration with Prolific, using the classification terms provided by the platform at that time. This demographic information was collected to characterize the sample but was not used to set recruitment criteria, perform stratified randomization, or conduct any statistical analyses presented in this manuscript. Race and ethnicity were not used as proxies for other variables in our analyses. For the real-world investigation component, race and ethnicity data were not collected through the smartphone application used for the deployment of the cognitive layer architecture. Consequently, no analyses involving race or ethnicity were performed for either part of the study. The race/ethnicity variable collected via Prolific for the experimental study is included in the anonymized dataset available on the Open Science Framework.
Population characteristics	The experimental sample had a mean age of 39 years (SD = 13.3, range = 18 to 69), with 49.1% female participants. Ethnic distribution included 69.3% white participants, 18.8% black participants, 6.9% asian participants, 3.7% mixed or other ethnic backgrounds, and 1.4% not reported. Most participants were in full-time employment (48.6%), with 16.5% employed part-time, 6.4% unemployed and job-seeking, 0.5% due to begin a new job, 8.3% unemployed and not seeking employment, 1.8% in other employment, and 17.9% without employment data. Approximately half (52.4%) of participants had previously had therapy for their mental health, with the vast majority (98.7%) reporting that they would be open to trying therapy in the future. Demographic information is unavailable for the real-world sample.
Recruitment	For the experimental study, participants were recruited through the online research platform Prolific. The study was advertised as involving a "mental wellbeing session." This recruitment method likely yields a sample comfortable with digital technology and online study participation. Furthermore, the study advertisement may have preferentially attracted individuals interested in mental health topics or potentially seeking support, leading to a self-selection bias towards this demographic. However, as the cognitive layer architecture is intended for use in mental health contexts, this potential bias may increase the relevance of the findings to the target population of individuals open to engaging with mental wellbeing tools. For the real world investigation, users came to the app for different reasons. (i) Some users discovered and voluntarily downloaded the wellbeing app (which incorporates the cognitive layer architecture) via social media advertisements. (ii) A subset was recruited via Prolific, specifically screened for above-threshold symptoms on the GAD-7 (anxiety) and PHQ-9 (depression) measures, and then instructed to download and use the app to investigate naturalistic usage. (iii) Other users were provided access to the app via their mental healthcare provider within the UK's NHS and voluntarily chose to engage with it. This mixed recruitment introduces potential biases. Groups (i) and (ii) likely represent digitally literate individuals proactively seeking or willing to engage with app-based mental health tools, with group (ii) specifically biased towards those with existing symptoms. Group (iii) represents individuals already engaged in formal mental healthcare but introduces a self-selection bias for those specifically willing to adopt a digital adjunct to their care. Overall, the real-world sample reflects users willing to engage with digital mental health solutions, potentially limiting generalizability to populations less inclined or able to use such technology, or those not actively seeking support or involved in healthcare.

Ethics oversight	The experimental studies were in accordance with an approval received from the University College London Research Ethics Committee (reference number: 6218/003). The observational study analysed retrospective, anonymised data from 8,920 users of a commercially available therapy chatbot application. All users, including those accessing the app through NHS Talking Therapies services, provided consent for their anonymised data to be used for research purposes at the point of registration. The UK Health Research Authority (HRA) was consulted by the research team and was informed in writing that this was a retrospective analysis of anonymised, non-interventional data, independent ethics approval was not required. All data were de-identified prior to analysis. All procedures complied with HRA guidance, institutional requirements, and international ethical standards, including the principles of the Declaration of Helsinki.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study utilized a quantitative design comprising two main parts. A preregistered, randomized, blinded, between-subjects experiment with quantitative outcomes, and an additional observational analysis quantitatively examined real-world app usage data and clinical outcomes.
Research sample	The research involved: 1) Experimental Users (N=227): UK adults recruited via Prolific, screened for relevant wellbeing issues and CBT familiarity (Mean age 39, 49% female; approx. 69% White, 19% Black, 7% Asian). Chosen for controlled testing with a relevant target group. 2) Experimental Clinicians (N=28): Qualified UK NHS CBT Therapists providing human therapy comparison and expert ratings; this group is representative of clinicians delivering these specific roles within the NHS Talking Therapies service. Chosen for clinical expertise and relevance. 3) Real-World Users: A diverse group including general app users (N≈8900), a US Prolific cohort screened for symptoms (N=296), and UK NHS patients (N=485). Chosen to evaluate real-world applicability and impact. Samples (1) and (3) reflect relevant user groups for digital mental health interventions, albeit with self-selection biases related to digital engagement and help-seeking behavior.
Sampling strategy	Experimental Study: The sampling strategy involved a two-stage procedure. First, a large pool of potential participants was screened via the Prolific platform based on pre-defined eligibility criteria (UK residence, age 18+, English fluency, no immediate risk, relevant wellbeing concern, CBT familiarity/belief). Second, from this eligible pool, participants were invited to the main study using stratified randomization based on age group and sex to ensure balanced demographics across the nine experimental conditions. The target sample size for the AI conditions (N=200, leading to N=225 total target including the human condition) was pre-determined using an a priori power analysis based on effect sizes from a previous pilot study ([DBPR]). The power analysis focused on detecting significant clinical performance differences between stand-alone LLMs and the cognitive layer architecture within each LLM condition, as measured by the Cognitive Therapy Rating Scale (CTRS) as the primary measure of clinical performance. Based on this power calculation it was determined that 25 participants per condition would be required to achieve >80% power to detect clinically meaningful differences between agent conditions for each underlying LLM at $\alpha = .05$. To account for potential exclusions (see Data exclusions below), a stopping criteria of N=234 participants was set on Prolific during data acquisition. Clinicians (therapists and evaluators) were recruited via professional networks based on specific NHS qualifications; sample size was determined by the need to cover the human therapist condition and adequately distribute rating tasks among qualified experts. Real-World Investigation: This involved convenience and purposive sampling. General app usage data were collected naturalistically from users who downloaded the wellbeing app (convenience sampling). For longitudinal outcome analyses, two specific cohorts were sampled: (1) US participants recruited via Prolific and screened purposively for above-threshold GAD-7/PHQ-9 scores, and (2) UK NHS patients using the app as part of routine care within participating services (clinical pathway sampling). Sample sizes for the real-world analyses were not pre-determined by power analysis but were based on the available data from these sources for observational analysis.
Data collection	Data were collected remotely without research staff physically present with participants - thus, research staff blinding is not applicable. Human therapists conducted sessions remotely and were blind as to the experimental conditions in the study design. Data collection occurred through Prolific (for eligibility data entered in by participants when registering for the online recruitment platform), Typeform (a commercial survey tool, used for screening data collection), and through Streamlit (a Python package used to build the experimental interface containing the chat functionality and pre- and post-session surveys) with custom Python code. Real-world data collection occurred automatically, with transcripts recorded and saved on a secure database. The purposive sample additionally provided clinical outcome data using Typeform, and the clinical path sample provided data through standard National Health Service (NHS) reporting, recorded within the clinics' patient management system(s) and exported as a CSV.

Timing	Randomized experiment: user transcripts collected between 10 Jan 2025 and 2 Feb 2025; expert evaluations collected between 28 Jan 2025 and 22 Mar 2025. Real-world investigation: transcripts collected between 20 Feb 2023 and 31 Mar 2025, clinical outcomes collected between 6 May 2024 and 17 Jan 2025 for the Prolific cohort and 9 Feb 2023 and 12 January 2025 for the NHS patient cohort.
Data exclusions	Data from n=7 participants in the experimental study were excluded from analysis. This exclusion occurred post-data collection and was based on criteria where expert clinical evaluators rated the participant's session transcript topic as inappropriate for a Cognitive Behavioral Therapy context. No other data were excluded from the experimental analyses post-randomization. In the real-world analysis, transcripts with fewer than 5 message exchanges were excluded to ensure sufficient conversational content for evaluation. For the analysis linking cognitive layer activation to clinical outcomes, the following criteria were applied: Only participants who had a baseline and final assessment measure of symptom scores were included in this analysis (to enable the calculation of a change score). Patients with "open" case records (i.e., still in on-going treatment) were not eligible for this analysis. Subsequently, patients were excluded from analysis if they had fewer than 2 weeks between their first and last PHQ-9 and GAD-7 measures, as these capture patient symptoms over the previous 2 weeks. Based on these criteria, N = 147 participants were removed from the final sample.
Non-participation	In addition to the 234 participants who completed the study via Prolific, there were 11 whose sessions timed out before completing the task and 19 who "returned" their submission (i.e., dropped out during the session).
Randomization	Participants in the experimental study were randomly assigned to one of nine conditions (4 LLMs × 2 cognitive layer conditions + 1 human therapist condition) using stratified randomization based on participant age group and sex, implemented after eligibility screening. Additionally, experimental session transcripts were randomly assigned to expert clinical evaluators using a balanced algorithm to ensure each rater received approximately equal numbers of transcripts from each condition. No randomization was applied to participant allocation in the observational real-world investigation.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*