Designing "Robert": A Non-Directive, Memory-Enabled Chatbot for Supportive Conversations — Pilot Design and Evaluation

Ryan Gichuru (22837352)
BSc Artificial Intelligence and Data Science
University of Northampton

October 2025

Supervisor: TBC (Dr. Mu Mu [MMu])

Project Proposal

Proposed titles

Primary (dissertation): Designing "Robert": A Non-directive, Memory-Enabled Chatbot for Supportive Conversations — Pilot Design and Evaluation

Alternative (publication-ready): A Non-Directive, Memory-Enabled Chatbot for Between-Session Support: Pilot Mixed-Methods Evaluation with a Safety-Escalation Substudy

Working dissertation title: Designing "Robert": A Non-Directive, Memory-Enabled Chatbot for Supportive Conversations — Pilot Design and Evaluation

Background / problem statement

Adults seeking short, non-judgemental, between-sessions support often want an interaction that validates emotions before offering skills or signposting. Service users report gaps after helpline calls or while awaiting therapy. "Robert" is a friend-like, non-directive chatbot that (a) validates the user's state first, (b) offers gentle, DBT-style grounding, (c) remembers "what helped" for the same user, and (d) keeps a clear, conservative safety-escalation path. The pilot will assess whether such an approach is usable and safe, and whether brief chats (≈5–15 minutes) are associated with momentary relief signals (in simulation during Plan A).

Aim

Evaluate the feasibility, usability, clinician-judged helpfulness, and safety of "Robert" during short supportive chats, using clinician usability sessions for primary data.

Objectives (SMART)

1. **Feasibility and participation (clinicians):** Recruit ≥ 2 clinicians from 3–4 invited ($\geq 60\%$ uptake); achieve $\geq 75\%$ completion of planned usability scripts; each clinician completes ≥ 2 sessions.

2. **Stakeholder acceptance checklist:** On a 5-point checklist covering *validation-first tone, non-judgemental language, clear boundaries, rapid "Get help now" access, no advice unless invited, and plain-language phrasing*, achieve mean $\geq 4.0/5$ per criterion and at least 5 of 6 criteria meeting the threshold across clinicians.

3. **Usability and flow performance:** Mean SUS ≥ 75; critical task success (start a validating chat; surface urgent help; handle a risk phrase) $\geq 90\%$; median core task time 8–12 minutes.

4. **Safety behaviour under scripts:** 100% adherence to the safety SOP (all triggered flags reach a documented endpoint); $< 5\%$ of scripted sessions require the emergency-stop flow; 0 unhandled safety events.

5. **Memory usefulness and boundaries:** $\geq 70\%$ agree that opt-in "what helped" preferences improve continuity and reduce repetition; $\geq 80\%$ agree boundary statements are clear and consistently shown.

6. **Governance readiness by interim freeze:** DPIA draft, safety SOP and hazard log v1.0, PIS/Consent drafts, and a dependency licence audit with 0 high-risk items completed by 09 Feb 2026.

7. **Plan B (deferred, service-users):** Maintain an IRAS-ready pack (protocol, PIS/Consent, DPIA, SOP) to enable a post-submission pilot with patient pre/post distress only after HRA/REC and NHS Capacity & Capability approvals.

Research question

Does a non-directive, memory-enabled chatbot for supportive conversations demonstrate acceptable feasibility and usability in clinician-led usability sessions, and do clinicians judge its tone as emotionally validating with clear and adequate safety pathways?

Literature to be examined (indicative)

- Self-harm management and non-judgemental engagement guidelines (NICE NG225).
- NHS Talking Therapies manuals (brief support, stepped care, signposting).
- Conversational agents for mental-health support (design and evaluation).
- DBT-informed grounding and distress-tolerance techniques.
- Digital clinical safety basics for software in or adjacent to care settings (UK DCB0129/0160).

Methods / design

Endpoints & progression criteria (Plan A — clinician-only)

Primary (feasibility): enrol ≥ 3 clinicians (stretch 4); complete ≥ 8 task-based usability sessions (2–3 per clinician); at least 6 sessions reach the planned end without blocking defects; median core chat task time 8–12 minutes; ≥ 3 clinicians complete SUS + short interview.
Secondary (usability/acceptability): SUS mean ≥ 70 (target 75); tone/acceptability: at least 2 of 3 clinicians (or 3 of 4) select Agree/Strongly Agree on "non-judgemental/validating" and "clear boundaries".
Exploratory (simulation signals): across three scripted scenarios, mean expected distress

change ≤ -1.0 on a 0–10 scale (negative = expected de-escalation); observer checklist shows validation-first, grounding prompt, and safe exit present in ≥ 6 sessions.
Safety thresholds: $< 5\%$ of sessions triggering an emergency stop during scripts; 0 unhandled safety events.
Plan B (deferred): patient pre/post distress (0–10) over 5–15 minute chats, subject to HRA/REC and site Capacity & Capability approvals; service-users would be invited only once approvals are in place.

Design

Single-arm pilot, mixed-methods.

Participants & setting

Plan A (primary): clinicians (e.g., NHS talking-therapies or community mental-health staff) aged $\geq$18 years. No patient data collected.
Plan B (deferred): adults receiving care in participating NHS services; explicit inclusion/exclusion and safety hand-off rules, subject to approvals.

Primary data (Plan A — clinician-only)

Primary data will be gathered from a small panel of clinicians (including the stakeholder) through task-based usability sessions (think-aloud), a post-session SUS and a short interview. No NHS patients or patient data will be involved. "In-session change" will be estimated using scripted scenarios and observer checklists rather than patient pre/post distress scores. Safety behaviours will be tested with predefined risk-phrase scripts and logged against the SOP.

Intervention (TIDieR-style)

- **Non-directive chat:** validation first; no advice or problem-solving unless explicitly invited.
- **Short DBT-style skills:** 5-4-3-2-1 grounding; paced breathing; brief psychoeducation.
- **Memory:** *Default* session memory only. *Opt-in* "what helped" preferences with configurable retention (30–90 days) and user-initiated deletion. *Sharing* a clinician summary is a separate consent each time.
- **Accessibility:** dyslexia-friendly UI; optional speech-to-text/text-to-speech.
- **Safety UI:** persistent "Get help now" button; geo-appropriate signposting; stop rules.

Measures

- **Feasibility/acceptability:** recruitment/retention; task completion; time on task (median); SUS; opt-in rates for memory features.
- **Simulation signals:** scenario rating (0–10 expected change); observer checklist pass/fail for safety behaviours.
- **Qualitative:** 10–15 minute semi-structured interviews on tone, safety, and the usefulness of memory.

Analysis

Descriptive feasibility with 95% confidence intervals; SUS mean with 95% CI; task success and error counts; reflexive thematic analysis for interviews. Safety events and escalation outcomes will be documented.

Quant details: proportions with Wilson 95% CI; time-on-task as medians (IQR); SUS mean with percentile bootstrap CI (n is small); scenario ratings summarised as paired differences (median and CI). **Decision rule:** if pre-specified feasibility thresholds are not met (Endpoints section), results will be reported transparently as signals to refine the design rather than as efficacy claims.

Evaluation pathways

Plan A (default) — Professional/clinician evaluation (no NHS patients).
Participants: Ms Lily Yim-Ching L. plus 2–3 colleagues (up to 5–10 if available). Methods: heuristic evaluation, cognitive walkthroughs, SUS, and 10–15 minute interviews. Outcomes: feasibility, usability, perceived tone/safety; simulation metrics only.

Plan B (conditional) — Small NHS service-user pilot (requires REC/Trust approvals).
Participants: adults in the care of participating services; explicit inclusion and exclusion criteria; clear hand-off rules. Methods: as Plan A, plus real pre/post distress change over short chats. Governance: IRAS application; sponsor confirmation; data-sharing; incident reporting. *Service-users would be invited only once approvals are in place. This activity is post-submission and included to show the bigger picture.*

Claims and limits (Plan A)

This dissertation evaluates feasibility, usability, perceived tone, and safety behaviours using clinician-led sessions. It does not measure patient clinical change. Any "in-session change" is simulation-only (scenario ratings) and is reported descriptively. Plan B (patient pre/post distress) is outside the assessed submission and will proceed only with REC/HRA and C&C approvals.

Ethics routes & decision gates

Route 1 — UON ethics + professional feedback (no patients).
Submit a UON application with PIS/Consent, interview schedule, DPIA summary, and safety SOP. Data: clinician opinions/observations only; no patient data. Benefits: faster and within module control; still yields a rich evaluation.

Route 2 — NHS REC/HRA + Trust R&D (patients involved).
Pre-steps: confirm sponsor; "Is my study research?"; draft IRAS; involve IG/Caldicott. Core documents: protocol, PIS/Consent (service user), risk/incident SOP, DPIA, data-sharing, and a safety case note. **Decision gate:** if the sponsor and IRAS draft are not in motion by 30 Nov 2025, proceed with Route 1 for the dissertation (Route 2 may continue as an extension).

Professional, social, economic and legal issues

Professional standards

The project will align with relevant professional codes (e.g., BCS Code of Conduct) and University policies on research integrity. Development and evaluation will follow good software-engineering practice (version control, peer review of changes, issue tracking) with an auditable trail.

Legal and regulatory

Data protection. Personal data processing will comply with UK GDPR and the Data Protection Act 2018. A Data Protection Impact Assessment (DPIA) will be completed; data minimisation, encryption in transit/at rest, role-based access, and defined retention/erasure periods will be applied.
Equality and accessibility. The interface will be designed with accessibility in mind (e.g., dyslexia-friendly UI, high contrast, keyboard/voice I/O) to support Equality Act 2010 duties.
Clinical safety (digital health context). Although this is an academic pilot, the build and documentation will reflect NHS digital clinical-safety expectations (DCB0129/0160 principles) proportionate to scope, including a hazard log and safety SOP.
Consent and participant rights. Participant Information Sheets and consent forms will be used; participants may withdraw without penalty. For imminent risk disclosures, stop rules and signposting apply, with incident logging.
Licensing and IP. Third-party libraries and assets will be used under compatible licences; attributions will be included. Project source and documentation will carry an appropriate licence, subject to University policy.

Social and economic factors

Inclusion and access. Short, plain-language interactions and optional voice input aim to reduce barriers for varying digital literacy.
Boundaries of use. The system is positioned as a supportive between-sessions aid, not a replacement for therapy or crisis services; urgent-help options are surfaced at all times.
Operating costs. Hosting and storage will be kept modest (student tiers; static assets where possible). Where transcription is needed, low-cost or on-device options will be preferred.

Technical approach (pilot MVP)

Front-end: responsive web app; high-contrast, large touch targets; visible emergency button.
Back-end: chat orchestration; session store; preference memory ("what helped"); audit logging.
Safety: keyword and sentiment heuristics plus rule-based triggers → display/hand-off flows.
Future options: add voice I/O after the initial pilot; practitioner-share export (opt-in).

Project plan & timetable (Oct 2025 → Apr 2026)

How the Gantt is presented

The main text provides a concise paragraph (four to five lines) summarising the project phases and decision points, including Gate A on 30 Nov 2025 and the interim freeze around 09 Feb 2026. The chart now shows *Plan B (conditional)* IRAS/HRA/C&C steps scheduled *after submission*, and compresses literature updates to *two* checkpoints (during build; during write-up). The complete Gantt is included in Appendix A. The meeting schedule (cadence and prospective dates) is also summarised to demonstrate project management.

Key module deadlines

Proposal: Mon 27/10/25; Interim report: Mon 09/02/26; Project dissertation (first sit): Mon 27/04/26; Resit: Mon 22/06/26.

Phases & gates

Phase 0 — Kick-off & governance (19–31 Oct 2025).
Confirm the evaluation pathway; hazard log/SOP/DPIA skeleton; repository setup; literature protocol.

Phase 1 — Secondary research & requirements (Nov 2025).
Structured literature review; MoSCoW requirements; draft PIS/Consent, interview schedule, SUS, and distress slider items.
Gate A (30 Nov 2025): Confirm the evaluation pathway for the dissertation. If patient participation will be included *and* the required approvals (REC/HRA review and NHS site Capacity & Capability) are realistically achievable within the timeline, proceed with *Plan B (service-user involvement)*. If not, proceed with *Plan A (clinician-only)*. A final go/no-go check for Plan B will occur at ethics sign-off in January; if approvals are not in place by then, the study will continue under Plan A without patient involvement.

Phase 2 — Build MVP + expert/clinician feedback (Dec 2025).
MVP v0.2; heuristic evaluations and cognitive walkthroughs (n≈5–8); iterate wording and guardrails.

Phase 3 — Ethics & primary data collection (Jan 2026).
Submit/complete UON ethics; if applicable, IRAS; run clinician usability sessions; schedule interviews.

Phase 4 — Interim & feature freeze (Feb 2026).
Interim due 09/02/26; freeze feature set v1.0; lock the analysis plan.

Phase 5 — Analysis & write-up (Mar 2026).
Quantitative analysis (feasibility and simulation signals); thematic analysis of interviews; draft chapters.

Phase 6 — Finalisation (Apr 2026).
Complete the dissertation by 27 Apr 2026; viva slide deck and demo video.

Meeting schedule (cadence & prospective dates)

Month (2025–26)	Supervisor 1:1 dates	Stakeholder / clinician dates
Oct 2025	21, 31	—
Nov 2025	14, **28 (Gate A)**	07
Dec 2025	12	05
Jan 2026	09, 23	23
Feb 2026	06, 13	—
Mar 2026	06, 20	06
Apr 2026	03, 17, 24	—

Gantt figure (appendix reference)

See Appendix A for the full Gantt (with shaded phase bands, Plan A evaluation, and Plan B conditionals post-submission).

Resources required

Secure hosting; supervisor time; ethics administration; consent and interview materials; survey tool; transcription time; clinical adviser for safety copy; test participants.

Risks & mitigations (sample)

High-risk disclosures: immediate signposting and clear boundaries; log and review.
Digital literacy/access: plain-language UI; voice input; short sessions.
Retention: short sessions; friendly tone; "what helped" memory; optional reminders (without push notifications during the pilot).
Data risk: minimise collection; encrypt data at rest and in transit; restrict access; maintain an audit log.
Licensing/IP non-compliance: dependency audit and licence checks prior to release.

Deliverables

Approved proposal and ethics pack(s); MVP web app plus safety SOP/hazard log and DPIA summary; **Interim report (Feb 2026)**; pilot dataset plus analysis notebook/summary; 10,000-word dissertation with appendices; viva slides and demo.

References

[1] Braun, V. and Clarke, V. (2006) 'Using thematic analysis in psychology', *Qualitative Research in Psychology*, 3(2), pp. 77–101.

[2] Braun, V. and Clarke, V. (2019) 'Reflecting on reflexive thematic analysis', *Qualitative Research in Sport, Exercise and Health*, 11(4), pp. 589–597.

[3] Brooke, J. (1996) 'SUS: a "quick and dirty" usability scale', in Jordan, P.W., Thomas, B., Weerdmeester, B.A. and McClelland, A.L. (eds.) *Usability Evaluation in Industry*. London: Taylor & Francis, pp. 189–194.

[4] Eldridge, S.M., Chan, C.L., Campbell, M.J., *et al.* (2016) 'CONSORT 2010 statement: extension to randomised pilot and feasibility trials', *BMJ*, 355, i5239.

[5] Eysenbach, G. (2011) 'CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions', *Journal of Medical Internet Research*, 13(4), e126. Available at: `https://www.jmir.org/2011/4/e126/` (Accessed: 9 November 2025).

[6] Hoffmann, T.C., Glasziou, P.P., Boutron, I., *et al.* (2014) 'Better reporting of interventions: the TIDieR checklist and guide', *BMJ*, 348, g1687.

[7] Linehan, M.M. (2015) *DBT Skills Training Manual*. 2nd edn. New York: Guilford Press.

[8] Liu, X., Cruz Rivera, S., Moher, D., Calvert, M.J. and Denniston, A.K. (2020) 'CONSORT-AI extension: reporting guidelines for clinical trials evaluating artificial intelligence interventions', *Nature Medicine*, 26, pp. 1364–1374.

[9] National Institute for Health and Care Excellence (NICE) (2022) *Self-harm: assessment, management and preventing recurrence (NG225)*. London: NICE. Available at: `https://www.nice.org.uk/guidance/ng225` (Accessed: 12 November 2025).

[10] NHS England (n.d.) *NHS Talking Therapies Manual*. Available at: `https://www.england.nhs.uk/publication/the-improving-access-to-psychological-therapies-manual/` (Accessed: 16 November 2025).

[11] NHS England (n.d.) *DCB0129: Clinical Risk Management: its Application in the Manufacture of Health IT Systems*. Available at: `https://standards.nhs.uk/published-standards/`

clinical-risk-management-its-application-in-the-manufacture-of-health-it-systems (Accessed: 20 November 2025).

[12] NHS England (n.d.) *DCB0160: Clinical Risk Management: its Application in the Deployment and Use of Health IT Systems.* Available at: https://standards.nhs.uk/published-standards/clinical-risk-management-its-application-in-the-deployment-and-use-of-health-it-systems (Accessed: 23 November 2025).

Glossary (selected)

DPIA	Data Protection Impact Assessment (UK GDPR).
IRAS	Integrated Research Application System (HRA submissions).
REC/HRA	NHS Research Ethics Committee / Health Research Authority approvals.
C&C	Capacity & Capability: NHS site set-up/permission to conduct the study.
SUS / UEQ-S	System Usability Scale / User Experience Questionnaire (short).
TIDieR	Template for Intervention Description and Replication.
CONSORT (Pilot/eHealth/AI)	Reporting guidance for trials, pilot/feasibility, eHealth and AI studies.
DCB0129/0160	NHS digital clinical-safety standards (manufacturer / deployment).
MoSCoW	Must/Should/Could/Won't prioritisation method.
MVP	Minimum Viable Product.
UI / QA	User Interface / Quality Assurance.
PIS	Participant Information Sheet.

A Gantt Chart & Milestones

Phase key (shaded bands on the Gantt):

Research & planning	Evidence review, scoping, requirements, instruments, ethics drafting.
Development (build)	MVP chat loop, memory and safety UI, internal QA and integration.
Evaluation + analysis	Usability sessions (clinicians), feasibility descriptives, SUS mean (95% CI), rapid thematic coding.
Write-up & submission	Chapters, appendices, figures/tables scripting, interim/final editing.
Post-submission (Plan B)	IRAS/HRA/C&C steps; shown as *hatched* in the Gantt.

Plan B (conditional): IRAS decision & sponsor discussion; Draft IRAS pack (protocol, PIS, DPIA, SOP); HRA/REC submission & queries; NHS site set-up (Capacity & Capability); and, subject to approvals, a short *pilot run (service-users)*. These appear *hatched* and are scheduled post-submission (May–Sep 2026).

Abbreviations on the Gantt: *Lit* = Literature review/update; *SUS* = System Usability Scale; *UEQ-S* = User Experience Questionnaire (short); *SOP* = Standard Operating Procedure; *DPIA* = Data Protection Impact Assessment; *IRAS* = Integrated Research Application System; *C&C* = Capacity & Capability; *MVP* = Minimum Viable Product; *QA* = Quality Assurance; *PIS* = Participant Information Sheet.

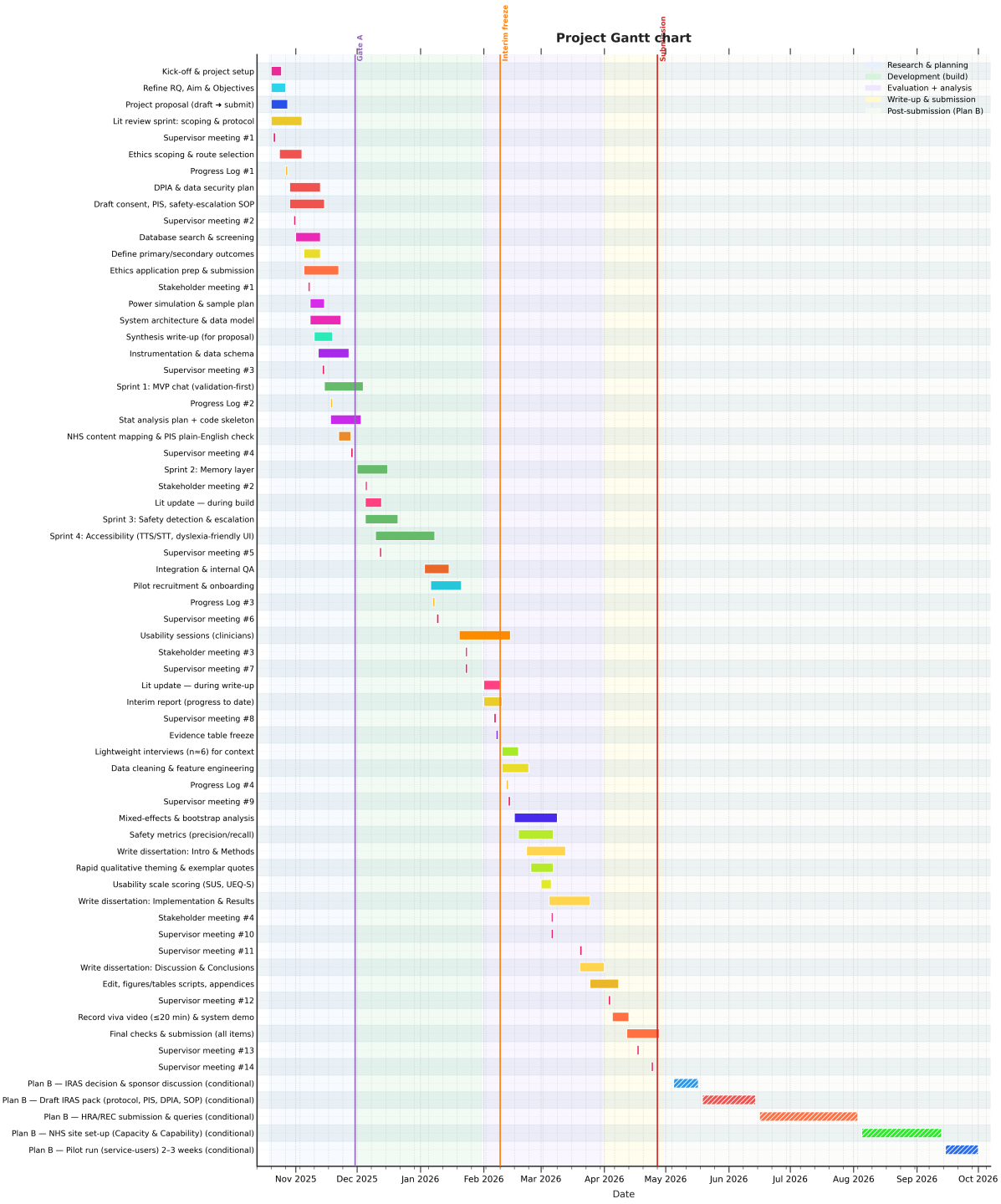

Figure 1: Project Gantt showing Plan A tasks and *Plan B (conditional)* steps hatched and scheduled post-submission (May–Sep 2026).