Adaptive Conversational Agents for Motivation and Engagement in Task-Related Goal Planning Dialogues

Chenxi Yu
Heriot-Watt University
dy2010@hw.ac.uk

David A. Robb
Heriot-Watt University
D.A.Robb@hw.ac.uk

The psychological quality of a conversational agent's dialogue style shapes user engagement and experience, yet controlled experimental evidence for motivational effects remains limited, particularly for brief task-related single-session interactions. This study presents a within-subjects experiment examining whether two Self-Determination Theory-grounded dialogue styles, a warm, autonomy-supportive *Friend* condition and a structured, competence-oriented *Coach* condition, can enhance task-related state motivation (users' in-session drive to pursue their stated goal) and user engagement compared to a non-adaptive *Generic* baseline condition. All three conditions employed identical session-scoped conversational memory, isolating dialogue style as the sole manipulated variable. The evaluation employed a mixed-method approach combining validated self-report scales, exploratory behavioural indicators and qualitative thematic analysis. Both adaptive conditions statistically significantly outperformed the SDT-neutral *Generic* baseline in this controlled comparison in terms of improved engagement and motivation (both with large measured effect sizes). The two SDT styles did not differ significantly, though preference, behavioural and qualitative indicators converge tentatively on competence support mattering more than relatedness in task-focused planning. These findings provide empirical evidence that psychologically informed dialogue design enhances engagement and motivation in brief goal-oriented interactions, with practical implications for conversational AI design.

conversational agents, self-determination theory, dialogue adaptation, motivation, engagement

1. INTRODUCTION

1.1. Motivation

Conversational agents (CAs) are increasingly deployed across productivity, well-being support and learning contexts, with growing potential to provide personalised support at scale. However, reviews of CA research identify persistent challenges: many CAs fail to meet user expectations, with fostering positive experiences and sustaining engagement remaining open challenges (Diederich et al. 2022). A growing body of evidence suggests that how an agent communicates is as consequential as what it communicates (Chaves and Gerosa 2021): conversational style shapes user perceptions, emotional responses and willingness to continue engaging. Despite this, most systems optimise for task completion while leaving the motivational quality of interaction largely unaddressed.

Self-Determination Theory (SDT) provides the theoretical framework for addressing this gap (Ryan and Deci 2000). SDT proposes that design features supporting autonomy, competence and relatedness enhance task-related state motivation (users' in-session motivational orientation toward their goal) and engagement; recent empirical work confirms that SDT constructs meaningfully predict why users prefer and engage with chatbots (Fröbel and Kenning 2025). Prior CA research demonstrates potential for agents to support psychological needs in clinical and long-term behaviour-change contexts (Bickmore and Picard 2005), while SDT-informed work has demonstrated how chatbots support autonomy, competence and relatedness in domain-specific applications such as language learning and coaching (Annamalai et al. 2023; Terblanche 2025). However, short-session productivity contexts such as daily goal planning represent common real-world applications yet remain under-explored in motivational CA research. Brief sessions (eight to ten minutes) present distinct challenges, including limited time to establish rapport and a need for

© Yu et al.

immediate motivational impact, creating the gap this study addresses.

1.2. Research Approach and Contributions

This short paper examines this gap through a controlled within-subjects experiment ($N = 24$) comparing two SDT-grounded dialogue styles, a warm, autonomy-supportive *Friend* condition and a structured, competence-oriented *Coach* condition, against a non-adaptive *Generic* baseline condition in brief goal-planning sessions. The evaluation (described in Section 3.4) combined the User Engagement Scale Short Form (UES-SF; O'Brien et al. 2018) and the Motivation for Cognition State scale (MFC-State; Blaise et al. 2021) as self-report questionnaires, alongside behavioural indicators and qualitative feedback. Specifically, the study addresses the following research questions:

- **RQ1:** Do SDT-grounded adaptive dialogue conditions significantly outperform a *Generic* baseline on user engagement in brief goal-planning sessions?
- **RQ2:** Do SDT-grounded adaptive dialogue conditions significantly outperform a *Generic* baseline on task-related state motivation in brief goal-planning sessions?
- **RQ3:** Do the two adaptive styles (*Friend* and *Coach*) differ significantly from each other on engagement and task-related state motivation?

Using a purpose-built, locally deployed goal-planning CA, this paper makes three contributions: (1) controlled experimental evidence that SDT-grounded prompt-based dialogue design significantly enhances engagement and task-related state motivation in brief single-session interactions, extending SDT CA research from long-term interventions into short productivity contexts; (2) a demonstration that motivationally effective dialogue adaptation is achievable through prompt engineering alone without fine-tuning or interface modification, making the approach directly deployable in existing LLM-based CA systems; and (3) a replicable mixed-method evaluation framework combining validated self-report scales, exploratory behavioural metrics and qualitative thematic analysis, suited to short-session motivational CA assessment.

2. BACKGROUND

2.1. Self-Determination Theory

We ground our work in SDT because it provides a theoretical framework with experimentally validated, manipulable constructs that map directly onto conversational design (Ryan and Deci 2000). SDT proposes that three psychological needs, autonomy (volition), competence (capability) and relatedness (social connection), when satisfied, enhance intrinsic motivation and engagement. Critically for our design, Sheldon and Filak (2008) showed in a $2 \times 2 \times 2$ factorial experiment that each need can be independently manipulated and independently predicted enjoyment, engagement and learning with largely additive effects. Yang and Aurisicchio (2021) demonstrated this operationalisability specifically within CA design through qualitative exploration of how competence, autonomy and relatedness are supported or undermined in user experiences. In our current study, the *Friend* and *Coach* conditions were selected because they map cleanly onto distinct SDT constructs: *Friend* supports autonomy and relatedness through warm, collaborative dialogue emphasising user choice; *Coach* supports competence through structured, goal-focused guidance and clear action steps.

2.2. Motivation, Engagement and Behavioural Indicators

Motivation and engagement were selected as primary outcomes because prior CA research predominantly measures general satisfaction or task completion, leaving task-related state motivation under-examined (Diederich et al. 2022). Task-related state motivation refers to users' moment-to-moment motivational orientation toward a specific goal, is separate from stable traits and is measured using the ten-item Motivation for Cognition State scale (MFC-State) (Blaise et al. 2021). Engagement captures the multidimensional quality of the interaction experience across focused attention, perceived usability and reward. It is measured using the nine-item User Engagement Scale Short Form (UES-SF) (O'Brien et al. 2018). Measuring both provides a fuller picture than either alone. Behavioural indicators such as conversational turn count can further strengthen construct validity by revealing engagement depth beyond self-reported states alone (Touré-Tillery and Fishbach 2014).

2.3. The Research Gap

Short-session goal-planning was selected as the task context for this study because it represents a common real-world use case for AI assistants that remains under-explored in motivational CA research. A systematic review of CA personalisation in healthcare found that personalisation features were commonly implemented without a theoretical framework and with limited evaluation of their motivational impact (Kocaballi et al. 2019), consistent with a broader pattern in which short-session, theoretically grounded approaches to motivational CA

design remain under-developed. Prior CA work in long-term relational contexts focuses on month-long interventions (Bickmore and Picard 2005) or domain-specific multi-session applications (Annamalai et al. 2023). Whether SDT-grounded dialogue design can produce measurable effects within a single brief interaction remains an open question.

3. METHOD

3.1. Conversational Agent Design

A purpose-built, locally deployed goal-planning CA was developed for this study using Python and Flask for the server layer, LangChain for conversation pipeline orchestration and the OpenAI GPT-4o-mini API for language model inference (temperature = 0.7). The application, session management and data storage were hosted on a local research machine, with all session and participant data stored in a local SQLite database. Conversational transcripts were deleted automatically at session end; only anonymised per-session behavioural metrics (turn count, session duration) were retained (see Appendix A for system architecture).

Three experimental conditions (*Generic*, *Friend*, *Coach*) were differentiated through distinct system prompts applied to an otherwise identical underlying pipeline. Session-scoped memory (ConversationBufferMemory) retained each participant's stated goals and prior exchanges within a session and was cleared automatically at session end; this held memory availability constant across conditions and ensured outcome differences reflected SDT linguistic framing rather than differential context accumulation. Condition presentation order was enforced server-side; out-of-order completion attempts were automatically rejected, ensuring Latin Square counterbalancing compliance. Conversational turn count and session duration were logged automatically in real time to the local database per session.

A particular design challenge was establishing a valid SDT-absent ground truth for the *Generic* condition. Because large language models fine-tuned with RLHF carry inherent warmth and collaborative framing, a naively uninstructed baseline would still contain partial SDT-supportive language, potentially confounding comparison with the SDT-grounded conditions. The *Generic* prompt therefore applied explicit constraints prohibiting collaborative pronouns, empathic validation, motivational framing and user-choice language, producing a rigorously SDT-absent condition rather than simply a less enthusiastic variant of the adaptive styles. The *Generic* condition presented a factual and neutral tone. The *Friend* (autonomy- and relatedness-supportive) condition used collaborative language and empathic validation. The *Coach* (competence-supportive) condition used first-person assertive framing and explicit capability language. The prompt templates were developed iteratively through pilot testing and constraint verification to ensure each condition reliably produced its intended dialogue style. Appendix B contains a table summarising the prompt templates, a link to the prompt templates used and dialogue snippets exemplifying each condition.

3.2. Participants and Data Collection

A total of 24 participants were recruited from the School of Mathematical and Computer Sciences at Heriot-Watt University via convenience sampling. Participants ranged in age from nineteen to fifty years (mean = 23.4, SD = 6.4, median = 21.5); the median better reflects the central tendency of the predominantly undergraduate sample. Of the 24 participants, nineteen were male (79.2 per cent) and five were female (20.8 per cent); twenty-three were undergraduate students and one was a postgraduate student. The gender distribution reflects the STEM population from which participants were drawn: the 20.8 per cent female representation in our sample closely mirrors the 20.3 per cent female representation recorded in the UK information technology workforce as of 2024 (WISE/IET 2024), consistent with national patterns of gender representation in STEM disciplines rather than a sampling artefact.

Initial recruitment yielded twenty-eight participants; four were excluded for providing uniform extreme ratings across all questionnaire items, leaving $N = 24$ as our usable sample. No further participants were recruited beyond the 24 usable, as this satisfied the Latin Square counterbalancing requirement: with three conditions, the design requires a sample size that is a multiple of six and $N = 24$ provides four participants per order. All procedures were approved by the Heriot-Watt University School of Mathematical and Computer Sciences Ethics Committee.

3.3. Design and Procedure

A within-subjects design was selected because it controls for individual differences while maximising statistical power at a small sample size. Six counterbalanced orders (four participants per order) ensured uniform condition distribution.

Participants completed three eight to ten minute goal-planning sessions, one per condition, in a single sitting of approximately twenty-five to thirty minutes total. Conditions were presented under neutral labels (A, B, C) to prevent expectation bias.

In each session participants brought a real personal goal of their own choosing (for example starting an exercise routine, learning a language or improving study habits) and used a browser-based chat interface to work with the agent towards an actionable plan. A session typically involved the participant stating their goal, the agent eliciting relevant context, and an exchange in which concrete steps were proposed and refined. Participants could accept the resulting plan or request changes. Appendix C reproduces the on-screen instructions and Appendix D presents worked dialogue examples for each condition.

After each condition session, participants completed post-condition questionnaires. A final qualitative survey collecting condition preferences and open-ended feedback was administered after all three conditions. The independent variable was dialogue style (*Generic*, *Friend*, *Coach*). The primary dependent variables were user engagement (UES-SF) and task-related state motivation (MFC-State).

3.4. Measures

Engagement was measured using the UES-SF (O'Brien et al. 2018), comprising nine items each rated on a 5-point response scale (1 = Strongly Disagree to 5 = Strongly Agree) across the Focused Attention, Perceived Usability and Reward subscales, yielding a continuous mean score per participant in the range [1.0 to 5.0]. The aesthetic appeal subscale was excluded as inappropriate for text-based CA evaluation. Task-related state motivation was measured using the MFC-State (Blaise et al. 2021), comprising ten items each rated on a 5-point response scale and yielding a continuous mean score per participant in the range [1.0 to 5.0]. Following the behavioural measurement principles outlined by Touré-Tillery and Fishbach (2014), exploratory interaction metrics such as conversational turn count were logged automatically to capture engagement depth beyond self-reported states alone (Venkatesh et al. 2018). Qualitative data comprising condition preferences and open-ended feedback were collected via final survey and analysed using inductive thematic analysis (Braun and Clarke 2006).

3.5. Analysis

All analyses were conducted in Python using `scipy.stats` (Virtanen et al. 2020) and `statsmodels.stats` (Seabold and Perktold 2010). As not all score distributions were normally distributed as assessed by Shapiro-Wilk tests (two of six condition-outcome combinations violated normality: UES-SF in *Friend*, $W = 0.881$, $p = .009$; UES-SF in *Coach*, $W = 0.846$, $p = .002$), non-parametric tests were applied consistently across all research questions. Internal consistency, assessed using Cronbach's alpha, was high for both scales prior to analysis (UES-SF $\alpha = 0.915$; MFC-State $\alpha = 0.893$), confirming that both instruments measured their respective constructs reliably in this sample.

Table 1: *Key outcomes by condition ($N = 24$). UES-SF = User Engagement Scale Short Form; MFC-State = Motivation for Cognition State scale. All values are continuous scores in the range from [1.0 to 5.0].*

	UES-SF		MFC-State	
Condition	M (SD)	Mdn	M (SD)	Mdn
Generic	3.19 (0.97)	3.11	2.68 (0.91)	2.45
Friend	4.12 (0.79)	4.28	3.58 (0.85)	3.70
Coach	4.35 (0.68)	4.61	3.67 (0.87)	3.60
*Adaptive (*pooled)	4.23 (0.61)	4.33	3.62 (0.77)	3.65

For RQ1 (engagement) and RQ2 (motivation), Wilcoxon signed-rank tests (Wilcoxon 1945) compared *Generic* against a pooled Adaptive condition, computed as each participant's mean of their *Friend* and *Coach* scores; the Wilcoxon statistic is reported as T. For RQ3 (style differentiation), a Friedman test (Friedman 1937) assessed omnibus differences across all three conditions, followed by post-hoc pairwise Wilcoxon tests with Holm-Bonferroni correction (Holm 1979). Effect sizes were computed as $r = |Z|/\sqrt{N}$ (Rosenthal 1991; Fritz et al. 2012) using the same formula across all RQs to enable direct comparison, with $r \geq .50$ interpreted as a large effect size (Cohen 1988). All tests were two-tailed at a significance level of .05, with bootstrapped 95 per cent confidence intervals reported for mean differences.

For the inductive qualitative analysis a single coder used open coding. As the codebook developed, additional passes through the data allowed further attribution of newly developed codes. Overarching themes were developed to represent the major topics.

4. RESULTS

RQ1: Engagement (UES-SF). Table 1 shows adaptive conditions substantially outperformed *Generic*. A Wilcoxon signed-rank test confirmed significantly higher engagement, $T(Wilcoxon) = 29.0$, $p < .001$, $r = 0.70$ (large), $\Delta = 1.04$ scale points, 95 per cent CI [0.63, 1.48]. Figure 1 illustrates this separation: *Generic* shows a lower, wider distribution while both adaptive conditions shift markedly upward with tighter spread, reflecting both the direction and consistency of the effect.

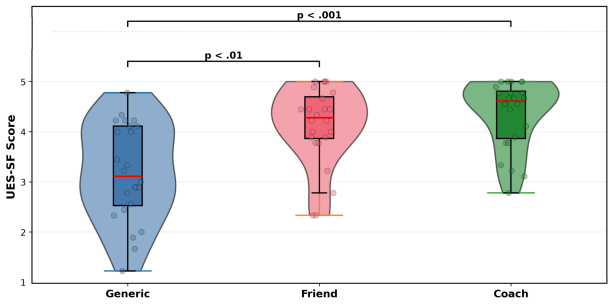

Figure 1: *UES-SF engagement scores by condition. Raincloud plots combine distribution shape, quartiles and individual data points. Generic shows lower engagement with wider spread; both adaptive conditions shift upward. Significance bars indicate RQ3 post-hoc comparisons (see text for details).*

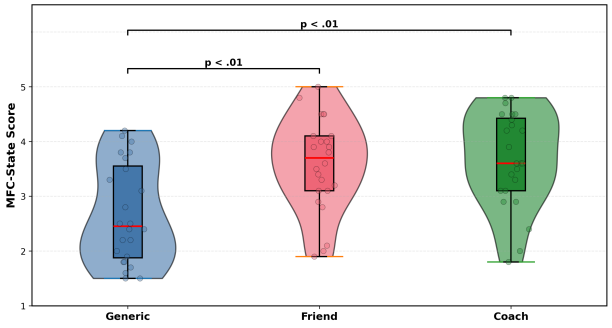

Figure 2: *MFC-State motivation scores by condition. Generic is concentrated in the lower range (2.0–3.5); Friend and Coach shift upward (3.0–4.5), with the two adaptive conditions overlapping substantially. Significance bars indicate RQ3 post-hoc comparisons (see text for details).*

RQ2: Motivation (MFC-State). Adaptive conditions outperformed *Generic* on task-related state motivation, $T(Wilcoxon) = 31.0$, $p = .001$, $r = 0.69$ (large), $\Delta = 0.95$, 95 per cent CI [0.46, 1.42]. Figure 2 shows *Generic* concentrated in the lower range (2.0–3.5) with *Friend* and *Coach* shifted upward (3.0–4.5).

RQ3: Style Differentiation. Friedman tests revealed significant omnibus effects for engagement, $\chi^2(2) = 21.07$, $p < .001$, $W = 0.44$ and motivation, $\chi^2(2) = 11.77$, $p = .003$, $W = 0.245$. Table 2 presents post-hoc pairwise comparisons. A consistent pattern of *Generic* $<$ *Friend* $\approx$ *Coach* emerged across both outcomes: both adaptive conditions significantly outperformed *Generic* with large effect sizes ($r > 0.6$) on both outcomes, while *Friend* and *Coach* did not significantly differ on either measure (UES-SF: $r = 0.46$, $p = .204$; MFC-State: $r = 0.13$, $p = .511$). The *Friend*-versus-*Coach* engagement difference, although not statistically significant, could indicate a potentially meaningful trend that warrants further investigation with a larger sample.

Table 2: *RQ3 post-hoc pairwise comparisons, Holm-Bonferroni corrected. ΔM = mean difference; r = effect size. Statistically significant pairwise differences in bold.*

	UES-SF			MFC-State		
Pair	ΔM	p	r	ΔM	p	r
Gen $<$ *Fri*	**+0.93**	**.007**	**.66**	**+0.90**	**.007**	**.64**
Gen $<$ *Coa*	**+1.16**	**.001**	**.80**	**+0.99**	**.007**	**.62**
Fri $\approx$ *Coa*	+0.23	.204	.46	+0.08	.511	.13

The qualitative questions and behavioural measures yielded results which provide more evidence about the differences and similarities between the two adaptive conditions. As a manipulation check, 85 per cent of participants reported conditions felt qualitatively distinct, confirming successful style differentiation; preferences favoured *Coach* (40 per cent), *Friend* (25 per cent), *Generic* (10 per cent), no preference (20 per cent) and prefer not to answer (5 per cent).

Exploratory analysis found *Friend* generated more conversational turns than *Generic* ($M = 9.8$ vs $M = 6.3$, $r = 0.76$, $p < .001$), consistent with *Friend*'s more exploratory dialogue style; *Coach* ($M = 7.5$) did not significantly differ from *Friend*.

Inductive thematic analysis of responses from 20 of 24 participants (four did not complete the final survey) identified three themes: interpersonal tone (conditions distinguished primarily by warmth, with *Friend* described as warm and human-like, *Generic* as plain or boring); social-task balance (participants valued agents balancing conversational engagement with task efficiency, with *Coach* frequently described as achieving this, for example '[*Coach*] felt more balanced between giving guidance and information, the others were either too simple or too much'); and planning quality (*Coach* was associated with clearer, more actionable plans, for example 'better structure and clearer steps').

5. DISCUSSION AND LIMITATIONS

5.1. Interpretation of Results

The engagement (RQ1) and motivation (RQ2) results suggest that psychological need support can be expressed through linguistic and interactional features of dialogue, including the collaborative phrasing and autonomy-supportive choice language of the *Friend* condition and the competence-oriented encouragement of the *Coach* condition (see Appendix D for dialogue snippets), and that these effects are achievable through prompt-based design without changes to interface layout or model architecture. This aligns with Chaves and Gerosa's (2021) conclusion that how a CA communicates is as consequential as what it communicates. This is

consistent with prior work showing that conversational style shapes user motivation and experience (Ciechanowski et al. 2019) and extends it by demonstrating that informationally adequate but stylistically neutral dialogue substantially underperforms SDT-grounded alternatives in a controlled within-subjects comparison.

Our findings also give quantitative support to Yang and Aurisicchio (2021), whose qualitative CHI study identified how CA design features support or undermine the three SDT needs; the present experiment tests those principles under controlled conditions and finds measurable motivational effects. This responds directly to the gap identified by Kocaballi et al. (2019), whose review found CA personalisation commonly implemented without theoretical grounding or evaluation of motivational impact.

Although *Friend* and *Coach* did not differ significantly on either outcome (Table 2), three independent indicators converge. *Coach* was numerically higher on engagement, was the most frequently preferred condition (40 per cent) and was associated in the qualitative data with clearer plans and a better balance of guidance and information; *Friend* elicited the longest conversations without converting them into higher scores. We read this cautiously as suggesting the three needs are not equally weighted in this task: in an explicitly task-focused context, competence support may carry more weight than relatedness support, and warmth without structure risks being experienced as verbosity. The alternative reading is that both styles satisfied overlapping needs, *Friend*'s warmth implicitly communicating competence and *Coach*'s structure conveying investment in the user, consistent with Sheldon and Filak's (2008) finding that need support is largely additive. For design practice the tentative implication is to default to competence-supportive framing in goal-directed CAs while keeping warmer styles user-selectable, since preferences were heterogeneous (*Friend* 25 per cent, no preference 20 per cent). An adequately powered study should test this directly.

5.2. Limitations and Future Work

Several limitations constrain interpretation. The sample consisted of predominantly male, undergraduate STEM students from a single institution, meaning that generalisation to broader populations should be done with caution. Effect sizes reflect the difference between SDT-aligned and SDT-neutral conditions; these should not be read as the expected gain over a default assistant and generalisability to real-world deployment contexts should be interpreted accordingly. The study examined a single task domain and a single LLM (GPT-4o-mini); the *Generic* condition may also have retained residual warmth inherent to RLHF-trained models, reducing the observed style difference. The study did not include per-condition measures of need satisfaction; the SDT mechanism account is therefore consistent with but not directly tested by the data, and accounts based on general warmth or pleasantness cannot be fully excluded without post-session autonomy, competence and relatedness ratings (omitted to preserve statistical power at our sample size). Our use of a within-participants design introduced risks of participant fatigue and effect carry-over. However, these were mitigated by keeping the typical total experimental session duration to approximately twenty-five to thirty minutes and balancing condition presentation order. Future work should examine longitudinal effects, individual-difference moderation (particularly personality-style fit (Hoegen et al. 2019)), per-condition need-satisfaction measurement and alternative task domains including health behaviour support and professional coaching.

6. CONCLUSION

This paper investigates whether conversational agents designed around Self-Determination Theory could enhance user motivation and engagement in brief goal-planning interactions. The results indicate that they can. Both adaptive conditions consistently outperformed a *Generic* baseline condition on engagement and task-related state motivation.

While differences between the *Coach* and *Friend* SDT conditions were not statistically significant, participants distinguished and appreciated them, and preference, behavioural and qualitative measures converged tentatively on competence-supportive framing as the stronger default for goal-directed tasks, a hypothesis a larger study should test directly.

The contributions are threefold: 1) a controlled experimental framework for isolating dialogue-style effects in LLM-based agents; 2) evidence that psychologically informed dialogue design produces substantial engagement and motivation effects in brief goal-planning sessions when only linguistic framing varies; and 3) a replicable evaluation approach for short-session motivational CAs. These findings have practical implications for the design of educational assistants, student support tools, health behaviour coaching applications and professional productivity tools where brief goal-oriented interactions are the primary modality. More broadly, they suggest that effective conversational AI requires not only task capability, but interaction design grounded in an account of human motivation.

REFERENCES

Annamalai, N., Eltahir, M.E., Zyoud, S.H., Soundrarajan, D., Zakarneh, B. and Al Salhi, N.R. (2023) Exploring English language learning via Chabot: A case study from a self determination theory perspective. *Computers and Education: Artificial Intelligence*, 5, 100148.

Bickmore, T.W. and Picard, R.W. (2005) Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, 12(2), 293–327.

Blaise, M., Marksteiner, T., Krispenz, A. and Bertrams, A. (2021) Measuring motivation for cognitive effort as state. *Frontiers in Psychology*, 12, 785094.

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

Chaves, A.P. and Gerosa, M.A. (2021) How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 37(8), 729–758.

Ciechanowski, L., Przegalinska, A., Magnuski, M. and Gloor, P. (2019) In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems*, 92, 539–548.

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences* (2nd edn). Hillsdale, NJ: Lawrence Erlbaum Associates.

Diederich, S., Brendel, A.B., Morana, S. and Kolbe, L. (2022) On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. *Journal of the Association for Information Systems*, 23(1), 96–138.

Friedman, M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200), 675–701.

Fritz, C.O., Morris, P.E. and Richler, J.J. (2012) Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, 141(1), 2–18.

Fröbel, L. and Kenning, P. (2025) What motivates people to prefer and love chatbots? An empirical investigation based on self-determination theory. In: *Proceedings of the 58th Hawaii International Conference on System Sciences*. Waikoloa, HI, 7–10 January 2025. 1762–1771.

Hoegen, R., Aneja, D., McDuff, D. and Czerwinski, M. (2019) An end-to-end conversational style matching agent. In: *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. Paris, 2–5 July 2019. New York: ACM. 111–118.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.

Kocaballi, A.B., Berkovsky, S., Quiroz, J.C., Laranjo, L., Tong, H.L., Rezazadegan, D., Briatore, A. and Coiera, E. (2019) The personalization of conversational agents in health care: Systematic review. *Journal of Medical Internet Research*, 21(11), e15360.

O'Brien, H.L., Cairns, P. and Hall, M. (2018) A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies*, 112, 28–39.

Rosenthal, R. (1991) *Meta-analytic procedures for social research* (Revised edn). Applied Social Research Methods Series, Vol. 6. Newbury Park, CA: Sage Publications.

Ryan, R.M. and Deci, E.L. (2000) Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–78.

Seabold, S. and Perktold, J. (2010) Statsmodels: Econometric and statistical modeling with Python. In: *Proceedings of the 9th Python in Science Conference*. Austin, TX, 28 June – 3 July 2010. 92–96.

Sheldon, K.M. and Filak, V. (2008) Manipulating autonomy, competence, and relatedness support in a game-learning context: New evidence that all three needs matter. *British Journal of Social Psychology*, 47(2), 267–283.

Terblanche, N.H.D. (2025) Smooth talking: Generative versus scripted coaching chatbot adoption and efficacy comparison. *International Journal of Human–Computer Interaction*, 41(14), 9109–9122.

Touré-Tillery, M. and Fishbach, A. (2014) How to measure motivation: A guide for the experimental social psychologist. *Social and Personality Psychology Compass*, 8(7), 328–341.

Venkatesh, A., Khatri, C., Ram, A., Guo, F., Gabriel, R., Nagar, A., Prasad, R., Cheng, M., Hedayatnia, B., Metallinou, A., Goel, R., Yang, S. and Raju, A. (2018) On evaluating and comparing conversational agents. *arXiv preprint* arXiv:1801.03625.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P. and SciPy 1.0 Contributors. (2020) SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), 261–272.

Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80–83.

WISE/IET (2024) *Workforce data: December 2024*. The IET WISE Campaign. Available from: `https://wise.theiet.org/resources-and-impact/research-and-statistics/workforce-data-december-2024/` [Accessed July 2026].

Yang, X. and Aurisicchio, M. (2021) Designing conversational agents: A self-determination theory approach. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Yokohama, 8–13 May 2021. New York: ACM. 1–16.

APPENDICES

A. EXPERIMENTAL PLATFORM ARCHITECTURE

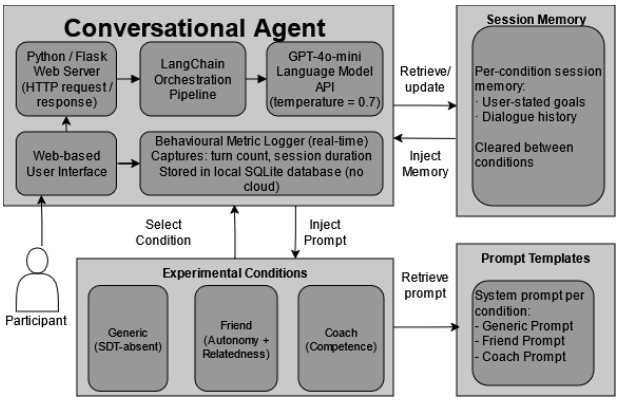

Figure 3: *Architecture of the locally hosted experimental platform. Three SDT-grounded conditions share an identical pipeline; condition-isolated session memory clears at session end. Full prompt templates are provided in Appendix B.*

B. PROMPT TEMPLATES

Table 3 summarises the three conditions. Full anonymised prompt templates are available at:

`https://osf.io/ecbhq/?view_only=e298d2d6c60a4e8abc771e562f7a365f`

Table 3: *Key design features of the three conditions.*

	Generic	***Friend***	***Coach***
SDT targets	All three needs suppressed (baseline)	Relatedness + Autonomy (primary); Competence (secondary)	Competence (primary); Autonomy (secondary)
Dialogue style	Neutral; encyclopaedic; third-person; no encouragement	Warm; casual; empathic; emoji for relatedness	Direct; assertive; first-person; commitment-seeking
Example utterance	*'Common frequency: 2–4 sessions weekly.'*	*'You've got this! Want me to help plan?'*	*'I'm giving you three sessions. Can you commit?'*

C. PARTICIPANT INSTRUCTIONS

Task instructions:

1. Choose a real personal goal (e.g. 'I want to start exercising').
2. Chat naturally with the agent, answer questions, share your thoughts.
3. Develop a plan with the agent.
4. Refine if needed using the Refine button or the input field.
5. Indicate commitment after receiving a plan (e.g. "Yes, I'll do this" or "No, this won't work").
6. Complete the post-session questionnaire before proceeding to the next condition.

Guidance:

- Be honest and authentic, there are no right answers.
- Express your true level of commitment.
- You may use the same goal or different goals across the three conversations.

D. DIALOGUE SNIPPETS

These snippets are from test dialogues with an author acting as participant. Same user inputs across all three conditions; responses differ solely by SDT-grounded system prompt. Emoji are denoted in brackets; **User** = participant, *Agent* = condition.

***Generic*, SDT-neutral baseline**

`User: 'I want to start exercising. I used to run but stopped six months ago.'`

`Agent: 'Standard exercise frequency is typically 2-4 sessions per week, 20-40 minutes each. Common types include cardiovascular activities, strength training, and flexibility work. People`

often begin with lower-intensity options after a break.'

User: 'I got busy and lost motivation.'

Agent: 'Fluctuations in motivation due to life demands are common. Factors may include time constraints or changes in routine. Many people find it helpful to identify obstacles and consider how to manage them while reintroducing exercise. Reflecting on previous experiences may provide useful insights.'

User: 'Just give me a plan.'

Agent: 'Common planning elements include activity selection, frequency determination, and duration estimation. Individuals typically establish frequency first (2-5 sessions weekly), then select activities based on preferences and resources. A typical structure involves a re-familiarisation phase, a practice phase, and an application phase. Habit formation generally takes 4-8 weeks.'

Friend, Autonomy and Relatedness-supportive

User: 'I want to start exercising. I used to run but stopped six months ago.'

Agent: 'Hey, that's awesome! [smile] What made you stop running in the first place?'

User: 'I got busy and lost motivation.'

Agent: 'Yeah, that totally happens. Life can get really hectic! So what's making you wanna pick it back up now?'

User: 'Just give me a plan.'

Agent: 'I get it! So here's what I'm thinking [smile] -- maybe 2-3 times a week, 20-30 minutes? Fit that in whenever works for you, morning or lunch break! And to keep it fun, mix it up -- run one day, walk or stretch another. There are great apps to help too! [strong] How does that sound? Wanna tweak anything?'

Coach, Competence-supportive

User: 'I want to start exercising. I used to run but stopped six months ago.'

Agent: 'Good. What's your current activity level? Are you doing any physical activity right now? I need to know where you're starting.'

User: 'I got busy and lost motivation.'

Agent: 'Alright. I'm addressing both. Here's what I'm giving you: three sessions per week, Monday, Wednesday, Friday mornings, 20 minutes each. Start with walking. That doesn't work? Tell me now and I'll adjust. Three sessions is the minimum. Can you commit?'

User: 'Just give me a plan.'

Agent: 'I'm setting you up with this: Week 1-2, gym Monday, Wednesday, Friday, 20 minutes -- start walking, just build the routine. Week 3-4, add bodyweight work after your walk -- 10 push-ups, 20 squats, 30-second plank each session. That doesn't work? Tell me now. Can you commit?'

E. AI USE DECLARATION

Two AI tools were used in this work. Claude (Anthropic) was used for grammar correction, vocabulary improvement and word choice during manuscript preparation; its use was limited to editing author-written text, with no content generation. GPT-4o-mini (OpenAI) provided the language model inference underlying all three experimental conditions, as described in Section 3.1.

All intellectual work, including research design, theoretical framing, experimental protocol, statistical analysis and interpretation of findings, is the authors' original contribution. AI assistance was limited to language editing and to the experimental platform described in this paper.