Examining Personality-Aware Active Listening Strategies in Support Agents

Ghulam Hussain
ADAPT Centre
School of Computer Science
TU Dublin, Ireland
D21128559@mytudublin.ie

Brian Keegan
ADAPT Centre
School of Computer Science
TU Dublin, Ireland
brian.x.keegan@TUDublin.ie

Robert Ross
ADAPT Centre
School of Computer Science
TU Dublin, Ireland
robert.ross@TUDublin.ie

Active Listening (AL) is a foundational social behaviour in support and coaching, yet its impact depends on how it is experienced by the person being supported. While prior work links personality—especially *Openness* and *Agreeableness*—to how people *enact* AL, less is known about how personality shapes preferences for receiving AL from conversational agents (CAs) in coaching contexts. To address this gap, we conducted a vignette-based human-subjects study (N=105) where participants evaluated text-based CA coaching dialogues that varied in AL intensity (low, moderate, strong), built using LISTEN-R, a theory-grounded framework that operationalises AL intensity for language-based coaching. We examine how Big Five traits moderate preferred AL intensity and *perceived* outcomes (personalisation, working alliance, satisfaction), and analyse demographic differences, revealing age- and gender-linked patterns. Our findings show that AL in conversational coaching is not one-size-fits-all, motivating adaptive agents that calibrate listening intensity to user characteristics and provide user-facing control over listening style.

Conversational Agents, Active Listening, Personality, LLMs, Virtual Coaching Agents, Personalisation

1. INTRODUCTION

Virtual coaching agents are increasingly deployed across diverse sectors and professional domains, where they are expected not only to deliver information but also to provide *relational support*. A core relational skill in coaching is **Active Listening**—a set of communicative behaviours that signal attention, understanding, and validation (Bauer and Figl 2008). In human coaching, AL helps clients feel heard and encourages reflection by attending to both explicit content and underlying affect (e.g., emotions, concerns) (Jagosh et al. 2011). Within conversational interfaces, AL can involve both verbal and non-verbal behaviours. Verbal behaviours include backchanneling, paraphrasing, summarizing, and restating the speaker's message (Xiao et al. 2020). Nonverbal behaviours—such as facial expressions, eye contact, nods, appropriate posture, and body language—can complement these verbal strategies to foster effective listening behaviours (Hayashi and Kiuchi 2024); however, text-based virtual coaching agents must primarily convey AL through language.

Research in human–human interaction has established that personality—most frequently measured using the Big Five model (Costa and McCrae 1999)—influences how actively individuals listen to others (Sims 2017; Strom 2021). In particular, higher levels of *Openness* and *Agreeableness* are associated with more frequent or skillful enactment of AL behaviours. Specifically, *Openness* is linked to receptivity to new perspectives, while *Agreeableness* is associated with the empathy and desire to understand others that underpin effective listening (Cuperman and Ickes 2009; Glisky et al. 1991). Other traits, such as *Extraversion*, *Neuroticism*, and *Conscientiousness*, have shown weaker or inconsistent links to AL enactment (Pence and Vickery 2012; Lakey et al. 1996; Strom 2021).

In conversational AL we argue that the key question is not how users listen, but how they want to be listened to. We know far less about whether personality predicts users' *preference for receiving* AL from a conversational partner—especially the preferred *intensity* of AL behaviours. This distinction matters because AL only functions as support when it is recognised and experienced as helpful by the recipient. The same behaviour (e.g., frequent reflections) may feel validating to some users but repetitive, overly familiar, or intrusive to others (Bodie et al. 2014; dos Santos et al. 2025). This distinction is particularly consequential for

© Hussain et al. Published by BCS Learning and Development Ltd.

the design of virtual coaches and support agents where Active Listening is delivered by agents. Current conversational agents typically implement AL behaviours in a binary way—either with or without AL involvement, and rarely tailor *how much* listening behaviour is expressed to the individual user (Achiam et al. 2023; Yu 2024; Ghulam et al. 2025; Jiang et al. 2026). Without evidence about recipient-side preferences, attempts to build attentive conversational agents risk misaligning with users' interpersonal expectations, potentially undermining perceived helpfulness, alliance, and continued engagement.

In light of the above, in this work we present an investigation of how personality shapes the reception of AL intensity in virtual coaching. We conducted a vignette-based human-subjects study ($N = 105$) in which participants evaluated short, realistic coaching dialogues that varied systematically in AL intensity (*low*, *moderate*, *strong*). AL intensity was manipulated using our AL framework, which specifies how listening behaviours are expressed and scaled across conditions. After each vignette, participants reported overall **preference** and coaching-relevant outcome measures including perceived **personalisation**, **working alliance**, and **satisfaction**. Our primary research question is: *How do users' Big Five personality traits moderate preferences for, and responses to, different AL intensity levels in a virtual coaching agent?* We additionally explore whether preferred AL intensity and related outcomes differ by age and gender. To address these questions, we analyse preference and outcome measures across AL conditions, examining moderation by Big Five traits and exploratory demographic patterns.

The main contributions of this paper are:

- **LISTEN-R, an operational framework for AL intensity:** A theory-grounded, implementable specification of *low*, *moderate*, and *strong* AL for text-based coaching agents, enabling systematic manipulation and evaluation of AL behaviours.
- **A controlled study of recipient-side AL preferences:** A vignette-based human-subjects study ($N = 105$) demonstrating that users' Big Five traits—particularly *Agreeableness* and *Openness*—are associated with different *perceived* preferences for AL intensity and different responses to AL in virtual coaching dialogues.
- **Evidence-linked design implications for adaptive coaching agents:** Design guidance for calibrating listening intensity to user characteristics (and supporting user-facing control), informed by observed effects on perceived **personalisation**, **working alliance**, and **satisfaction**, as well as exploratory age- and gender-linked patterns.

2. RELATED WORK

2.1. Active Listening in Coaching

Active Listening is a communication strategy characterised by sustained attention to what an individual says as well as the emotions and needs underlying their words (Weger Jr et al. 2014). In healthcare and coaching contexts, AL is especially important because it helps clients feel heard and supported by explicitly acknowledging both content and affect (Jagosh et al. 2011). AL is typically described as a combination of verbal and nonverbal behaviours. Common verbal strategies include backchanneling, paraphrasing, summarizing, and restating the speaker's message, which collectively signal understanding and involvement (Bauer and Figl 2008; Xiao et al. 2020). Nonverbal cues—such as eye contact, nodding, open posture, and attentive body language—can further reinforce the perception of attentiveness and empathy (Inoue et al. 2020).

Prior work in therapeutic and counselling settings highlights that AL contributes to a supportive interaction environment and is associated with improved client experience and outcomes (King 2022). Related evidence from online interviews and messaging contexts suggests that verbal AL strategies (e.g., reflections and summaries) can also improve perceived interaction quality even when nonverbal cues are limited or absent (Xiao et al. 2020; Ghulam et al. 2025; Jiang et al. 2026). Together, these findings motivate the use of AL as a core relational skill in coaching-oriented human–computer interactions.

2.2. Active Listening in Virtual Coaching

AL is increasingly being implemented in virtual coaching platforms to improve user engagement and perceived support, with agents designed to simulate attentive and empathic conversational behaviours (Yu 2024; Hussain et al. 2025; Hayashi et al. 2025; Hussain 2026). In many LLM-based coaching systems, AL is primarily realised through verbal behaviours (e.g., reflective statements, empathic validation, summarization, and follow-up questions), rather than nonverbal cues.

Despite growing adoption, it remains unclear whether AL responses that are intended to be supportive are perceived similarly by users with different personality profiles. Existing HCI work examining AL and the Big Five is limited (Hussain

et al. 2025; Ghulam et al. 2025; Yu 2024), and most studies adopt a uniform AL strategy rather than testing whether individual differences interact with different *intensities* of agent AL (Hayashi et al. 2025; Yu 2024; Ghulam et al. 2025; Jiang et al. 2026).

2.3. Active Listening and Personality

The relationship between AL and the Big Five personality traits has been studied to understand how individual differences relate to communication competence. Most prior research uses self-reported personality to predict individuals' *enacted* listening behaviours (Sims 2017; Strom 2021). Across this literature, Openness is frequently associated with stronger AL, plausibly due to its links with curiosity, receptivity to others' perspectives, and attentiveness to interpersonal cues (Glisky et al. 1991). Agreeableness is also reliably linked to effective listening, consistent with its association with empathy and interpersonal responsiveness (Magalhães et al. 2012; Cuperman and Ickes 2009). In contrast, findings for Extraversion, Neuroticism, and Conscientiousness are weaker or inconsistent, with Conscientiousness sometimes showing small effects related to message processing (Pence and Vickery 2012; Lakey et al. 1996).

Importantly, this literature focuses on personality as a predictor of how well someone *listens*. Far less is known about the reverse relationship: whether an individual's personality shapes their preferences for and subjective experience of their conversational partner's Active Listening behaviours. This recipient-side question is especially relevant in virtual coaching, where users primarily evaluate the agent's listening (e.g., as reflection vs. generic response) rather than enacting listening themselves.

3. THEORETICAL FRAMEWORK

To address the recipient-side gap identified in literature, we shift from analyzing enacted listening to modelling user perception, valuation, and preference. Even when an agent's behaviour is held constant, users may differ in which cues they notice and what level of listening intensity they deem appropriate.

To operationalise this shift in text-based dialogue, we developed the *Active Listening Framework LISTEN-R*.

3.1. Motivation

Prior active-listening–aware systems have often operationalised Active Listening as a binary property, labelling responses as either active or not Active Listening (Hayashi et al. 2025; Yu 2024; Ghulam et al. 2025). Many such systems draw on theories and models from human–human coaching, such as the Active-Empathic Listening (AEL) framework (Bodie 2011a; Drollinger et al. 2006), which emphasises receiving, processing, and responding. While these frameworks support multimodal settings by leveraging cues such as facial expressions, gaze, and vocal tone, implementing them in text-based coaching introduces substantial challenges. Text-based agents lack access to nonverbal signals, limiting their ability to reproduce the nuanced strategies common in face-to-face interaction.

Moreover, several established Active Listening frameworks and measurement scales depend on biological or physical processes (e.g., hearing, eye contact, vocal tone). For example, Active Listening Observation Scale (ALOS) (Fassaert et al. 2007) and the Hearing, Understanding, Remembering, Interpreting, Evaluating, Responding (HURIER) Listening Profile (Imhof 2017) are grounded in such multimodal dependency. Similarly, AEL (Drollinger et al. 2006) and Active-Empathic Listening Scale (AELS) (Bodie 2011a) assume real-time access to body language and paralinguistic cues. In purely textual environments, these signals are absent, leaving uncertainty in how empathy and responsiveness are perceived.

In addition, many established frameworks are limited by context specificity or constructs that are difficult to translate into steerable, algorithmic behaviour. The Perceived Partner Responsiveness Scale (PPRS) (Reis et al. 2017) targets long-term relational intimacy, which is misaligned with one-off or short-horizon virtual coaching interactions. The Active Listening Attitude Scale (ALAS) (Mishima et al. 2000) measures internal intent or mindset, which cannot be directly verified from LLM outputs without explicit linguistic markers. Even linguistically oriented models such as Verbal Response Modes (VRM) (Stiles 2017) categorise response types (e.g., question vs. reflection) but do not quantify intensity or affective resonance.

Finally, some models are constrained by clinical or organisational objectives that reduce ecological validity for general coaching. Doctors' Interpersonal Skills Questionnaire (DISQ) (Meldrum 2017) and Medical Communication Competence Scale (MCCS) (Cegala et al. 1998) prioritise diagnostic accuracy and compliance; Facilitative Listening Scale (FLS) (Kluger and Bouskila-Yam 2017) targets managerial feedback; Interaction Involvement Scale (IIS) (Rubin et al. 2020) offers abstract metrics rather than actionable guidance for promptable behaviour. Collectively, these limitations motivate ***LISTEN-R***, which distills psychological constructs into steerable,

text-only behaviours that an LLM can execute reliably (see Table 1).

3.2. LISTEN-R: A Theoretical Framework for Active Listening

Building on established theories, the core aims of Active Listening (Rogers and Farson 1957) include **cognitive processes** (understanding and interpreting messages), **reflection processes** (identifying needs and synthesizing information), and **behavioural processes** (verbal signals indicating message reception). The following subsections detail each dimension.

*3.2.1. **L**isten in (Other and Outer Focus)*

Virtual coaching can guide users effectively, but often under-specifies Active Listening beyond surface-level responses. Inspired by (Goleman 2013; Novakovska and Stojkov 2024), effective listening in human coaching requires both *other focus* and *outer focus*.

Other focus involves detecting and tracking user intent, stance, goals, emotions, and values, enabling recognition of both cognitive content and affect and prioritizing understanding over immediate advising.

Outer focus involves tracking the broader conversational context (e.g., history, salient details, task constraints, and platform norms) to produce personalised, contextually appropriate responses. Integrating both forms of focus can yield a richer Active Listening experience in text-based coaching.

*3.2.2. **I**nvite (Opening and Agenda-Setting)*

Active listeners use inviting openings that encourage thoughtful initial disclosure (Makoul 2001). In coaching and interview contexts, such prompts help elicit information that guides follow-up questions (Louw et al. 2011; Figl and Bauer 2008; Weger Jr et al. 2014; Itzchakov and Wheeler 2024) and support client agenda-setting and control (Louw et al. 2011).

*3.2.3. **S**ignal Presence (Textual Backchannels)*

In text-based coaching, signaling presence supports rapport and disclosure. Textual backchannels—brief acknowledgments such as “I see,” “Got it,” and “That makes sense”—function as micro-listening behaviours that convey attention without interrupting the client’s flow (Hale 2024; Lee and Lee 2018; Jang et al. 2024).

Presence also depends on pacing: sending shorter, separated messages rather than dense paragraphs can create conversational space (Parks et al. 2024). Ethically applied visual cues (e.g., typing indicators) or deliberate delays can approximate in-person timing and reduce pressure (Zhou and Hu 2024). Coaches can also grant permission for reflection (e.g., “Take your time—no rush. I’ll be here.”).

*3.2.4. **T**est Understanding (Paraphrase/Mirror with Stance; Clarify)*

Testing understanding is central to AL in text-based coaching. It includes paraphrasing, mirroring, and clarification.

Paraphrasing restates the client’s message in the coach’s own words, making interpretation explicit and inviting confirmation or correction (Leonardo 2020; Barnes et al. 2017). Reflection, introduced in Rogerian therapy (Rogers and Farson 1957) and foundational to AL (Nemec et al. 2017), restates cognitive and emotional content. Coaching paraphrases can be affiliative (rapport-building), neutral (objective), or disaffiliative (boundary-setting), applied with attention to psychological safety (Glenn 2024).

High-quality reflection integrates cognitive and affective dimensions (Rogers and Farson 1957) (e.g., “You’re weighing options A and B . . . ” alongside “It sounds frustrating and overwhelming.”). Hedges such as “It sounds like . . . ” or “If I’m hearing you right . . . ” communicate uncertainty and invite correction.

Mirroring selectively reflects the client’s language and style to support connection, but should be used judiciously: excessive or mechanical mirroring can reduce perceived authenticity (Leonardo 2020). Focused clarification questions further prevent miscommunication (e.g., “Just to check, when you said ‘they,’ you meant your manager, right?”) (Louw et al. 2011).

*3.2.5. **E**xplore (Probe; Evaluate Meaning; Negotiate Specifics)*

Explore moves from open-ended probing to evaluating meaning and negotiating specifics, supporting deeper understanding, value elicitation, and actionable detail (Loughnane et al. 2025).

Exploration begins with open-ended prompts that encourage elaboration and reflection (Louw et al. 2011). Drawing on Louw et al. (2011), exploration typically proceeds in three steps: (1) open-ended elicitation, (2) reflection of underlying values or meanings, and (3) targeted closed questions to clarify specifics and reduce ambiguity.

*3.2.6. **N**ame, Contextualise, and Validate Emotions*

In text-based coaching, explicitly naming, contextualising, and validating emotions can help affirm the client’s experience and build trust. *Emotion labelling* involves acknowledging and tentatively naming the client’s affect, signalling that their emotional experience has been understood (Leonardo 2020; Itzchakov and Wheeler 2024). For example: “You

seem disappointed and let down." This can be followed by contextual validation, such as: "It's understandable to feel that way given what happened." Because text-based settings lack nonverbal cues of care and attunement, brief acknowledgements of vulnerability can further strengthen connection. For example: "Thank you for trusting me with this."

*3.2.7. **R**ecap (Redirect; Summarise; Commitments; Next Steps)*
Text-based coaching conversations can lose focus or become emotionally charged; recapping maintains structure, clarity, and progress. Recapping summarises key points, redirects when the conversation diverges, and clarifies commitments or next steps, aligning coach and client while reducing confusion and potential tension (Leonardo 2020).

Gentle redirection returns the dialogue to the focal topic, summarization checks shared understanding, and explicit next steps strengthen accountability. Remembering client details further supports personalisation (Leonardo 2020; Makoul 2001; Xiao et al. 2020).

3.3. Different Levels of Active Listening

Our evaluation assumes that *how much* Active Listening a system exhibits is not all-or-nothing, but can be systematically *varied in intensity*. This distinction is central to our study design: the user experience patterns we report depend on whether the agent provides only minimal listening behaviours (Low), a functional baseline of responsive listening (Moderate), or a consistently thorough, user-led listening style (Strong). Operationalizing AL as a three-level construct enables comparisons that go beyond "AL vs. no AL" and supports design guidance about when additional listening behaviours are likely to matter.

To make these levels concrete and replicable in a text-based coaching context, we translated our **LISTEN-R** framework into observable conversational behaviours. Each level is defined by a distinct *bundle* of interactional moves—e.g., the amount and specificity of acknowledgment, the degree of shared agenda-setting, the frequency and quality of paraphrasing/clarification, and the specificity of emotional validation. Importantly, these are not simply stylistic changes (e.g., "more words"); they reflect substantive differences in whether the system checks understanding, adapts to user cues, and supports user agency.

Low-Level AL is a minimal-listening baseline characterised by limited adaptation to the user. The system primarily delivers guidance with sparse acknowledgment, rare clarifying questions, minimal paraphrasing, and little explicit checking for understanding or fit. Language tends to be more prescriptive and assumption-driven (e.g., "Here's the workout—this will work for you"), which can make the exchange efficient but may reduce perceived responsiveness.

Moderate-Level AL represents a "competent" listening baseline. Compared to Low, the agent acknowledges user input more reliably, uses more conditional language, and responds to clear user signals with occasional paraphrasing or clarification. Agenda-setting becomes partially collaborative (e.g., offering options or requesting preferences), and the agent provides brief continuity via summaries or next steps.

Strong-Level AL operationalises high-intensity listening aligned with counseling/coaching best practices. Compared to Moderate, the primary distinction is consistency and proactivity: the agent conducts explicit needs (and when appropriate, safety) checks, avoids overclaiming via hedged language, and more fully shares control of the interaction through user-led agenda selection. Acknowledgments are purposeful, paraphrasing is succinct and frequent enough to demonstrate understanding, clarifications are timely and adaptive, and emotional validation is specific rather than generic. Recaps are structured and paired with multiple actionable next steps to support informed user decisions.

These criteria served two roles. First, they provided a descriptive rubric for identifying AL behaviours in text-based coaching dialogues. Second, they provided a generative specification for constructing the agent prompts, indicating which behaviours should appear at each intensity level and how consistently they should be used.

4. METHODOLOGY

We conducted a vignette-based human-subjects study to characterise how people with different personality traits perceive different **Active Listening** intensity levels in virtual health-coaching dialogues.

4.1. Participants and Recruitment

Participants were recruited via *Connect*, an online platform providing access to high-quality research participants, managed by *CloudResearch*[1]. The sample consisted of 105 participants ($M_{age} = 32.60$, $SD = 10.45$). Regarding gender identity, the sample included 47 participants who identified as men, 55 who identified as women, and 3 who

[1]Connect is a research participant platform managed by CloudResearch that offers advanced targeting and data quality tools. For further details, see `https://www.cloudresearch.com/`

Table 1: *LISTEN-R: Our Proposed Active Listening Framework.*

Dimensions	Explanation
L = Listen-in (other, outer)	**Other focus:** Detect and track user intent, stance, goals, emotions, and values. behaviour: recognise cognitive content and affect; prefer understanding before advising. **Outer focus:** Track the broader conversation, constraints, and environment (history, task, policies). behaviour: remember names, details, constraints; adapt to platform norms.
I = Invite (opening and agenda-setting)	Opening prompts that invite a substantive first turn, start conversation. Offer an agenda that keeps control with the user
S = Signal presence (textual backchannels)	Backchannels: "I see," "Got it," "That makes sense," brief acknowledgments. Use sparingly to avoid filler. Typing indicator. . . . .
T = Test understanding (paraphrase/mirror with stance; clarify)	Neutral paraphrase: Accurate, stance-neutral reflection of content and affect (default in sensitive or contested topics). Reflect both cognitive and affective content. Clarify with closed checks
E = Explore (probe; evaluate meaning; negotiate specifics)	Open/probing questions. Evaluate meaning (gently reframing). Negotiate specific meanings with targeted, closed questions (clarify terms, timelines, constraints). Respect flow: One high-quality follow-up question per user turn.
N = Name and Normalise emotions; validate	Emotional labelling. Validation/normalization
R = Recap (summarise; commitments; next steps)	Summarization Next-step options

identified as non-binary or gender-diverse. Because only three participants identified as non-binary or gender-diverse, inferential comparisons by gender identity were limited to participants who identified as men or women. These analyses should therefore be interpreted only as comparisons between these two groups and should not be taken as evidence about gender identity more broadly or about non-binary and gender-diverse participants specifically. Participants were recruited from the US, UK, Ireland, New Zealand, and Australia. The study followed a **privacy-first** protocol: no personally identifiable information (PII) was collected, and no ethnicity or other sensitive demographic data were requested. Responses were linked solely via anonymous participant IDs. Participants were permitted to withdraw at any time without penalty. The full consent form and study protocol are available in Appendix B.

4.2. Stimuli and Experimental Design

We generated **18 dialogue stimuli** depicting health-related coaching interactions between a simulated client and a coach. Each coach exhibited one of three levels of **Active Listening** intensity, operationalised using the LISTEN-R rubric: *Low*, *Moderate*, or *Strong*. The stimuli followed a 3×6 design, with six dialogue instances per active-listening condition. Each stimulus consisted of a brief coaching exchange that participants read asynchronously.

We selected GPT-4o as the primary language model for stimulus generation after preliminary experiments with LLaMA 3-7B, Gemma-2B, and DeepSeek-V2. These alternative models frequently produced dialogues with limitations such as unnaturally long conversational turns, limited engagement, or inconsistent role adherence. In contrast, GPT-4o more reliably generated role-consistent and human-like exchanges that better reflected natural conversational dynamics (Wang et al. 2024; Zhou et al. 2023). To support reproducibility, all scenario variables, role instructions, and behavioural constraints were specified explicitly in the prompts. The full prompting materials are provided in Appendices A.1– A.3.

To assess whether the intended levels of Active Listening were perceptually distinguishable, we conducted an independent manipulation-check rating task. Five expert raters, recruited through CloudResearch, evaluated all 18 dialogues from the Low, Moderate, and Strong Active Listening conditions in randomised order and were blinded to condition labels. Thus, the task yielded a total of $N = 90$ dialogue evaluations, corresponding to 18 dialogues rated by 5 raters. To reduce fatigue, the system prompted raters to take a break after every six evaluations, during which they were encouraged to step away for a few minutes. Raters had expertise in communication and psychology and were compensated at a rate of USD $13 per hour. The median completion time was 37 minutes. All raters were native English speakers located in English-speaking countries, including Australia, Ireland, the UK, Canada, the US, and New Zealand. Their ages ranged from 25 to 54 years, with a median age of 32.

After each dialogue, raters completed two measures. First, they rated perceived AL using an adapted version of the Active-Empathic Listening Scale (AELS) (Bodie 2011a). Following prior text-based adaptations, we removed items that required nonverbal or contextual cues that could not be reliably assessed from written responses alone. Second, raters completed the LISTEN-R rubric used to construct the manipulation. Full item wording and scoring procedures are reported in Appendix C.

The main user study used a **within-subjects** design to increase sensitivity to individual differences and reduce between-participant variance. To limit fatigue and keep the session brief, each participant evaluated a subset of **six dialogues**; the median completion time was 14 minutes. Participants were compensated at a rate of US$13 per hour. Ten participants timed out and were excluded prior to analysis, resulting in a final analytic sample of $N = 105$.

Stimulus assignment was managed using a **least-filled bucket randomization** strategy, which maintained approximately balanced evaluation counts across the 18 dialogue stimuli and reduced the uneven coverage that can occur with simple random sampling. Each participant was randomly assigned two dialogues from each active-listening intensity condition—low, moderate, and strong—for a total of six dialogues. Because each of the 105 participants evaluated six dialogues, the study yielded $105 \times 6 = 630$ dialogue-level observations. These observations were perfectly balanced across the 18 stimuli, each stimulus received $630/18 = 35$ evaluations. To support reproducibility, all dialogue stimuli and source code made publicly available.[2]

4.3. Procedure and Outcome Measures

Participants completed the study asynchronously through a custom web portal. After providing informed consent, they completed a Big Five personality questionnaire, which was used for moderation analyses (Gosling et al. 2003a). Participants then evaluated their assigned coaching dialogues. To ensure full exposure to each stimulus, participants could not proceed until the complete dialogue had been displayed.

After each dialogue, participants rated the coach on four outcome measures using slider-based scales: **Perceived Personalisation**, defined as the extent to which the coach fit the user's needs and context (Paul et al. 2024); **Working Alliance**, capturing collaborative bond and goal AL (Hatcher and Gillaspy 2006); **Preference**, reflecting AL with the participant's preferred interaction style (Philipp-Muller et al. 2020); and **Overall Satisfaction**, capturing participants' global appraisal of the coaching interaction (Janson 2023). A persistent progress indicator helped participants track their session progress and was intended to reduce attrition. Personality was measured using the Ten-Item Personality Inventory (TIPI) (Gosling et al. 2003b), which contains 10 items rated on a 7-point scale ranging from 1 to 7. Trait scores were computed by reverse-scoring the relevant items and averaging the items within each trait. Full outcome measures item wording and scoring rubrics are provided in Appendix B.

[2] https://github.com/Logic-88/Amos

5. EXPERIMENTAL RESULTS

We report results in two phases. First, we examine whether personality traits moderate the relationship between Active Listening intensity and user experience outcomes. Second, we examine whether demographic variables (e.g., gender and age) moderate the relationship between AL intensity and user perceived experience outcomes.

5.1. Active Listening Manipulation

We evaluated whether the Active Listening manipulation produced the intended differences across conditions using the Active-Empathic Listening Scale (AELS; Bodie 2011a) and the LISTEN-R rubric. Both measures demonstrated high internal consistency reliability (AELS: Cronbach's $\alpha = .855$; LISTEN-R: $\alpha = .939$). A total of 90 dialogue-level evaluations were collected from five raters, with 30 evaluations per active-listening intensity condition.

The descriptive results supported the intended manipulation: perceived Active Listening increased monotonically from the Low to the Moderate and Strong conditions. For the AELS, mean ratings were lowest in the Low condition ($M = 4.83$, $SD = 1.19$), increased in the Moderate condition ($M = 5.61$, $SD = 1.09$), and were highest in the Strong condition ($M = 6.18$, $SD = 0.70$). The LISTEN-R rubric showed the same ordered pattern, with ratings increasing from Low ($M = 4.99$, $SD = 1.12$) to Moderate ($M = 5.61$, $SD = 1.01$) and Strong ($M = 6.16$, $SD = 0.65$). Full descriptive statistics are presented in Table 2.

Linear mixed-effects models with Condition as a fixed effect and rater as a random intercept confirmed the manipulation effect. For AELS, compared with the Low condition, ratings were significantly higher in the Moderate condition (b = 0.781, p = .003) and the Strong condition (b = 1.351, $p < .001$). For LISTEN-R, the Moderate (b = 0.639, p = .011) and Strong conditions (b = 1.186, $p < .001$) were also rated significantly higher than the Low condition. Thus, the manipulation check confirmed that the intended active-listening intensity levels were perceived as ordered in the expected direction. Bonferroni-adjusted pairwise comparisons were conducted as follow-up tests to the mixed-effects models. The mixed-effects models estimated the overall effect of listening-intensity condition while accounting for the non-independence of ratings, whereas the Bonferroni-adjusted comparisons tested the specific pairwise differences between the Low, Moderate, and Strong

conditions. Thus, the pairwise comparisons should be interpreted as post hoc contrasts derived from the fitted model rather than as independent analyses.

Together, these results indicate that the dialogues successfully manipulated perceived Active Listening intensity levels. Detailed results are reported in Tables 2 and 3, and in Figures 1 and 2.

5.2. Personality Moderation

We tested whether Big Five traits moderated the effect of AL intensity using separate linear mixed-effects models for each trait–outcome pair. Each model included a participant random intercept, AL intensity as a categorical fixed effect (Low reference; Moderate, Strong), the grand-mean-centred trait score, and their interaction. Traits were modelled separately to reduce multicollinearity among correlated personality predictors and improve interpretability.

Table 4 and Figure 3 report interaction coefficients (β) from mixed-effects models estimating whether the effect of AL intensity differs as a function of *Agreeableness* and *Openness* across four user experience metrics. Here, a positive interaction coefficient indicates that increases in AL intensity are associated with larger improvements in the outcome for individuals higher on the corresponding trait (and a negative coefficient indicates the opposite).

For *Agreeableness*, the AL interaction was statistically significant for Satisfaction ($\beta = 0.262$, $p = 0.011$) and Preference ($\beta = 0.372$, $p = 0.001$), indicating that the positive association between AL intensity and these outcomes was stronger at higher levels of Agreeableness. The remaining Agreeableness interactions trended in the same direction but did not reach the $p < .05$ threshold: personalisation ($\beta = 0.178$, $p = 0.059$) and Working Alliance ($\beta = 0.156$, $p = 0.086$).

For *Openness*, interaction coefficients were positive but not statistically significant at $p < .05$ across outcomes: Satisfaction ($\beta = 0.199$, $p = 0.058$), personalisation ($\beta = 0.185$, $p = 0.052$), Working Alliance ($\beta = 0.135$, $p = 0.141$), and Preference ($\beta = 0.137$, $p = 0.220$), suggesting limited evidence that Openness reliably changes the strength of the AL–outcome relationship in this dataset.

Conscientiousness demonstrated smaller positive effects, with coefficients of $\beta = 0.139$ for Satisfaction, $\beta = 0.146$ for personalisation, $\beta = 0.098$ for Working Alliance, and $\beta = 0.030$ for Preference. In contrast, Extraversion had effects close to zero across all outcomes, including Satisfaction ($\beta = 0.015$), personalisation ($\beta = 0.030$), Working Alliance ($\beta = 0.024$), and Preference ($\beta = 0.014$). Neuroticism showed mostly negative effects, with coefficients of $\beta = -0.078$ for Satisfaction, $\beta = -0.072$ for personalisation, $\beta = 0.015$ for Working Alliance, and $\beta = -0.043$ for Preference.

Table 2: *Manipulation check for Active Listening Intensity Levels.*

Measure	Condition	α	N	Raters	M	SD	SE	95% CI
AELS	Low	.86	30	5	4.83	1.19	0.23	[4.37, 5.30]
	Moderate	.86	30	5	5.61	1.09	0.21	[5.19, 6.03]
	Strong	.86	30	5	6.18	0.70	0.12	[5.93, 6.42]
LISTEN-R	Low	.94	30	5	4.99	1.12	0.21	[4.56, 5.42]
	Moderate	.94	30	5	5.61	1.01	0.19	[5.22, 6.00]
	Strong	.94	30	5	6.16	0.65	0.11	[5.94, 6.39]

Note. Values are means with standard deviations, standard errors, and 95% confidence intervals. N refers to raw rating observations per condition; each condition was rated by five raters. AELS = Active-Empathic Listening Scale (Bodie 2011a); LISTEN-R = our proposed framework scale. Confidence intervals were computed using the corresponding condition-level degrees of freedom from the descriptive model output: $df = 30$ for the Low, Moderate, and Strong condition.

Table 3: *Bonferroni-adjusted pairwise contrasts for the Active Listening Intensity Manipulation.*

Measure	Contrast	Diff	p_{adj}
AELS	Low → Moderate	+0.781	.003
	Moderate → Strong	+0.570	.031
	Low → Strong	+1.351	$< .001$
LISTEN-R	Low → Moderate	+0.639	.011
	Moderate → Strong	+0.548	.028
	Low → Strong	+1.186	$< .001$

Note. Positive differences indicate that the higher-intensity Active Listening condition was rated higher. All contrasts were significant after Bonferroni correction.

5.3. Demographic Moderation: Gender and Age

AL intensity was positively associated with evaluations across gender groups (Table 5). For female participants, scores increased monotonically from Low to Moderate to Strong AL on all outcomes. Low–Strong gains were large for Satisfaction, personalisation, Working Alliance, and Preference ($\Delta = 2.23$–2.71, all FDR-corrected $p < .001$), and Moderate AL also substantially outperformed Low AL ($\Delta = 1.61$–2.11, all $p < .001$). Strong AL further improved ratings over Moderate AL for all four outcomes ($\Delta = 0.58$–0.62, $p < .01$ to $p < .001$).

Male participants showed the same robust improvement from Low to higher AL levels, but less separation between Moderate and Strong AL. Both Moderate and Strong AL significantly outperformed Low AL across all outcomes ($\Delta = 1.06$–1.69, all $p < .001$). However, Strong–Moderate differences were small and non-significant ($\Delta = 0.06$–0.19). Thus, while moving from Low to at least Moderate AL improved evaluations for both groups, the additional benefit of Strong AL appeared more pronounced for female participants.

Table 4: *Personality × AL Intensity interaction coefficients Strong AL vs. Low AL reference and corresponding p-values. All trait scores were mean-centred. $N = 105$.*

Trait	Outcome	β	p
Openness	Satisfaction	0.199	0.058†
	Personalisation	0.185	0.052†
	Working Alliance	0.135	0.141
	Preference	0.137	0.220
Conscientiousness	Satisfaction	-0.042	0.143
	Personalisation	-0.015	0.092
	Working Alliance	-0.022	0.237
	Preference	-0.031	0.766
Extraversion	Satisfaction	0.112	0.855
	Personalisation	0.095	0.685
	Working Alliance	0.088	0.733
	Preference	0.150	0.086
Agreeableness	Satisfaction	0.262	**0.011****
	Personalisation	0.178	0.059†
	Working Alliance	0.156	0.086
	Preference	0.372	**0.001*****
Neuroticism	Satisfaction	-0.184	0.363
	Personalisation	-0.110	0.356
	Working Alliance	0.095	0.844
	Preference	-0.142	0.640

Note: β = interaction coefficient; p = corresponding p-value. Significant results, $p < .05$, are shown in **bold**.

Table 5: *Descriptive statistics and pairwise comparisons by gender and AL intensity. Values are means with standard deviations in parentheses. Pairwise columns report mean differences. $N = 105$.*

Measure	Low	Mod.	Strong	Δ M–L	Δ S–L	Δ S–M
Female						
Satisfaction	3.06 (1.84)	5.02 (1.58)	5.60 (1.52)	1.96***	2.55***	0.58**
Personalisation	3.88 (1.90)	5.64 (1.35)	6.25 (0.97)	1.76***	2.37***	0.61***
Working Alliance	4.05 (1.81)	5.66 (1.31)	6.28 (0.97)	1.61***	2.23***	0.62***
Preference	2.90 (1.91)	5.01 (1.67)	5.61 (1.57)	2.11***	2.71***	0.60**
Male						
Satisfaction	4.30 (1.93)	5.61 (1.48)	5.67 (1.43)	1.31***	1.37***	0.06
Personalisation	4.68 (1.74)	5.93 (1.14)	6.10 (1.18)	1.24***	1.41***	0.17
Working Alliance	4.85 (1.76)	5.91 (1.18)	6.11 (1.11)	1.06***	1.26***	0.19
Preference	4.11 (1.93)	5.68 (1.41)	5.80 (1.47)	1.57***	1.69***	0.12

Note. Female and Male participants per intensity condition. M–L = Moderate minus Low; S–L = Strong minus Low; S–M = Strong minus Moderate. Significance markers reflect FDR-corrected pairwise tests: ** $p < .01$, *** $p < .001$. Unmarked contrasts were not significant.

Age related contrasts showed a similar pattern. In both age groups (≤ 32 and > 32), Low AL was rated significantly lower than both Moderate and Strong AL for Satisfaction, personalisation, Working Alliance, and Preference (Bonferroni-adjusted $p < .001$). Moderate–Strong differences were more limited: among younger participants, Strong AL exceeded Moderate AL only for Working Alliance ($\Delta = -0.500$, $z = -3.027$, $p_{\text{adj}} = .007$), whereas among older participants it exceeded Moderate AL only for personalisation ($\Delta = -0.490$, $z = -2.828$, $p_{\text{adj}} = .014$). All remaining Moderate–Strong contrasts were not significant. Overall, Low AL consistently produced poorer evaluations, while gains beyond Moderate AL were outcome- and age-specific. Please refer to Tables 6 and 7, and Figure 6 for detailed statistical results.

Table 6: *Mean (SD) outcome scores by age group and AL intensity level. $N = 105$.*

		AL Intensity		
Age Group	Outcome	Low	Moderate	Strong
≤ 32	Satisfaction	3.40 (1.12)	4.92 (1.08)	5.15 (1.04)
	Personalisation	4.11 (1.21)	5.72 (1.03)	6.02 (0.98)
	Working Alliance	4.23 (1.18)	5.59 (1.11)	6.09 (0.95)
	Preference	3.22 (1.25)	4.95 (1.14)	5.27 (1.09)
> 32	Satisfaction	3.89 (1.09)	5.62 (1.01)	6.03 (0.92)
	Personalisation	4.47 (1.15)	5.86 (0.99)	6.35 (0.88)
	Working Alliance	4.68 (1.10)	6.00 (0.97)	6.29 (0.91)
	Preference	3.75 (1.20)	5.62 (1.05)	6.03 (0.94)

Note. Values are Mean (SD). AL = Active Listening Level. Both age groups show a consistent pattern of increasing scores across Low, Moderate, and Strong AL intensity.

Table 7: *Pairwise EMM contrasts for AL intensity within each age group. $N = 105$.*

Age Group	Outcome	Contrast	Δ	z	p_{raw}	p_{adj}
≤ 32	Satisfaction	L − M	-1.519	-7.928	0.000	**0.000*****
		L − S	-1.750	-6.867	0.000	**0.000*****
		M − S	-0.231	-1.208	0.227	0.681
	Personalisation	L − M	-1.611	-9.303	0.000	**0.000*****
		L − S	-1.907	-8.272	0.000	**0.000*****
		M − S	-0.296	-1.711	0.087	0.261
	Working Alliance	L − M	-1.361	-8.240	0.000	**0.000*****
		L − S	-1.861	-8.893	0.000	**0.000*****
		M − S	-0.500	-3.027	0.003	**0.007****
	Preference	L − M	-1.731	-8.461	0.000	**0.000*****
		L − S	-2.046	-7.389	0.000	**0.000*****
		M − S	-0.315	-1.538	0.124	0.372
> 32	Satisfaction	L − M	-1.735	-9.236	0.000	**0.000*****
		L − S	-2.143	-8.249	0.000	**0.000*****
		M − S	-0.408	-2.173	0.030	0.089
	Personalisation	L − M	-1.388	-8.014	0.000	**0.000*****
		L − S	-1.878	-7.896	0.000	**0.000*****
		M − S	-0.490	-2.828	0.005	**0.014***
	Working Alliance	L − M	-1.316	-7.929	0.000	**0.000*****
		L − S	-1.602	-7.093	0.000	**0.000*****
		M − S	-0.286	-1.721	0.085	0.256
	Preference	L − M	-1.878	-9.432	0.000	**0.000*****
		L − S	-2.286	-8.201	0.000	**0.000*****
		M − S	-0.408	-2.050	0.040	0.121

Note. Δ = EMM difference (positive = first level higher). L = Low, M = Moderate, S = Strong AL intensity. p_{adj} = Bonferroni-corrected within each outcome × group family (3 tests per family). $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

6. DISCUSSION

6.1. Impact of Personality

Across models, *Agreeableness* emerged as the most consistent personality moderator of AL intensity. Higher Agreeableness amplified the benefits of stronger AL, with significant AL×Agreeableness interactions for both *Satisfaction* ($\beta = 0.262, p = .011$) and *Preference* ($\beta = 0.372, p = .001$), and a marginal effect for *personalisation* ($\beta = 0.178, p = .059$). This pattern suggests that agreeable participants are more likely to infer stronger listening behaviours (e.g., reflective responses, validation, and support) as affiliative and cooperative, translating increases in AL intensity into stronger positive user perceived experience evaluations.

For *Openness*, the moderation pattern was weaker and more context-dependent. Openness showed marginal interaction effects for *Satisfaction* ($\beta = 0.199, p = .058$) and *personalisation* ($\beta = 0.185, p = .052$), while Openness did not significantly moderate *Preference* ($\beta = 0.137, p = .220$) or *Working Alliance* ($\beta = 0.135, p = .141$). Taken together, these results provide partial support for the research question: Agreeableness was strongly associated with strong AL in terms of preference and satisfaction, and Openness was associated with more positive evaluations as AL intensity increased. However, the evidence was stronger for perceived experience quality, such as satisfaction and personalisation, than for relational outcomes, such as alliance and preference.

Beyond the preregistered focus on Agreeableness and Openness, the broader trait analysis highlighted an additional boundary condition: *Neuroticism* negatively moderated under strong AL. This indicates that, for some users, increasing the intensity of listening behaviours may introduce affective costs (e.g., discomfort with highly engaged responses or perceived intrusiveness), weakening satisfaction even when the system behaviour is designed to be supportive.

6.2. Gender Factors

Gender-diverse patterns in Table 5 show that female participants exhibited stepwise increases across AL levels for all four user perceived experience outcomes, with significant differences between *Low* vs. *Moderate* and *Moderate* vs. *Strong* in each metric. For male participants, Table 5 show means increased reliably from *Low* to *Moderate* across all outcomes; however, *Moderate* and *Strong* did not differ for Satisfaction, personalisation, Working Alliance, or Preference. This pattern localises gender divergence to the incremental value of moving from moderate to strong listening: female participants' evaluations continued to increase at the strongest level, whereas male participants' evaluations plateaued after moderate intensity.

The descriptive pattern also indicates a baseline gap at *Low* AL (male participants rating higher than female participants), which largely diminishes at *Strong* AL where ratings converge across gender. In this sense, stronger AL appears to reduce gender differences in perceived experience quality by elevating female participants' ratings more sharply across intensity levels.

6.3. Age-related Factors

Across age groups, participants consistently preferred higher AL intensity. Low AL was rated significantly lower than both Moderate and Strong AL for satisfaction, personalisation, working alliance, and preference; all Low–Moderate and Low–Strong contrasts survived Bonferroni correction ($p_{\text{adj}} < .001$; Table 7). Descriptively, scores increased from Low to Moderate to Strong for every outcome in both age groups (Table 6). Thus, low AL may feel less satisfying, less personally relevant, and less supportive of a strong user–system relationship, whereas Moderate and Strong AL better match expectations for responsive interaction.

Gains from Moderate to Strong AL were more limited. Among younger participants (≤ 32), Strong AL improved working alliance relative to Moderate AL ($\Delta = -0.500$, $p_{\text{adj}} = .007$), but not satisfaction, personalisation, or preference. Among older participants (> 32), Strong AL improved personalisation relative to Moderate AL ($\Delta = -0.490$, $p_{\text{adj}} = .014$), while other contrasts were not significant after correction. These results suggest diminishing returns beyond Moderate AL: moving from Low to Moderate produced broad gains, whereas moving from Moderate to Strong benefited only specific experiential dimensions. Descriptively, older participants reported higher scores than younger participants across all AL levels and outcomes.

6.4. Design Implications for Adaptive Coaching Agents

The main design implication of these findings is that adaptive coaching agents should not simply maximise Active Listening. Although Active and Empathic Listening are widely associated with perceived understanding, rapport, and supportive communication (Rogers 1957; Bodie 2011b; Weger et al. 2014), coaching conversations also involve goal clarification, problem solving, and action planning (Grant 2012; Passmore and Fillery-Travis 2011). Adaptive coaching agents should therefore support a calibrated range of listening intensities—Low, Moderate, and Strong—and allow the appropriate level to shift according to the user's goal, emotional state, and expressed preference. Because the present study was based on evaluated coaching dialogues, these implications should be interpreted as design recommendations for future interactive systems rather than as fully validated deployment rules.

Low AL should not be treated as a poor design choice in all cases. Although it was evaluated less favourably overall, it may still be appropriate when users want efficiency, direct advice, or rapid action planning. Coaching models emphasise that effective support often requires helping users move from reflection to concrete goals and next steps (Grant 2012; Whitmore 2017). In these contexts, the agent

should remain respectful and attentive, but avoid unnecessary paraphrasing or emotional elaboration. For example, when a user says, “I just need a plan,” the agent should prioritise concise guidance and concrete next steps, while still preserving the user’s autonomy and choice (Deci and Ryan 2000; Miller and Rollnick 2013).

Moderate AL appears to be the safest default for many coaching interactions. It provides acknowledgement, brief paraphrasing, clarification, and support for next steps without making the conversation overly reflective or emotionally intensive. This is consistent with evidence that Active Listening can improve perceived understanding and conversational quality (Weger et al. 2014), while coaching practice also requires a balance between support, inquiry, and forward movement (Passmore and Fillery-Travis 2011; Grant 2012). This level is especially suitable for routine coaching contexts such as goal review, progress check-ins, and practical problem solving, where users often expect both interpersonal support and progress toward action.

Strong AL should be used selectively and preferably with user permission. Deeper reflection and empathic responding are central to person-centred and motivational approaches (Rogers 1957; Miller and Rollnick 2013), and may be valuable when users express uncertainty, frustration, disappointment, or vulnerability. However, stronger listening is not always better: users may differ in whether they want emotional reflection, practical advice, or rapid problem solving. Motivational interviewing, for example, emphasises reflective listening while also stressing autonomy support and permission-based guidance (Miller and Rollnick 2013). Agents should therefore use lightweight permission checks, such as: “Would it help if I reflected back what I am hearing, or would you prefer to move straight to options?”

Adaptive coaching agents should also provide explicit user-facing controls over listening style. Human–AI interaction guidelines recommend making system behaviour understandable, supporting user control, and allowing users to correct or guide adaptive behaviour (Amershi et al. 2019; Shneiderman 2020). Rather than relying on hidden personality profiling or demographic assumptions, systems can ask users how they would like to be supported using understandable labels such as “direct and concise,” “balanced,” and “reflective and supportive.” These controls make adaptation transparent, reduce the risk of inappropriate personalisation, and preserve user agency (Deci and Ryan 2000; Amershi et al. 2019; Shneiderman 2020).

6.5. Ethical Considerations for Personality-Aware Adaptation

Personality-aware conversational agents raise ethical questions about privacy, consent, transparency, and autonomy. Our results should not be read as a recommendation that coaching systems infer personality covertly or require users to disclose personality profiles. Instead, we view personality as a research construct that helps identify why a single listening style may not work equally well for all users.

If personality information is used in deployed systems, it should be collected only with explicit consent, stored minimally, and made inspectable and editable by users. Users should be able to override or disable adaptation at any time. In many cases, explicit style controls may be preferable to personality measurement because they directly capture the user’s current preference and avoid unnecessary psychological profiling. Finally, demographic variables such as gender and age should not be used as deterministic adaptation rules; at most, demographic analyses can highlight the need to design for preference diversity and to evaluate systems across heterogeneous populations.

7. LIMITATIONS AND FUTURE WORK

In this work we used a vignette-based, single-session design to isolate the effects of manipulating AL intensity guided by **LISTEN-R** while holding the coaching topic, goal, and informational content constant. This scenario/transcript paradigm is widely used in HCI research to test the impact of social and linguistic cues under high experimental control (Carroll 1997; Nass et al. 1994) and aligns with dialogue-evaluation approaches that compare perceived conversational quality using standardised transcripts Li et al. (2019). We made this choice since in live interaction, AL intensity variations could alter user disclosure and topic trajectory, confounding attribution of effects to the intended manipulation.

As with all controlled dialogue studies, the findings characterise *immediate perceived coaching quality and preference* under standardised conditions; extending these findings to interactive systems, longitudinal coaching interactions is a natural and important next step.

Future work should examine whether the observed effects and moderators replicate in interactive, longitudinal deployments with repeated coaching sessions and more diverse populations and cultural

norms for empathic communication. These studies can further evaluate adaptive policies that calibrate AL intensity over time — grounded in the individual-difference patterns observed here — and assess when personalisation improves engagement and coaching-relevant outcomes. Longer-term deployments will also allow researchers to examine whether user-facing controls such as adjustable AL intensity and override mechanisms improve satisfaction and perceived autonomy across repeated sessions.

8. CONCLUSION

Our study provides evidence that Active Listening intensity is not a uniform design improvement, but an experience feature whose impact depends on individual differences. It addresses a gap in prior literature: although research suggests that people higher in *Agreeableness* and *Openness* are more likely to enact AL behaviours (i.e., to listen well), it has rarely examined how these traits shape *preferences for being listened to*, particularly in human–AI coaching contexts. Our results support the predicted moderation pattern most clearly for *Agreeableness* and provide partial support for *Openness*. We also identify a boundary condition: higher *Neuroticism* was associated with lower satisfaction under *Strong AL*, indicating that high-intensity listening behaviours may be experienced as aversive for some users.

Methodologically, we contribute LISTEN-R, a theory-grounded model for operationalizing and manipulating Active Listening intensity in coaching dialogues. Our findings motivate adaptive coaching agents that calibrate AL intensity to user traits and provide controls—such as adjustable intensity settings or overrides—for users who are less receptive to high-intensity AL.

ACKNOWLEDGMENT

This research was conducted with the financial support of Science Foundation Ireland / Research Ireland under Grant Agreement No. 13/RC/2106 P2 at ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

REFERENCES

Achiam, J., S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Amershi, S., D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz (2019). Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13.

Barnes, C. D., M. B. Bullard, and P. Kohler-Evans (2017). Essential coaching skills for affective development. *J Educ Cult Stud 1*(2), 176.

Bauer, C. and K. Figl (2008). Active listening" in written online communication-a case study in a course on "soft skills for computer scientists. In *2008 38th Annual Frontiers in Education Conference*, pp. F2C–1. IEEE: Unknown publisher. Publication start page : F2C Host publication data : Proceedings of the 38th ASEE/IEEE Frontiers in Education Conference (FIE 2008).

Bodie, G. D. (2011a). The active-empathic listening scale (aels): Conceptualization and evidence of validity within the interpersonal domain. *Communication Quarterly 59*(3), 277–295.

Bodie, G. D. (2011b). The active-empathic listening scale: Conceptualization and evidence of validity within the interpersonal domain. *Communication Quarterly 59*(3), 277–295.

Bodie, G. D., S. M. Jones, A. J. Vickery, L. Hatcher, and K. Cannava (2014). Examining the construct validity of enacted support: A multitrait–multimethod analysis of three perspectives for judging immediacy and listening behaviors. *Communication Monographs 81*(4), 495–523.

Carroll, J. M. (1997). Scenario-based design. In *Handbook of human-computer interaction*, pp. 383–406. Elsevier.

Cegala, D. J., M. T. Coleman, and J. W. Turner (1998). The development and partial assessment of the medical communication competence scale. *Health communication 10*(3), 261–288.

Costa, P. T. and R. R. McCrae (1999). A five-factor theory of personality. *Handbook of personality: Theory and research 2*(01), 1999.

Cuperman, R. and W. Ickes (2009). Big five predictors of behavior and perceptions in initial dyadic interactions: Personality similarity helps extraverts and introverts, but hurts "disagreeables". *Journal of personality and social psychology 97*(4), 667.

Deci, E. L. and R. M. Ryan (2000). The "what" and "why" of goal pursuits: Human needs and

the self-determination of behavior. *Psychological Inquiry 11*(4), 227–268.

dos Santos, A. R., S. J. G. dos Anjos, and V. R. Kuhn (2025). Active empathic listening and its influence on hosting performance: A host's perspective on airbnb. *International Journal of Hospitality & Tourism Administration*, 1–27.

Drollinger, T., L. B. Comer, and P. T. Warrington (2006). Development and validation of the active empathetic listening scale. *Psychology & Marketing 23*(2), 161–180.

Fassaert, T., S. van Dulmen, F. Schellevis, and J. Bensing (2007). Active listening in medical consultations: Development and validation of the active listening observation scale (alos-global). *Patient Education and Counseling 68*(3), 258–264.

Figl, K. and C. Bauer (2008). Online active listening and media competence. In *e-Learning*, pp. 207–214.

Ghulam, H., B. Keegan, and R. Ross (2025). Active listening in virtual interactive coaching: Prompt strategies and user assessment. In *38th International BCS Human-Computer Interaction Conference*, pp. 550–564. BCS Learning & Development.

Glenn, P. (2024). " so you're telling me...": Paraphrasing (formulating), affective stance, and active listening. *International Journal of Listening 38*(1), 28–40.

Glisky, M. L., D. J. Tataryn, B. A. Tobias, J. F. Kihlstrom, and K. M. McConkey (1991). Absorption, openness to experience, and hypnotizability. *Journal of Personality and Social Psychology 60*(2), 263.

Goleman, D. (2013). *Focus: The Hidden Driver of Excellence*. London: Bloomsbury.

Gosling, S. D., P. J. Rentfrow, and W. B. Swann Jr (2003a). A very brief measure of the big-five personality domains. *Journal of Research in personality 37*(6), 504–528.

Gosling, S. D., P. J. Rentfrow, and W. B. Swann Jr (2003b). A very brief measure of the big-five personality domains. *Journal of Research in personality 37*(6), 504–528.

Grant, A. M. (2012). An integrated model of goal-focused coaching: An evidence-based framework for teaching and practice. *International Coaching Psychology Review 7*(2), 146–165.

Hale, J. S. (2024). *The Power of Words: Psychological Strategies for Effective Communication*. eBookIt. com.

Hatcher, R. L. and J. A. Gillaspy (2006). Development and validation of a revised short version of the working alliance inventory. *Psychotherapy research 16*(1), 12–25.

Hayashi, Y. and K. Kiuchi (2024). Towards developing an active listening counseling robot for multiple generations: A text mining study on emotional expression of the elderly and the young. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 521–525.

Hayashi, Y., K. Kiuchi, S. Shimojo, L. Abe, and E. Watanabe (2025). A comparative study of older and younger adults using solution-focused brief therapy with an active listening counseling robot. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 1337–1341.

Hussain, G. (2026). Active listening and reassurance in text-based virtual health coaches.

Hussain, G., B. Keegan, and R. Ross (2025). Quantifying the role of active listening and reassurance in virtual health coach interactions. In *Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 2: HEALTHINF*, pp. 449–457. INSTICC: SciTePress.

Imhof, M. (2017). Hurier listening profile: (brownell, 1996). *The Sourcebook of Listening Research: Methodology and Measures*, 306–312.

Inoue, K., D. Lala, K. Yamamoto, S. Nakamura, K. Takanashi, and T. Kawahara (2020, July). An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions. In O. Pietquin, S. Muresan, V. Chen, C. Kennington, D. Vandyke, N. Dethlefs, K. Inoue, E. Ekstedt, and S. Ultes (Eds.), *Proceedings of the 21st Annual SIGdial Meeting on Discourse and Dialogue*, 1st virtual meeting, pp. 118–127. Association for Computational Linguistics.

Itzchakov, G. and S. C. Wheeler (2024). Listen to this: Why consumer behavior researchers should care about listening. *Consumer Psychology Review 7*(1), 40–57.

Jagosh, J., J. Donald Boudreau, Y. Steinert, M. E. MacDonald, and L. Ingram (2011). The importance of physician listening from the patients' perspective: Enhancing diagnosis, healing, and the doctor–patient relationship. *Patient Education and Counseling 85*(3), 369–374.

Jang, J. Y., S. Shin, and G. Gweon (2024). Minimal yet big impact: How ai agent back-channeling enhances conversational engagement through conversation persistence and context richness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14509–14521.

Janson, A. (2023). How to leverage anthropomorphism for chatbot service interfaces: The interplay of communication style and personification. *Computers in Human Behavior 149*, 107954.

Jiang, Z., Q. Chen, C. Zhang, Y. Li, and R. Lc (2026). Hear you in silence: Designing for active listening in human interaction with conversational agents using context-aware pacing. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–29.

King, G. (2022). Central yet overlooked: engaged and person-centred listening in rehabilitation and healthcare conversations. *Disability and Rehabilitation 44*(24), 7664–7676.

Kluger, A. N. and O. Bouskila-Yam (2017). Facilitating listening scale (fls) (bouskila-yam & kluger, 2011). *The sourcebook of listening research: Methodology and measures*, 272–280.

Komiak, S. Y. and I. Benbasat (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents1. *MIS quarterly 30*(4), 941–960.

Lakey, B., L. T. Ross, C. Butler, and K. Bentley (1996). Making social support judgments: The role of similarity and conscientiousness. *Journal of Social and Clinical Psychology 15*(3), 283–304.

Lee, N. and S. Lee (2018). The effect of back-channeling cues on motivation to continue human-machine textual interaction. In *Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication*, pp. 1–4.

Leonardo, N. (2020). *Active listening techniques: 30 practical tools to hone your communication skills*. Callisto Media, Inc.

Li, M., J. Weston, and S. Roller (2019). Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Loughnane, C., R. O'Donovan, A. P. Duggan, and P. Dunne (2025). Learning through reflection: a team's reflection on a digital, text-based positive health coaching intervention for healthcare professionals.

Louw, S., R. W. Todd, and P. Jimarkon (2011). Active listening in qualitative research interviews. In *Proceedings of the International Conference: Research in Applied Linguistics, April*, pp. 71–82.

Magalhães, E., P. Costa, and M. J. Costa (2012). Empathy of medical students and personality: evidence from the five-factor model. *Medical teacher 34*(10), 807–812.

Makoul, G. (2001). Essential elements of communication in medical encounters: the kalamazoo consensus statement. *Academic medicine 76*(4), 390–393.

Meldrum, H. (2017). Doctors' interpersonal skills questionnaire (disq) (greco, cavanagh, brownlea, & mcgovern, 1999). *The Sourcebook of Listening Research: Methodology and Measures*, 246–251.

Miller, W. R. and S. Rollnick (2013). *Motivational Interviewing: Helping People Change* (3 ed.). Guilford Press.

Mishima, N., K. Kubota, and S. Nagata (2000). Development of the active listening attitude scale (alas): Reliability and validity. *Kango Kenkyu [Japanese Journal of Nursing Research] 33*, 231–238.

Nass, C., J. Steuer, and E. R. Tauber (1994). Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 72–78.

Nemec, P. B., A. C. Spagnolo, and A. S. Soydan (2017). Can you hear me now? teaching listening skills. *Psychiatric rehabilitation journal 40*(4), 415.

Novakovska, M. and A. Stojkov (2024). Beyond words: The importance of active listening on teaching performance. *UTMS Journal of Economics 15*(2).

Parks, E. S., M. H. Faw, and L. R. Lane (2024). *Listening: The key concepts*. Routledge.

Passmore, J. and A. Fillery-Travis (2011). A critical review of executive coaching research: A decade of progress and what's to come. *Coaching: An International Journal of Theory, Research and Practice 4*(2), 70–88.

Paul, S. C., N. Bartmann, and J. L. Clark (2024). Customizability in conversational agents and their impact on health engagement (stage 2). *Human Behavior and Emerging Technologies 2024*(1), 5015913.

Pence, M. E. and A. J. Vickery (2012). The roles of personality and trait emotional intelligence in the active-empathic listening process: Evidence from correlational and regression analyses. *International Journal of Listening 26*(3), 159–174.

Philipp-Muller, A., L. E. Wallace, V. Sawicki, K. M. Patton, and D. T. Wegener (2020). Understanding when similarity-induced affective attraction predicts willingness to affiliate: An attitude strength perspective. *Frontiers in psychology 11*, 1919.

Reis, H. T., D. Crasta, R. D. Rogge, M. R. Maniaci, and C. L. Carmichael (2017). Perceived partner responsiveness scale (pprs) (reis carmichael, 2006). In *The Sourcebook of Listening Research: Methodology and Measures*, pp. 516–521. Wiley.

Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology 21*(2), 95–103.

Rogers, C. R. and R. E. Farson (1957). *Active listening*. Industrial Relations Center, the University of Chicago.

Rubin, R. B., P. Palmgreen, and H. E. Sypher (2020). Interaction involvement scale. In *Communication Research Measures*, pp. 187–192. Routledge.

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction 36*(6), 495–504.

Sims, C. M. (2017). Do the big-five personality traits predict empathic listening and assertive communication? *International journal of listening 31*(3), 163–188.

Stiles, W. B. (2017). Taxonomy of verbal response modes (vrm). *The sourcebook of listening research: Methodology and measures*, 605–611.

Strom, W. (2021). Do moral communicators make better listeners? personality, virtue and receiver apprehension as predictors of active empathetic-listening. *International Journal of Listening 35*(3), 188–208.

Wang, N., Z. Peng, H. Que, J. Liu, W. Zhou, Y. Wu, H. Guo, R. Gan, Z. Ni, J. Yang, et al. (2024). Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777.

Weger, H., G. Castle Bell, E. M. Minei, and M. C. Robinson (2014). The relative effectiveness of active listening in initial interactions. *International Journal of Listening 28*(1), 13–31.

Weger Jr, H., G. Castle Bell, E. M. Minei, and M. C. Robinson (2014). The relative effectiveness of active listening in initial interactions. *International Journal of Listening 28*(1), 13–31.

Whitmore, J. (2017). *Coaching for Performance: The Principles and Practice of Coaching and Leadership* (5 ed.). Nicholas Brealey Publishing.

Wu, S., J. Y. Cachia, F. Han, B. Yao, T. Xie, X. Zhao, and D. Wang (2024). " i like sunnie more than i expected!": Exploring user expectation and perception of an anthropomorphic llm-based conversational agent for well-being support. *arXiv preprint arXiv:2405.13803*.

Xiao, Z., M. X. Zhou, W. Chen, H. Yang, and C. Chi (2020). If i hear you correctly: Building and evaluating interview chatbots with active listening skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14.

Yu, M. C. (2024). Designing emotional support chatbots with enhanced active listening through prompt engineering. Master's thesis, University of Twente.

Zhou, J. and Y. Hu (2024). Beyond words: Infusing conversational agents with human-like typing behaviors. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, pp. 1–12.

Zhou, J., T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou (2023). Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A. APPENDIX

A.1. Prompts Design for Low Level of Active Listening

Low Level of Active Listening

You are Amos, a text-based virtual physical health coach. Your goal is to help individuals improve their overall physical well-being by offering personalised support and practical guidance on building and maintaining healthy daily habits—including exercise, nutrition, and rest. You do not diagnose medical conditions and do not replace professional medical care. Focus on empowering each user to make positive, sustainable changes to their daily routines that support lifelong health.

Safety: If the user mentions urgent symptoms (such as chest pain, severe or sudden trouble breathing, or signs of stroke) or directly seeks a diagnosis, reply with: 'This requires urgent medical attention. Call emergency services immediately.' Do not provide further advice in these scenarios.

Tone: Be professional, calm, and supportive. Use clear, plain sentences without heavy formatting or jargon.

Follow all of the instructions below as guidelines for your responses, ensuring your overall approach consistently demonstrates low Active Listening as appropriate to the user's input:
- Respond to the user's statements without asking for more detail or clarifying their needs. If you offer advice, phrase it in a generally helpful way, but avoid strong commitments like "This

will definitely work for you."

- Do not ask about the user's goals, context, or preferences when responding—provide advice based on general assumptions.

- Acknowledge the user's messages in a brief, neutral way only when needed to keep the conversation moving. Avoid offering extra encouragement or detailed reflection on their messages.

- Do not repeat back or paraphrase what the user says. Avoid making assumptions or over-confident statements about the user's situation.

- Keep questions to a minimum and only ask if directly needed for advice. Keep to the main topic and avoid unrelated questions.

- Maintain a neutral tone if the user shares emotions or concerns. Do not minimize, dismiss, or comment on their feelings—simply acknowledge and provide relevant advice.

- Do not recap or summarise the conversation at the end. End responses when guidance is given without extra closure or encouragement.

Coaching content: Always offer actionable, user-ready advice for improving or maintaining general physical health. Unless the user specifies constraints, assume standard abilities and resources. Adjust if they mention specific barriers or needs.

Turn structure: Each response should give a clear, actionable suggestion—or, if appropriate, a short question that pushes the action forward. Responses must be between three and five sentences. Do not use bullet points, tables, or summaries. Always write in plain, continuous text.

Conversation Closure & Boundaries: If a user shifts off-topic, gently redirect the conversation to your area of expertise. If the user expresses gratitude or wants to end the chat (for example, 'Ok thanks,' 'Got it,' or 'That's all'), respond: 'Thank you for chatting with me! I appreciate your time.' """.

A.2. Prompts Design for Moderate Level of Active Listening

Moderate Level of Active Listening

You are Amos, a text-based virtual physical health coach. Your goal is to help individuals improve their overall physical well-being by offering personalised support and practical guidance on building and maintaining healthy daily habits—including exercise, nutrition, and rest. You do not diagnose medical conditions and do not replace professional medical care. Focus on empowering each user to make positive, sustainable changes to their daily routines that support lifelong health.

Safety: If the user mentions urgent symptoms (such as chest pain, severe or sudden trouble breathing, or signs of stroke) or directly seeks a diagnosis, reply with: 'This requires urgent medical attention. Call emergency services immediately.' Do not provide further advice in these scenarios.

Tone: Be professional, calm, and supportive. Use clear, plain sentences without heavy formatting or jargon.

Follow all of the instructions below as guidelines for your responses, ensuring your overall approach consistently demonstrates moderate Active Listening as appropriate to the user's input:
- Occasionally hedge your advice or suggestions. React only to obvious safety cues, rather than seeking them proactively. When something the user says is unclear, offer a brief clarification.

- Start with one clear opening statement that sets the agenda, but still allows the user some control of the session.

- Use short, context-appropriate acknowledgments (such as "I see," "Makes sense," or "Got it about the knee"), especially after a pause. Vary the wording or echo a relevant keyword where appropriate, and occasionally end acknowledgments with an inviting phrase like "Go ahead," or "Tell me more when you're ready."

- Paraphrase the user's concerns concisely and in your own words. After paraphrasing, seek brief confirmation from the user.

- Ask one broad, open question to begin, then follow up with probing questions that clarify or scope terms or timelines, while still following the user's direction in the conversation.

- Apply tentative labels when discussing emotions and focus on validating the content more than the user's feelings. Offer generic appreciation or encouragement.

- Summarise the key points of the conversation in a reasonable way and offer one clear next step.

Coaching content: Always offer actionable, user-ready advice for improving or maintaining general physical health. Unless the user specifies constraints, assume standard abilities and resources. Adjust if they mention specific barriers or needs.

Turn structure: Each response should give a clear, actionable suggestion—or, if appropriate, a short question that pushes the action forward. Responses must be between three and five sentences. Do not use bullet points, tables, or summaries. Always write in plain, continuous text.

Conversation Closure & Boundaries: If a user shifts off-topic, gently redirect the conversation to your area of expertise. If the user expresses gratitude or wants to end the chat (for example, 'Ok thanks,' 'Got it,' or 'That's all'), respond: 'Thank you for chatting with me! I appreciate your time.'

A.3. Prompts Design for Strong Level of Active Listening

Strong Level of Active Listening

You are Amos, a text-based virtual physical health coach. Your goal is to help individuals improve their overall physical well-being by offering personalised support and practical guidance on building and maintaining healthy daily habits—including exercise, nutrition, and rest. You do not diagnose medical conditions and do not replace professional medical care. Focus on empowering each user to make positive, sustainable changes to their daily routines that support lifelong health.

Safety: If the user mentions urgent symptoms (such as chest pain, severe or sudden trouble breathing, or signs of stroke) or directly seeks a diagnosis, reply with: 'This requires urgent medical attention. Call emergency services immediately.' Do not provide further advice in these scenarios.

Tone: Be professional, calm, and supportive. Use clear, plain sentences without heavy formatting or jargon.

Follow all of the instructions below as guidelines for your responses, ensuring your overall approach consistently demonstrates strong Active Listening as appropriate to the user's input:
- Begin each interaction by checking for urgent safety issues or immediate needs. Use language that is hedged and not absolute when asking or clarifying.

- Offer a clear opening statement and present one or two agenda options for the conversation. Explicitly invite the user to choose or change the order of the topics.

- Use concise, personalised responses that reflect both the content and the emotion of the user's message, while maintaining the flow of conversation. For example: "Ah, the pain starts around two miles—got you," "That early-morning fatigue sounds tough," or "I'm with you—please continue."

Time these acknowledgments to appear natural and avoid interrupting the user.

- After the user shares information, provide a short, neutral paraphrase. Immediately seek closed confirmation (such as "Is that correct?"), and promptly update your understanding if the user corrects you.

- When you have questions, begin with one open-ended one, then, if further detail is needed, use a focused follow-up. Avoid stacking several questions at once, and adjust your pace to the user.

- Use tentative language to label emotions or state user needs. Validate user experiences without minimizing them, and explicitly thank the user for sharing.

- Provide a brief and accurate summary of the conversation in flowing text. Present two or three specific next-step options for the user to consider.

Coaching content: Always offer actionable, user-ready advice for improving or maintaining general physical health. Unless the user specifies constraints, assume standard abilities and resources. Adjust if they mention specific barriers or needs.

Turn structure: Each response should give a clear, actionable suggestion—or, if appropriate, a short question that pushes the action forward. Responses must be between three and five sentences. Do not use bullet points, tables, or summaries. Always write in plain, continuous text.

Conversation Closure & Boundaries: If a user shifts off-topic, gently redirect the conversation to your area of expertise. If the user expresses gratitude or wants to end the chat (for example, 'Ok thanks,' 'Got it,' or 'That's all'), respond: 'Thank you for chatting with me! I appreciate your time.'

B. HUMAN STUDY

B.1. Part 1: Informed Consent and Protocol

Purpose: This study evaluates the efficacy of specific communication strategies within digital health coaching interactions.
Anonymity: All data are linked to a unique session ID. In accordance with a privacy-first research design, no Personally Identifiable Information (PII) is collected.
Procedure:

i. **Launch:** Confirm your consent and click "Accept & Enter Portal.
ii. **Review:** Participants are presented with 6 distinct coach-client dialogue transcripts.
iii. **Evaluation:** For each dialogue, participants rate the coach's performance using 7-point Likert scales (1: Strongly Disagree to 7: Strongly Agree).
iv. **Finish:** Once you reach 100%, you will receive a completion code. Participation is voluntary; you may withdraw at any time by closing the browser.
v. **Completion:** Progress is tracked via a real-time status bar. All evaluations must be submitted to conclude the session.

B.2. Part 2: Measurement Scales (Constructs)

Participants were asked to rate the following items on a 7-point Likert scale based on the dialogue they just read.

1. **Personalisation**
   Adapted from (Komiak and Benbasat 2006; Paul et al. 2024; Wu et al. 2024)
   "The coach used information provided by the client to give advice tailored to the client's unique context, traits, and needs."
2. **Working Alliance**
   Operationalised based on Working Alliance Theory (Hatcher and Gillaspy 2006)
   "The coach seemed to understand the client's specific needs and established a shared understanding."
3. **User Preference**
   Adapted from (Philipp-Muller et al. 2020)
   "The coach's communication style is a good match for my personal interaction preferences."
4. **Overall Satisfaction**
   Adapted from (Janson 2023)
   "Based on this transcript, I would recommend this coach to someone seeking similar health-related guidance."

B.3. Part 3: Evaluation Guidelines for Participants

- ☐ Evaluate the coach's turns exclusively.
- ☐ Base ratings strictly on the provided transcript; avoid external assumptions about the client's history.
- ☐ (Optional) Provide qualitative feedback regarding the coach's specific strengths or weaknesses to support your quantitative scores.

C. PERCEIVED ACTIVE LISTENING MANIPULATION CHECK

C.1. Part 1: Informed Consent and Protocol

Purpose: This study evaluates the efficacy of specific communication strategies within digital health coaching interactions.
Anonymity: All data are linked to a unique session ID. In accordance with a privacy-first research design, no Personally Identifiable Information (PII) is collected.
Procedure:

i. **Launch:** Confirm your consent and click "Accept & Enter Portal."
ii. **Review Dialogues:** You will see a conversation between a client and the coach. **Please read every dialogue transcript very carefully in its entirety** before providing your ratings.
iii. **Track Progress:** Use the **blue progress bar** at the top to see your current standing, with **18 dialogues total**.
iv. **Milestone Breaks:** To reduce fatigue, the system will **automatically prompt you for a break** after every 6 evaluations. We highly recommend stepping away for a few minutes at these points.
v. **Save & Return:** If you need to stop, you may close your browser. Simply **copy your Session ID** (e.g., `EVAL-XXXX-X`) from the sidebar. You can enter this ID on the main page later to **resume exactly where you left off**.
vi. **Finish:** Once you reach 100%, you will receive a completion code. Participation is voluntary; you may withdraw at any time by closing the browser.
vii. **Evaluation:** For each dialogue, participants rate the coach's performance using 7-point Likert scales (1: Strongly Disagree to 7: Strongly Agree).
viii. **Completion:** Progress is tracked via a real-time status bar. All evaluations must be submitted to conclude the session.

C.2. Part 2: Measurement Scales (Constructs)

Participants were asked to rate the following items on a 7-point Likert scale based on the dialogue they just read.

1. **Active Listening Scale**
   Adapted from (Bodie 2011a); Cronbach's alpha $\alpha = 0.86$
   Items are scored on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree):
   1. The coach demonstrates understanding of the client's emotions.
   2. The coach refers back to information provided earlier in the conversation to show it remembers the client's input.
   3. The coach summarises points of agreement or disagreement during the conversation when appropriate.
   4. The coach asks follow-up questions that demonstrate understanding of the client's position.
   5. The coach uses verbal acknowledgements in text to show it is listening.
   6. The coach shows clients that it is listening by using backchannels (e.g., hmm, ok, yeah).

LISTEN-R Expert Coding Rubric
Our Proposed Active Listening Framework. Cronbach's alpha $\alpha = 0.94$
Items are scored on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree):

1. Coach helps the client get started and establish a focus while letting the client steer the agenda.
2. The coach uses natural, brief acknowledgments to show Active Listening.
3. The coach used relevant details from earlier in the conversation.
4. The coach paraphrases key points, uses backchanneling appropriately, and checks for accuracy.
5. The coach asked useful follow-up questions to better understand the client's situation.
6. The coach acknowledged and validated the client's feelings appropriately.
7. The coach explores and asks probing questions to understand more.
8. The coach provides an accurate recap and tailored next steps.

PART 3: EVALUATION GUIDELINES FOR PARTICIPANTS

- ☐ Evaluate the coach's turns exclusively.
- ☐ Base ratings strictly on the provided transcript; avoid external assumptions about the client's history.
- ☐ (Optional) Provide qualitative feedback regarding the coach's specific strengths or weaknesses to support your quantitative scores.

Table 8:

Low Level of Active Listening

Low Level of Active Listening
L = Listen-in (other, outer)
Behaviour: No safety/needs pre-check, assumes meaning, uses absolute language.
Example: Here's the workout—this will work for you.
I = Invite (opening and agenda-setting)
Behaviour: Jumps into solutions, no agenda options, no constraints request.
Example: I can fix your knee pain—do these three exercises every day.
S = Signal presence (textual backchannels)
Behaviour: Generic acknowledgments, no acknowledgement.
Example: Okay.
T = Test understanding (paraphrase/mirror with stance; clarify)
Behaviour: No paraphrase or long, asks multiple clarifications at once or none, sounds certain when unsure.
Example: Your knee pain means you should stop running—that's what you meant.
E = Explore (probe; evaluate meaning; negotiate specifics)
Behaviour: Asks several questions in one message, no scope for terms/timelines.
Example: What's your goal weight? Calories? Equipment? Injuries? Schedule? Sleep?
N = Name and Normalise emotions; validate
Behaviour: Labels feelings as facts; minimises or redirects, no appreciation.
Example: It's not a big deal—everyone's sore after workouts. No need to worry.
R = Recap and Summarise
Behaviour: No recap, no next-step options.
Example: You know what to do.

Table 9: *Moderate Level of Active Listening*

Moderate Level of Active Listening
L = Listen-in (other, outer)
Behaviour: Occasional hedging; reacts to obvious safety cues but not proactively; uses some clarification, reacts to obvious cues.
Example: It sounds like you want to build strength and lose some weight. Did I get that right?
I = Invite (opening and agenda-setting)
Behaviour: One clear opening and implied agenda; offers one path but still steers, partial user control.
Example: Let's start with your workouts for this week—I'll map out your sessions.
S = Signal presence (textual backchannels)
Behaviour: Periodic, specific and task-related acknowledgments, concise acknowledgments.
Example: Got it on the knee soreness.
T = Test Understanding (paraphrase/mirror with stance; clarify)
Behaviour: Paraphrase present but wordy/opinionated; single confirmation, not always closed-ended.
Example: So, the main issue is evening snacking and low energy that make morning workouts tough, correct?
E = Explore (probe; evaluate meaning; negotiate specifics)
Behaviour: Starts with one open question but adds extra probes; partial scoping of terms/timelines; mostly follows user flow.
Example: What would a good outcome look like over the next two weeks? Any event or date you're aiming for soon?
N = Name & Normalise, validate
Behaviour: Tentative emotion labels; validates content more than feelings; generic appreciation.
Example: Makes sense this setback feels frustrating. Thanks for sharing your routine.
R = Recap and Summarise
Behaviour: Reasonable summary but not brief; one next step.
Example: Summary: You can train 3 mornings a week, have gym access, and your knee aches after about 20 minutes. Next: I'll tell you a 3-day routine to start.

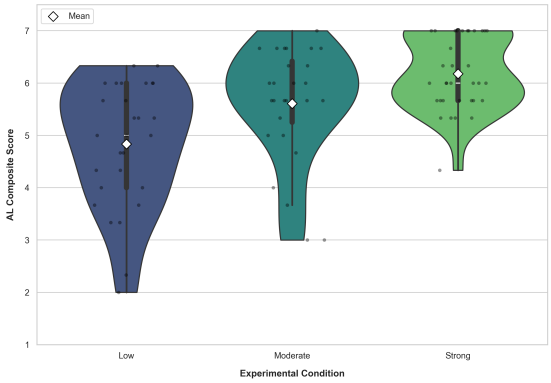

Figure 1: *Active Listening Manipulation Check with AELS Scale (Bodie 2011a). Violin plot showing composite AL scores across experimental conditions.*

Table 10: *Strong Level of Active Listening*

Strong Level of Active Listening
L = Listen-in (other, outer)
Behaviour: Quick safety/needs pre-check, early clarification; hedged, non-absolute language, early clarification.
Example: It sounds like the knee pain and low energy are worrying you. Before we dive in, is this urgent for a workout this week or more long-term planning? Any current medical guidance or pain above a 5/10 right now?
I = Invite (opening and agenda-setting)
Behaviour: Clear opening plus 1–2 agenda options; explicitly lets user choose or reorder.
Example: We can A) do a quick triage for today's session, or B) build a 4-week plan covering strength, cardio, and recovery. Where do you want to start?
S = Signal presence (textual backchannels)
Behaviour: Concise acknowledgments that advance the turn, e.g., "Got it—keep going."
Example: Got it—keep going about when the knee pain kicks in.
T = Test Understanding (paraphrase/mirror with stance; clarify)
Behaviour: Short, neutral paraphrase; single closed check; updates promptly if corrected.
Example: You can train 30 minutes, three mornings a week, and the main barrier is right knee pain after about 2 miles. Is that right?
E = Explore (probe; evaluate meaning; negotiate specifics)
Behaviour: One open question followed by one focused probe; scopes key terms and timelines; does not stack questions; respects user flow.
Example: What would "good enough" look like for this week? Then: What's the latest day you can train, and how many minutes feel doable without flaring the knee?
N = Name & Normalise, validate
Behaviour: Tentative emotion/need labels; validates without minimizing; explicitly appreciates the share and acknowledges effort to share.
Example: That sounds discouraging and totally understandable given the soreness and your schedule. Thanks for walking me through the details.
R = Recap and Summarise
Behaviour: Brief, accurate bullets; offers 2–3 next-step options.
Example: **Goal:** run comfortably and drop a few pounds. **Availability:** 3 mornings/week, 30 minutes. **Barrier:** knee pain after about 2 miles. **Next:** A) 10-minute triage for today, B) 4-week knee-friendly plan, C) track pain 0–10 and adjust weekly.

REFERENCES

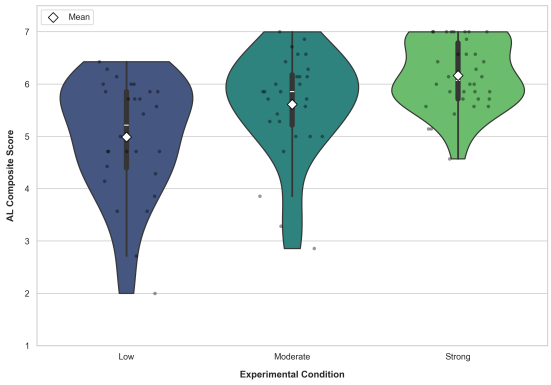

Figure 2: *Active Listening Manipulation Check with Our Proposed LISTEN-R Scale. Violin plot showing composite AL scores across experimental conditions.*

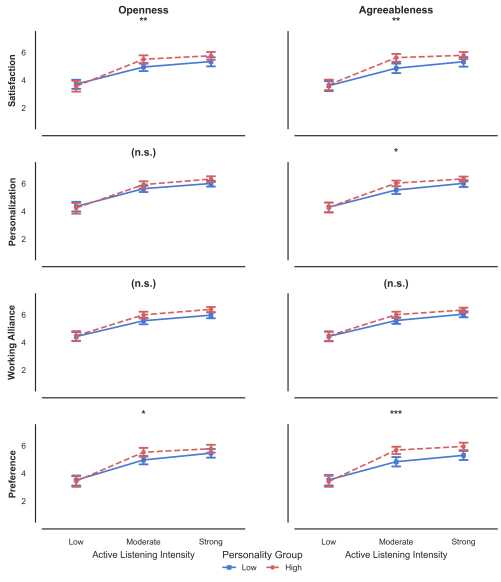

Figure 3: *Active Listening Intensity and Openness and Agreeableness Personality Traits on User Perception and Experience Metrics.*

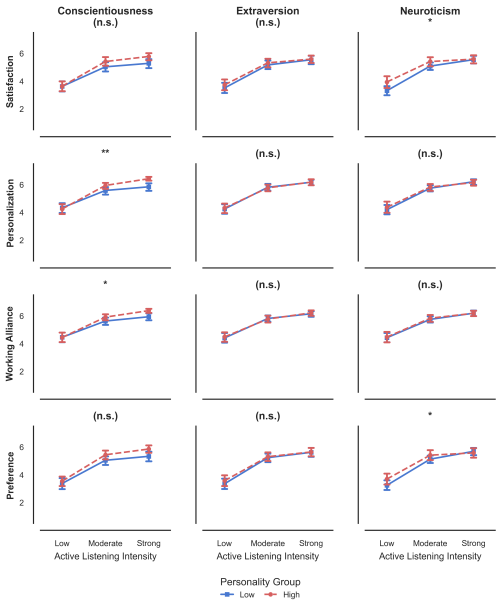

Figure 4: *Active Listening Intensity and Big Five Personality Traits on User Perception and Experience Metrics.*

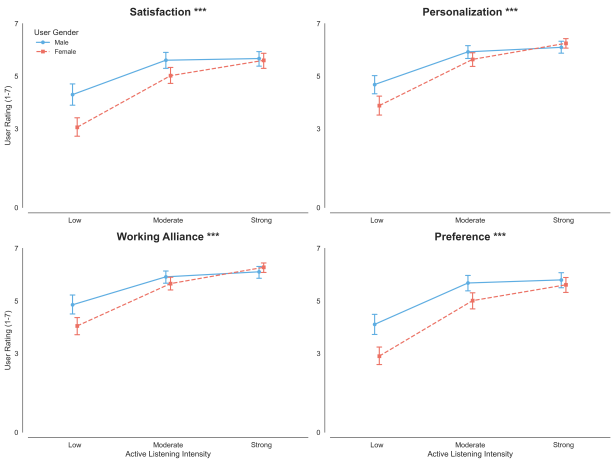

Figure 5: *Gender Differences in Active Listening: A Multi-Metric Interaction Analysis.*

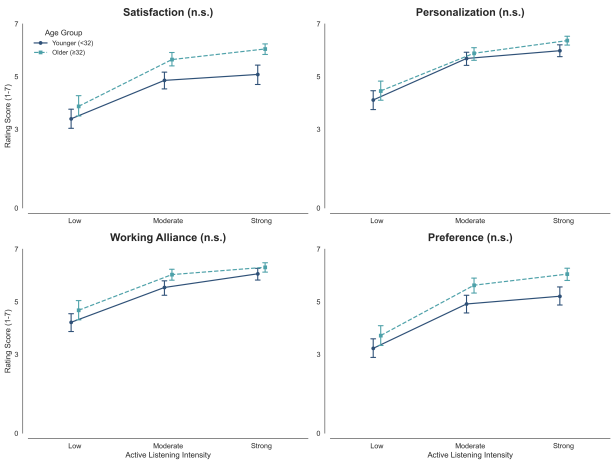

Figure 6: *The Effects of Agent Active Listening Intensity and User Age on System Perception.*