ChatGPT

NOCC-Like Cognitive/Noncognitive Assessment via LLM Conversations – Research Evidence and Design

Deliverable 1: Evidence Map of Key Studies

Domain	Reference (Authors, Year)	Key Findings / Conclusions	Measures or Constructs	Sample (N, Population)	Effect Sizes / Reliability	Limitati
Psychometrics (Trait Estimation)	**Matheson (2019)** – PeerJ	Emphasizes the importance of reliability in trait measures; reliability is the proportion of variance not due to error [1]. Low reliability (<0.8) attenuates correlations and requires larger samples [2] [3]. Repeated measures or multiple raters can substantially improve reliability (Spearman-Brown: 1 rater ICC=0.50 → 4 raters ≈0.80) [4]. Also highlights need for test-retest studies in validation.	General psychological traits (e.g. intelligence, personality)	N/A (Methodological analysis)	Recommends test-retest r ≥0.80 for stable traits [5]; example: ICC=0.95 for self-esteem one-week apart [6]. Averaging 5 measurements (~sessions) can raise reliability from ~0.50 to ~0.83 [4].	Concept no new c Focuses reliabilit address s LLM cont

Domain	Reference (Authors, Year)	Key Findings / Conclusions	Measures or Constructs	Sample (N, Population)	Effect Sizes / Reliability	Limitatio
Psychometrics (Validity & EMA)	**Kim et al. (2023)** – J. Med. Internet Res. (EMA study)	Demonstrated that aggregating repeated momentary assessments improves reliability and correlates with trait-level outcomes. Even ~5-7 brief EMA prompts yielded acceptable reliability in emotional clarity [7] [8]. The averaged EMA-based "drift rate" of negative affect correlated with life satisfaction and lower depression (r≈–0.27) [9]. Repeated measures help separate stable trait variance from state fluctuations [10] [11].	Emotional clarity (indirect RT measure) via EMA; also depression, anxiety scales	N=196 adults (2-week EMA, 5–6 surveys/day) [12]	EMA-based emotional clarity indicator had test-retest reliability ~0.7 with only ~4 prompts [7]. Correlations: drift rate vs. depression r=–0.27 [13]. G-theory shows resilience scales had G=0.90 across 3 occasions (trait) [14].	Specific t emotion sample (patients) generaliz Prelimin only [15].

Domain	Reference (Authors, Year)	Key Findings / Conclusions	Measures or Constructs	Sample (N, Population)	Effect Sizes / Reliability	Limitati
Chatbot-Administered Scales	**Dosovitsky et al. (2021)** – Front. Digit. Health	Established that standard mental health questionnaires can be delivered via chatbot with high feasibility and internal consistency. In a large sample, 99.8% completed the PHQ-9 via a text-based chatbot [16]. The chatbot PHQ-9 showed excellent internal consistency (Cronbach's α = 0.896) [17] and a one-factor structure with good fit [18], mirroring paper-and-pencil results. Demographic effects on scores were minimal (significant but negligible) [19]. Supports validity of chatbot-administered assessments.	PHQ-9 (Depression scale) via chatbot interface	N=3,902 adults (US/Canada, including >75 years) [16]	α = 0.896 [17]; Factor CFA: CFI=0.93, RMSEA=0.08 [18]. Completion rate 99.82% [16].	Only exa PHQ-9; s selectior Faceboo comparis person administ this stud

Domain	Reference (Authors, Year)	Key Findings / Conclusions	Measures or Constructs	Sample (N, Population)	Effect Sizes / Reliability	Limitatio
Noncognitive Trait (Grit)	**Credé et al. (2017)** – J. Pers. Soc. Psychol. (Meta-analysis)	Found that “grit” (perseverance and passion for long-term goals) has much weaker effects than initially claimed. Meta-analysis showed grit is not a unique construct: its perseverance facet correlates moderately with performance, but overlaps heavily with Conscientiousness; the “consistency of interest” facet had minimal predictive validity [20] [21]. Overall grit had only small to medium correlation with academic or job success (~r = .18). Suggests caution in interpreting grit scores; incremental validity over traditional traits is limited.	Grit scales (Grit-O, Grit-S); outcomes: GPA, performance metrics	88 studies, total N ≈ 66,000 (students, employees)	ρ (true-score correlation) grit–performance ≈0.18; Perseverance facet ~0.20, Consistency facet ~0.06 [20] [21]. Heterogeneity $I^2 > 70\%$.	Publicati possible. self-repo outcome construc concern not be di Conscien

NLP Inference of Personality	**Fan et al. (2023)** – J. Appl. Psychol.	Evaluated an AI chatbot interview that inferred Big Five personality traits from text responses. Machine-derived personality scores showed: (a) acceptable internal consistency and test-retest reliability at domain and facet level; (b) a factor structure comparable to self-report inventories; (c) **Convergent validity**: machine scores correlated with self-report Big Five ($r \approx .48$ on average) [22] ; (d) **Discriminant validity**: weaker, machine trait scores were intercorrelated (mean $r \approx .35$) [23] ; (e) **Criterion validity**: low – machine scores had weak correlations with GPA and peer-rated adjustment. In some cases, machine-inferred scores had incremental validity beyond self-report [24] . The psychometric properties were stable across subsamples, indicating generalizability.	Big Five personality (machine-inferred from 20–30 min chatbot interview) vs. IPIP-300 questionnaire	N=1,444 undergraduates (USA). Subsample n=407 with GPA & peer ratings; n=61 retested. [25] [26]	Convergent: $r \approx 0.45–0.55$ per trait [22] . Discriminant: machine Big5 inter-trait $r \approx 0.30–0.40$. Test-retest ICC over ~1 week ~0.70s (est.). Machine–GPA: small (ΔR^2 ~ .02).	Only Eng speaking students ended in content r in other c Criterion limited. I overfit to data patt biases).

Domain	Reference (Authors, Year)	Key Findings / Conclusions	Measures or Constructs	Sample (N, Population)	Effect Sizes / Reliability	Limitati
NLP Inference of Mental Health	**Fisher et al. (2025)** – Res Sq. (preprint, Systematic Review)	Meta-analysis of 50 studies using NLP/ML to detect depression from text (social media posts, interviews, transcripts). **Overall diagnostic performance was moderate**: pooled accuracy ~0.80, AUC ~0.79 [27] [28]. Precision ~0.78, recall ~0.76. Considerable heterogeneity: models using structured clinical interview transcripts and non-English texts performed best [29]. Text source was a significant moderator (e.g., clinical interview data > social media). The authors stress that while language-based depression detection is promising, **standardization and validation are needed before clinical use** [30] [31] (to avoid overfitting and to ensure generalizability).	Depression identification via language (various NLP features; outcome: clinical diagnosis or symptom scale)	50 independent samples (k=43 for accuracy; total N≈40,983) [27]	Pooled Accuracy = 0.80 (95% CI 0.76–0.83); AUC = 0.79 (0.70–0.85) [27] [28]. Higher accuracy in interviews (up to ~0.85) vs. social media (~0.75).	Many stu cross-sec potentia publicati (though as minim Preprint peer-revi Models n perform well on r world da bias).

Domain	Reference (Authors, Year)	Key Findings / Conclusions	Measures or Constructs	Sample (N, Population)	Effect Sizes / Reliability	Limitatio
Structured Conversation & Interviews	**Huffcutt et al. (2014)** – Personnel Psychology (meta-analysis)	High **structure** in interviews yields significantly better reliability and validity than unstructured formats. Asking all participants the same standardized questions and using **behaviorally anchored rating scales (BARS)** for scoring increases inter-rater agreement and criterion validity [32] [33] . Structured behavioral interviews (with BARS) predict job performance with correlations ~0.30–0.50 (vs ~0.20 for unstructured) [33] [34] . They also show less bias and "faking" than self-report surveys [35] . This demonstrates the effectiveness of a scripted-yet-flexible conversational protocol for assessment.	Structured employment interview (situational & behavioral questions); performance ratings	k=86 studies (job interview validity meta)	Validity (structured): r ~0.43 with job performance; unstructured ~0.20–0.30 [33] . Inter-rater reliability often >0.80 with structured panels. BARS use associated with higher predictive validity [36] .	Focused outcome translate psycholo indirectly extensive developm anchor e

Domain	Reference (Authors, Year)	Key Findings / Conclusions	Measures or Constructs	Sample (N, Population)	Effect Sizes / Reliability	Limitati
Motivational Interviewing (MI) for Assessment	**Miller & Rollnick (2013)** – MI 3rd Ed. (and MIA-STEP manual)	MI is an evidence-based conversational style that can be adapted for assessment. Key MI principles (express empathy, elicit "change talk," support self-efficacy) create a safe, engaging atmosphere [37] [38] . An "MI assessment sandwich" protocol has been used: start with open-ended MI dialogue to build rapport and understand client perspectives, then insert a structured assessment (standard questionnaires or rating tasks), and end with MI-based discussion of the results and motivation [39] [40] . This approach yields richer qualitative data and maintains client engagement, though it requires skillful facilitation. MI techniques have been shown to increase clients' honest disclosure and strengthen self-efficacy beliefs during sessions [37] [41] .	Motivational Interviewing (clinical interview) techniques applied to structured assessment (e.g., "change plan" discussion)	N/A (Clinical frameworks and training studies)	MI meta-analyses show moderate effects on behavior change (d~0.2–0.4) due to enhanced client language. No direct numeric outcome for assessment quality, but MI-consistent interviewer behavior correlates with client change talk and honesty [42] [43] .	Not a me tool per s requires maintain [44] . Har quantify psychom the conv output. F time-inte min sess

Ethics & AI in Mental Health	**Padilla (2026)** – CA Senate Press Release (SB-903)	Highlights emerging consensus that **AI chatbots must not be presented as licensed therapists or diagnostic tools**. Cites clinicians' warnings that AI "therapy" bots pose privacy risks, may misinterpret users (lacking human empathy and nonverbal cue detection), can give incorrect or harmful advice, and foster undue reliance [45]. Legislation (SB-903 in CA, 2026) is proposed to prohibit offering or advertising mental health services as "AI therapy" without a human professional accountable [46]. Also requires any AI assistance in care to have clear disclosure and affirmative informed consent from users, plus strong confidentiality protections [47] [48]. These guidelines underscore the need for transparency, user consent, and strict limits on the role of LLMs (only supportive, not	N/A (Ethics/ Governance policy)	N/A (Regulatory context)	–	Not emp policy-fo Details sp Californi not apply universa 2026, ma provisior and unte (awaiting enforcem

Domain	Reference (Authors, Year)	Key Findings / Conclusions	Measures or Constructs	Sample (N, Population)	Effect Sizes / Reliability	Limitatio
		replacing clinicians).				

Data Privacy & User Consent (AI Tools)	**Utah HB 452 (2025)** – Mental Health Chatbot Law	This state law requires: **Clear disclosure** that a mental health chatbot is AI and not a human therapist at the start of any conversation [49] ; **Data protections** – providers may not share or sell users' identifiable mental health data or chat content to third parties without consent [50] . Also bans targeted advertising within chatbot interactions [51] . Enforcement includes fines per violation. This exemplifies "data minimization" and privacy-by-design: collect only necessary information and ensure confidentiality. Aligns with ethical calls that AI systems in healthcare obtain explicit informed consent, explaining how data will be used and stored [52] [53] . Users should have the right to delete their data [54] . Overall, governance is moving toward stricter oversight to prevent misuse of sensitive	N/A (Law/ Governance)	N/A	–	Jurisdict specific (new – no evaluatio Focused chatbots therapy; cover ed settings

Domain	Reference (Authors, Year)	Key Findings / Conclusions	Measures or Constructs	Sample (N, Population)	Effect Sizes / Reliability	Limitatio
		conversational data.				

Sources: See inline citations 【3】【5】【13】【19】【21】【28】【33】【35】【37】【42】【46】【49】【50】 above for detailed references. The evidence map includes peer-reviewed papers and policy documents supporting the design of an LLM-based provisional assessment tool for cognitive and noncognitive skills.

Deliverable 2: Protocol Design Guidelines (5×30min Chat Sessions)

Overall Structure: We propose a **semi-structured interview protocol** spanning **5 sessions (≈30 minutes each)**. Each session has a specific focus (covering both trait **constructs** and current **state** factors) but follows a consistent MI-inspired style to build rapport and encourage authentic responses. The repetition across 5 sessions helps average out momentary fluctuations, increasing the reliability of trait estimates [4] while also allowing us to observe within-person changes (state sensitivity).

- **Session 1 – Establishing Rapport & Baseline:** (Focus: **Introduction and Self-Perception**) Begin with **Motivational Interviewing (MI)** core techniques – open-ended questions, affirmations, reflective listening [39] [38]. For example:
- "Tell me about a recent accomplishment you're proud of." (assesses self-efficacy and strengths)
- "What challenges have you been facing in learning or work?" (assesses stressors, baseline motivation)

Purpose: Build trust, let user articulate their goals/values. Collect baseline indicators of **self-efficacy**, **grit** (through examples of persistence), and **well-being**. Use MI to **support self-efficacy** ("confidence") by highlighting past successes [55] [38]. State checks: current mood, any acute stress today. Mini-task: A brief free-writing or storytelling prompt (to capture language style for NLP analysis of personality/well-being).

- **Session 2 – Cognitive Reflection:** (Focus: **Cognitive Abilities & Problem Solving**) Present a few short, engaging cognitive exercises through dialogue (recognizing the limits without full psychometric testing):
- E.g. a simple **fluid reasoning puzzle** presented in text, or a riddle, to see problem-solving approach (qualitatively).
- A **working memory** game using the chat (remembering details from earlier in conversation to recount later).

Purpose: Though we can't get precise IQ or processing speed via chat, we observe reasoning process, ability to handle information, and frustration tolerance. Trait focus: capturing reflective thinking, verbal reasoning (as proxy for fluid reasoning). State focus: note if user is tired, distracted (affecting performance). - If a user struggles, use **CBT-style probing**: "What went through your mind as you tried that problem?" to identify thought patterns (e.g. self-talk indicating confidence or doubt). This session's dialogue provides qualitative "evidence" for any cognitive ability rating, with caveats about precision.

- **Session 3 – Motivation & Grit:** (Focus: **Noncognitive – Motivation, Perseverance, Values**) Apply a **values clarification exercise (ACT-inspired)**: "What long-term goals or values are most

important to you right now?" Have the user describe why those matter. Then explore **grit** with MI-consistent probes:
- "Can you share a time you almost gave up on a project, but didn't? What kept you going?"
- "What typically motivates you to push through difficulties? What sometimes gets in the way?"

Purpose: Elicit narratives that reveal the **"perseverance of effort"** aspect of grit and the consistency of interests. Also assess **intrinsic/extrinsic motivation**, goal orientation, and energy levels. Use **reflective listening** to paraphrase their motivations ("It sounds like having a clear purpose really fuels you."). This session doubles as an intervention to bolster motivation (echoing MI's evocation of "change talk" related to persistence) [56] [37]. State element: current energy and engagement – e.g., ask them to rate their current motivation on a 1–10 scale (MI "importance/confidence ruler") and discuss why not lower/higher [41].

- **Session 4 – Social and Emotional Factors:** (Focus: **Well-being, Social Support, Stress Reactivity**) Begin with a brief **well-being check-in** each time (e.g., "How have you been feeling since we last spoke?" – capturing fluctuations in mood or stress [57]). In Session 4, delve deeper:
- "Who do you turn to for support when things get stressful?" (assesses **social support** networks, which is a context variable for performance and resilience)
- "Can you describe a recent stressful event and how you handled it?" (measures **coping strategies**, emotional regulation).

Use elements of **CBT** here: if they describe negative thoughts, gently challenge distortions ("What is an alternative perspective on that situation?"), which also gauges cognitive flexibility. Incorporate **ACT** if relevant: discuss acceptance vs. avoidance of difficult feelings. - Trait focus: **Emotional stability** (does the user ruminate or cope adaptively?), **social connectedness** (an indirect noncognitive trait). - State focus: Check if their stress this week is higher or lower than usual – an EMA-like query to correlate with state measures of anxiety or mood. The **evidence** from this session informs provisional ratings of well-being, coping/self-regulation skills, and whether current stress is affecting their engagement (important for validity of other trait measures).

- **Session 5 – Future Plans and Wrap-Up:** (Focus: **Goal-Directed Behavior, Summative Self-Evaluation**) In the final session, adopt a **strengths-based, structured summary**:
- Ask the user to reflect: "Over our conversations, what have you learned about how you approach challenges?" This meta-question prompts them to summarize their own traits (often aligning with Big Five or grit language, which can be compared to our inference).
- Present a **brief scenario** to gauge **achievement motivation** and **competitiveness** (if relevant to NOCC): e.g., "Imagine you have a chance to take on a project that is very competitive. How would you feel and what would you do?" (Look for signs of achievement striving vs. anxiety).
- **Attendance/self-discipline:** "What strategies do you use to keep yourself showing up to class/ work even when you don't feel like it?" (This addresses self-regulation and conscientiousness in context of attendance/effort).

Purpose: This session consolidates trait impressions (by explicitly discussing them) and checks understanding. It also measures **metacognition** and **insight** — how accurately do they see their own strengths and weaknesses? - Use an **MI approach to reinforce positives**: "It sounds like you've identified that once you commit to something, you follow through, which not everyone does – that's a real asset." (Reinforcing self-efficacy and grit) [37] [58]. - Finally, do a **well-being "exit" question**: "On a scale of 0–10, how would you rate your overall well-being this past month?" and "What might help improve it by one point?" This not only gives a self-rated outcome to compare with our inferred well-being score, but also leaves the user with a proactive mindset (an MI-consistent closing).

Trait vs. State Separation: We explicitly design the protocol to capture both **stable traits** and **state fluctuations**: - Trait indicators are gleaned from patterns and content across all sessions: e.g., if the user consistently describes overcoming obstacles (high grit) or frequently shows planning and organization (high conscientiousness). By aggregating evidence from 5 interactions, we improve the stability of these estimates [4]. A **Generalizability Theory** approach could be used post-hoc to ensure that variance attributable to the individual (across sessions) is high for trait constructs [10] [11]. - State indicators (like daily stress, mood) are checked in each session (short "how are you today" or situation-specific prompts). We expect these to **vary session-to-session**, and we will use those variations to validate that our system can detect state changes (e.g., if the user reports a major stressor in Session 4, the "stress" provisional score should reflect that). Over 5 sessions, we can **smooth out state noise** for trait scores (e.g., by taking an average or median of the trait-related responses), while also reporting on state trends.

Standardization vs. Personalization: The protocol is **semi-structured**: each session has set **core questions/tasks** (ensuring every user gets comparable prompts, aiding content validity [34]), but the **LLM is instructed to allow free-form responses and ask appropriate follow-up questions**. This is akin to a structured clinical interview which uses predefined questions but follows the interviewee's lead for probes [59]. By having the same topics for all, we achieve some standardization (critical for fairness and comparability [33]), yet the conversational nature ensures the user feels heard (enhancing engagement and honesty). The **five-session structure** also helps reduce social desirability or faking – as consistency of responses is harder to maintain if insincere across multiple interactions, and structured interviews are known to reduce faking compared to one-time surveys [35].

Use of Evidence-Based Techniques: We weave in **MI** (to engage and reduce defensiveness), **CBT** strategies (to explore thought patterns), and **ACT** (to clarify values) not as therapies, but as **structured conversational tools**. This approach is supported by evidence that these frameworks can increase the richness and accuracy of self-disclosure: - MI's empathetic, non-judgmental stance is likely to make users more comfortable talking about weaknesses or negative feelings (important for validity of self-assessments) [38] [37]. - For example, if a user expresses a self-doubt, the LLM (following MI) might reply, "It's understandable you feel that way; would you be open to exploring where that belief comes from?" – this encourages deeper reflection rather than a surface answer. - **Behaviorally Anchored Questions:** Many questions prompt concrete examples ("Describe a situation when…") similar to behavioral interview techniques, yielding observable "behaviors" in narrative form that we can anchor our ratings to [34] [60]. This is aligned with evidence from structured interviews that such questions improve reliability and predictive validity by focusing on actual past behavior [61].

Each session is thus **carefully scripted** (with flexibility) to serve a dual purpose: **data collection for measurement** and **maintaining a therapeutic alliance-like rapport**. The output of each session will be a transcript rich with "evidence"—stories, statements, response patterns—that our scoring rubric (Deliverable 3) can utilize. By Session 5, we expect to have a multi-faceted view of the person's cognitive and noncognitive profile, cross-checked against their own perceptions and consistent across multiple days.

Deliverable 3: Scoring and Aggregation Plan

To convert the qualitative conversation data into quantitative "provisional scores," we design a **Behaviorally Anchored Rating Scale (BARS) system** [62] [36] for each target construct, along with robust aggregation and validation rules:

1. 0–100 Behaviorally Anchored Rating Scales (BARS): For each construct (e.g., grit, fluid reasoning, self-efficacy, well-being), we develop a BARS with descriptive anchors at key points (e.g., 0, 25, 50, 75, 100): - **Anchor definitions:** Each anchor is phrased in terms of concrete behaviors or responses observed in the chat. For example, a **Grit** scale might say: - 0 = "Gives up immediately when faced with minor setbacks; explicitly states low determination." - 50 = "Persists at times but also switches tasks or interests when frustrated; gave at least one example of overcoming a challenge after an initial failure." - 100 = "Demonstrates exceptional perseverance; recounts multiple instances of sustained effort over long periods despite significant obstacles; language indicates unwavering commitment to goals." - These behavioral examples come from literature and our data: e.g., high grit anchor derived from user saying "I kept studying every day for 6 months for that exam even after failing twice". Using BARS is evidence-backed to improve rater agreement and predictive validity [36] . It reduces subjective guesswork by tying scores to specific evidence from the chat.

Each construct's BARS will be refined in pilot testing (using sample transcripts collected, possibly via crowdsourcing, similar to how Kell et al. (2017) gathered incidents to build BARS [63] [64]). Our aim is that any trained rater (human or AI) using the BARS on the same transcript segment would give similar scores – i.e., high inter-rater reliability [65] [33] .

2. Evidence Attribution ("Evidence Rules"): Every score assigned must be justified by at least one direct quote or interaction excerpt from the conversation. We implement an **"evidence tagging"** system: - As the LLM (or a subsequent analysis process) evaluates the conversation, it will **tag specific user statements** that correspond to BARS criteria. For instance, if rating **self-efficacy**, the system might tag the user's quote "I figured I could handle it, I've done something similar before" as evidence of high confidence (supporting a higher score) [38] [58] . - The final report for each session will include these snippets next to the provisional scores, ensuring transparency. For example: "Grit = 75 (High): evidenced by user statement 'I set a goal to write 500 words daily and haven't missed a day in 4 months'【hypothetical】." This mirrors the practice in structured interviews where raters note behavioral examples justifying their ratings [66] . - **No evidence, no score:** If a construct wasn't evidenced in a given session, no score is given that session (or it's marked as "insufficient data"), preventing unsupported inference. This rule enforces content validity – we only measure what was actually manifested [34] . - We will maintain an **"audit trail"** of these evidence-to-score linkages, which is critical for building user trust and for any human oversight to review how the AI arrived at a conclusion [67] [52] . This also mitigates risks of "hallucinated" judgments – the model must point to user's words as rationale.

3. Session-wise Scoring Procedure: After each session, the LLM (or a secondary analysis model) produces provisional scores for that session's data: - Use a **multidimensional rubric**: e.g., Session 3 (Motivation/Grit focus) will yield scores for Grit, Self-Efficacy, and perhaps Goal Orientation; Session 4 (Stress/Support focus) yields scores for Well-Being, Coping Skills, etc. We design a form for each session listing constructs expected and space for scores + evidence. - **Scaling 0–100:** Even though anchors are defined at 0, 25, 50, 75, 100, intermediate values can be given if evidence falls between anchor levels (just like a human interviewer might rate 3 out of 5 on a Likert anchored scale if partially meeting criteria). The 0–100 scale (in 5-point increments effectively) gives nuance. Research in assessment suggests finer scales can work if anchors are well-defined [68] [62] , but we will monitor that the model isn't over-precision. (We could round to nearest 5 or 10 to keep it coarse and reliable). - **Confidence flags:** The system can also note if a score is tentative (e.g., if evidence is weak or contradictory). This could be a simple confidence indicator (high/medium/low) next to the score, guiding how we weight it in the final aggregation.

4. Aggregating Over 5 Sessions: To derive a **final provisional score** per construct (the ultimate output akin to the NOCC profile), we need to combine session scores in a way that maximizes reliability and validity: - **Simple Average (Mean):** The baseline approach is to take the mean of all available session scores for a construct, under the assumption that each session is an independent sample of the true trait plus random error. Averaging improves reliability by canceling out measurement error [4] . For instance, if "Well-Being" was scored 60, 70, 50, 65, 55 across sessions, the average ~60 gives a more stable estimate.

- **Weighted Average:** Not all sessions are equally informative for all constructs. We might weight sessions more when the construct was a primary focus. E.g., a grit score from Session 3 (which directly targeted grit) might be given 2× weight compared to a minor mention in Session 1. We will define weights in the protocol (e.g., Session 3: grit weight 0.4 of final, other sessions combined 0.6). This is somewhat analogous to test blueprints where certain sections count more. - **Bayesian Updating:** Another approach is to treat the initial expectation of a score as a prior (maybe from self-report if available, or population average) and update with each session's evidence. However, given the small number of sessions, a simpler empirical Bayes – essentially weighted averaging with consideration of variance – might be enough. For transparency, we likely stick to averaging unless pilot data indicates a strong case for Bayesian modeling (which could unnecessarily complicate explainability). - **Outlier handling:** If one session score is an outlier (e.g., user was having a very bad day that drastically affected that session's scores), we might consider a median or truncated mean. This decision will be guided by validation – if including that session reduces test-retest reliability or criterion correlation, we'll adjust the aggregation method accordingly.

Example: Suppose "Self-Efficacy" was rated [Session1: 70, S2: 60, S3: 80, S4: 75, S5: 78] with each backed by quotes. The provisional overall **Self-Efficacy = ~73** (average). If Session 2 was an outlier due to a temporary slump the user had, the median is 75 – we might report 73 with a note on fluctuation, or even report a range. During validation (Deliverable 4) we'll see which aggregate correlates best with the criterion (e.g., a standard self-efficacy scale administered separately).

Ultimately, the aggregated score also comes with the **evidence portfolio** – we plan to show the user (and any human reviewer) a summary of why that score: e.g., "Your provisional grit score is 80/100, based on multiple examples you shared of sustained effort (like continuing your research project over a year despite setbacks) and consistent passion for your goals." This kind of feedback can enhance the user's understanding and acceptance of the assessment.

5. Rater Agreement and Consistency Checks: We incorporate several layers to ensure the scoring is consistent and not just one model's whim: - **LLM self-consistency:** We may prompt the LLM to rate the same session multiple times with slightly varied prompts or using different chain-of-thought methods, and see if scores converge. If there's high variance, that flags low confidence. (Similar to how ensemble methods work for stability). - **Multiple Models:** Use at least two different LLMs (or one LLM and one human) to score a sample of sessions. For example, if our main model is GPT-based, we could also have a smaller language model or rule-based system do a pass. Agreement can be measured via **intraclass correlation (ICC)** or Cohen's κ [69] [70]. If ICC is low (<0.6), we refine the BARS definitions and model prompts. - **Human Oversight:** Initially, during development, a psychologist or domain expert will independently score some transcripts using the same BARS. We expect fairly high alignment because of the explicit anchors. Structured interviews in research have shown inter-rater correlations often 0.8+ when using defined scoring criteria [65] [33]. Any systematic discrepancies (e.g., the LLM overestimates social anxiety compared to human ratings) will be analyzed and prompt-tuning or rubric changes made. - **Continuous Calibration:** Even after deployment, we could periodically sample conversations (with user permission) for human review to make sure the LLM's scoring hasn't drifted. This also ties into governance (Deliverable 5) – ensuring no "unchecked" AI assessment without some form of auditing, as recommended by emerging regulations.

6. Addressing Weak Indicators (e.g., Processing Speed): Some constructs, like **processing speed**, are hard to measure via text chat (since we are not using response latency due to user typing variability and potential system delays). If a construct is deemed not reliably measurable in our medium: - We either **exclude it from the scoring** (mark it as "unable to assess via chat – for true processing speed, a separate test may be needed") or - Use a **proxy indicator with caution**. For example, processing speed might be proxied by how quickly or succinctly the user responds to cognitive prompts or how many ideas they generate in a timed challenge. But we would tag such a score with **low confidence**. In interpretation guidelines, we state that this score is a rough guess and not on par with a real cognitive test. - An alternative approach: measure something related like **behavioral speed** in responding to questions (e.g., number of words typed in a minute on a simple task), but this enters the realm of reaction time which we said we mostly avoid for fairness (different typing speeds, etc.). - We might instead focus on **verbal fluency** (how easily do they express ideas) as an observable trait from chat, which correlates with certain cognitive abilities. - Any **construct where validity is weak in chat** will either be (a) supplemented with a **user-reported estimate** ("How fast do you usually finish tests or tasks compared to others?" – self-assessment) or (b) explicitly labeled in the report as "Not directly assessed: to be interpreted with caution."

7. Dynamic Adjustment and Learning: Our scoring system can improve over time: - If, during validation, we find that certain linguistic features predict a score well, we can incorporate those into the rubric (for example, frequent use of first-person singular and negative emotion words might inform a higher depression risk score [71] – we'd use this carefully along with the conversational content). - We'll also keep an eye on **bias** – e.g., if our LLM tends to score certain demographics differently given the same inputs (which we would test with simulated data), we must adjust. The BARS approach, focusing on concrete behaviors, inherently helps reduce subjective bias [65] [33], but the model's training data bias must be considered. We could employ bias mitigation prompts or even have a bias-checking script that reviews the evidence for irrelevant content (e.g., if the user's English fluency is low, does the model unfairly judge them as less intelligent? We would guard against that by focusing on content of answers, not language complexity per se).

In summary, the scoring and aggregation plan combines **structured expert-derived scales** (BARS), **transparent evidence linking**, and **multi-session integration** to yield provisional scores. Each score is traceable to the conversation (increasing trust and accountability), and our methods of averaging and rater consistency align with classical psychometric approaches to enhance reliability [4] and validity. By explicitly acknowledging uncertainty (through confidence levels and leaving some constructs unscored if needed), we avoid overstating what the LLM assessment can do. This framework will be further refined through pilot testing and the validation studies described next.

Deliverable 4: Validation Minimum Plan

To ensure our LLM-based provisional assessments are credible and "in alignment" with existing measures, we propose a pragmatic validation plan focusing on **convergent validity**, **reliability**, **sensitivity**, and **known-group validity**, executed with minimal burden:

1. Convergent and Criterion Validation against NOCC: Since the goal is to approximate the existing NOCC text report (which presumably includes standard questionnaires or teacher ratings of noncognitive skills), we will collect parallel data: - For an initial sample of users (with consent), administer the **original NOCC battery or equivalent standard scales** in a traditional format (e.g., Big Five questionnaire, Grit-S scale, self-efficacy scale, well-being survey). This can be done either before the first chat session or after the 5 sessions (to avoid priming the conversation). - Compute correlation coefficients (Pearson's r) between the LLM-derived scores and the corresponding survey scores. We expect strong positive correlations if our system is valid (e.g., LLM grit vs. Grit Scale self-report, LLM well-being vs. Satisfaction with Life Scale, etc.). A successful convergent validity might be $r \geq 0.5$–0.6 [22], akin to what Fan et al. found between chatbot-inferred and self-report personality ($r \sim .48$ on average) [23]. - We will also check **criterion-related validity** if available: for instance, if academic performance or attendance records are accessible (and ethically usable), do the LLM's "academic self-efficacy" or "conscientiousness" scores predict those outcomes similarly to NOCC reports? Even a modest correlation ($r \sim .3$) could indicate our conversational measure has real-world relevance [24] . - We need to be cautious: If the LLM was trained on common language patterns, high correlation might partly reflect shared method (language-based) variance. We might complement self-reports with **observer ratings**: e.g., have a teacher or peer provide a rating of the person's grit or social skills, and see if our LLM score correlates (this would be a more stringent test of criterion validity).

2. Test-Retest Reliability (Stability): Since we have 5 sessions, we can examine reliability in two ways: - **Internal Consistency across sessions:** If we treat each session's measurement of a trait as an item, we can compute something analogous to Cronbach's α or intra-class correlation. But given only 5 "items" (sessions), α may be unstable. Alternatively, use **ICC(1, k)** from a two-way mixed effects model (person as random effect, session as repeated measure) to see how much of the variance in scores is person vs. error. A high ICC (close to ≥0.7–0.8) would indicate our multi-session aggregation is reliably distinguishing people [5] [72] . If ICC is low, we need to refine consistency (maybe some sessions were too context-specific). - **Session 1 vs Session 5 correlation:** Since we hope the average of 5 is stable, we can also see how well the first session's provisional scores predict the final session's scores for the same individual. If our premise that averaging improves reliability holds, then any single session by itself might be less reliable – but by session 5 we also expect some **learning or engagement effect**. Ideally, if the trait is truly stable and measured well, Session1–Session5 correlation might be reasonably high (≥0.5–0.6 for traits like Big Five, which is analogous to a two-week test-retest of a personality inventory [73] [6]). For states like mood, we don't expect test-retest stability (that's good, mood should change). - If possible, have a subset of users **repeat the entire 5-session sequence** after a gap (say one month later) and see how well the overall scores replicate (this is a full test-retest of the assessment as a whole). This might be resource-intensive, so we might do it on a smaller sample. The correlation between Time1 composite scores and Time2 composite scores would indicate the temporal stability for traits (target > 0.7 for something like grit or personality over a month, since true change in a month is minimal). - We will use these reliability analyses to possibly adjust scoring: e.g., if one session consistently seems off (reducing reliability), we might drop or alter that session in future protocol revisions.

3. State Sensitivity (Construct Validity for State measures): We intentionally include state measures (stress, mood, etc.). To validate these: - During the 5-session period, we can introduce a **known stressor** or track a naturally occurring event. For example, ask users if they had any significant event (exams, deadline, personal issue) between sessions. Or even experimentally, one could schedule Session 4 right after a stressful task (with ethical limits). - Check if the LLM's state scores respond accordingly. If a user reports a spike in stress, does their "stress reaction" score increase, and their well-being score dip, compared to other sessions? This would demonstrate **sensitivity to change**, akin to responsiveness in psychometrics [74] [75] . - We can correlate LLM-generated state scores with external momentary measures: e.g., have users fill a quick mood rating (PANAS or stress VAS) before each session. Those should correlate with our conversation-inferred mood tone or stress level. Possibly, as seen in some studies, language sentiment can reflect depression/anxiety levels [76] [71] . If our system is picking up language cues, we expect positive mood words, faster responses, etc., to align with higher well-being scores. - We also examine if day-to-day variability in user self-report is captured. A successful result: "On days the user says they are very stressed (self-rating), our LLM's provisional stress/state-anxiety score is high" (concurrent validity for state). - We will document any failure to capture state changes (for instance, if the user was actually very upset but the system missed it because the topic didn't come up). That can inform us to adjust the protocol (maybe add a direct question "How are you today really?" in each session to ensure it gets captured).

4. Known-Groups Validity: Where feasible, test whether our assessment can differentiate between groups that it theoretically should: - If we have access to some grouping variable – e.g., "high performers" vs "at-risk" students, or people who have high attendance vs. those with low attendance (if this info is available from school records) – we can see if our scores reflect those differences. For example, do students with high attendance have higher “conscientiousness/self-discipline” scores from the LLM than those with chronic absenteeism? That would mirror known-group validity (similar to how traditional tests show differences between, say, honor students and dropouts on noncognitive scales). - Another example: perhaps compare a group that we expect to have **low grit** (maybe individuals who voluntarily report they often abandon projects) vs. a group with demonstrably **high grit** (e.g., marathon runners or long-term hobbyists). If our system is valid, it should assign significantly different grit scores to these groups, without being told about their background explicitly. It would be extracting that from how they talk about challenges. - Conduct t-tests or effect size (Cohen’s d) comparisons between groups. We’d want to see moderate to large differences for clear contrasts. For instance, known conscientious students vs. not, d maybe > 0.5 on the “self-management” score. - If such known-group differences are not observed in our scores but are known to exist via other measures, that indicates a potential validity problem (maybe our questions weren’t tapping the construct strongly enough or the scoring was too lenient).

5. Incremental Validity and Convergent Patterns: While not strictly “minimum,” we can explore if our conversation-based scores add any insight beyond traditional measures: - For example, does a combination of language style features and interaction markers yield any prediction of outcomes beyond the self-report scale? (Fan et al. found some incremental validity of chatbot scores beyond questionnaires in predicting certain criteria [24] .) - We can do regression analyses where an outcome (like GPA or teacher rating) is predicted by both the survey score and the LLM score to see if the LLM contributes uniquely. However, given our aim is not to replace but to approximate NOCC, this is secondary. - Another convergent pattern: Check the inter-correlations of our LLM trait scores. Are they in line with theory? E.g., grit should correlate positively with conscientiousness, modestly with self-efficacy; personality dimensions like extraversion and positive affect should correlate. If we find bizarre correlations (like our “openness” equivalent score correlates 0.9 with “neuroticism” equivalent), something’s off in our method (discriminant validity issue) [23] .

6. User Feedback and Face Validity: In validation, also gather qualitative feedback from participants: - Did they feel the conversation covered the topics well and that the provisional feedback resonated with them? If a user says “the chat really described me accurately” (or specifically, “I agree I am not very persistent, as the system noted”), that’s informal evidence of face validity and acceptance. - Conversely, if users consistently say a particular score or interpretation “doesn’t make sense,” we need to investigate. Either the scoring was wrong or we failed to communicate results with the right caveats. - We can include a brief survey at the end asking “On a scale of 1–5, do you feel the assessment described you accurately?” and an open comment. High agreement would be nice, but even critical feedback helps refine prompts.

Minimal Cost Considerations: We intend to leverage data that can be collected in-app and through existing instruments: - Using self-report scales as the primary comparison is cost-effective (no lab tests needed). Many such scales are free or low cost (Big Five IPIP, Grit-S scale (8-item) is free, etc.). - We will piggyback on any routine assessments if this is in an educational setting – e.g., if students already take a noncognitive survey, use that rather than administering new ones. - The sample size: We don’t necessarily need a huge N to detect moderate correlations (N~50–100 could detect r ~0.3–0.4 with power ~0.8). But to be safe and to check subgroups, aiming for ~N=100–150 for initial validation would be good. - If budget allows, oversample some known groups as mentioned to strengthen the known-groups analysis (even if it’s small N in each, large differences can be evident). - Much analysis (correlations, t-tests) can be done automatically once data is collected, so it’s more about coordination than expense.

Success Criteria: We will consider the validation successful if: - Most LLM scores show **convergent r ≥ 0.4–0.5** with corresponding traditional measures (comparable to other text-based inference studies, where r in .4–.6 range is common) [22] . - **Test-retest ICC ≥ 0.7** for stable traits across sessions. - State measures vary appropriately and correlate with external state indicators (qualitative confirmation or moderate r with daily mood). - No major biases or inconsistencies are found (or if found, they can be fixed). - Users report understanding and accepting the results (at least qualitatively).

If certain constructs don't meet these criteria, we will either recalibrate (adjust questions or scoring) and re-run a smaller validation, or consider dropping that construct from the "provisional" report. The aim is to be transparent about what is and isn't yet validated. For example, if "processing speed" proxy scores show near-zero correlation with an actual digit-symbol coding test, we will note that "Processing speed could not be assessed via chat with confidence," aligning with our earlier caution.

Finally, this validation is an **ongoing process**. Post initial validation, we move to more continuous validation: as more users go through the system, we can anonymously analyze correlation patterns and outliers, refining the model. This iterative approach ensures the provisional measurements inch closer to the psychometric standards of traditional tests [77] [78] while leveraging the rich, dynamic data of conversation.

Deliverable 5: Ethics and Safety Considerations

Implementing an LLM-based psychological assessment requires stringent ethical standards and governance to protect users. Our approach includes measures for **informed consent, privacy, data minimization, transparency about limitations, and misuse prevention**:

1. Non-diagnostic, Non-therapeutic Clarity: We explicitly frame the tool as a **"provisional self-assessment"** – **not** a clinical diagnosis, **not** therapy or counseling. All user-facing descriptions and conversations will reinforce this: - The AI assistant will introduce itself clearly as an AI and not a licensed professional [79] [49]. For example: "Hi, I'm an AI conversation partner designed to help you reflect on certain skills and feelings. I'm not a therapist or a doctor, but I can help summarize some of your strengths and challenges based on our chat." - We avoid language like "diagnose," "treat," or any claim of medical or definitive evaluation. If a user directly asks for a diagnosis (e.g. "Do I have ADHD?"), the assistant will respond with a disclaimer: "I cannot diagnose conditions. A licensed professional is needed for that. We're just exploring some of your tendencies." - This distinction is not only ethical but legally required in some jurisdictions (e.g., California SB-903 plans to prohibit calling such AI services "therapy" without a human provider [46]). Our compliance will meet those standards: e.g., "This tool is for educational and personal insight purposes only."

2. Informed Consent Process: Before starting the first session, users (or their guardians, if minors as applicable) must review and agree to a concise consent form. Key points in the **consent document (or welcome message)**: - **Purpose:** Explain that the chat sessions aim to estimate certain personal attributes (like problem-solving style, motivation) in an experimental/provisional manner [80] [81]. - **Procedures:** Outline that it involves 5 chat conversations of ~30 minutes each, and that they will be asked to discuss personal experiences and feelings. - **Risks & Benefits:** State that reflecting on challenges may cause mild emotional discomfort, but can also provide insight. Emphasize it is not therapy and not a substitute for professional help if needed (include resource info for mental health support). - **Privacy:** Describe how data will be handled (see below), and that no unauthorized person will see their identifiable responses. - **Voluntary Nature:** They can stop at any time or skip questions. Participation is voluntary and not for high-stakes use (e.g., not for a grade or job selection – and if it ever were, that would be a separate consent due to higher risk of coercion). - **Consent confirmation:** The user might have to click "I Agree" or explicitly type a consent acknowledgment. In case of minors, ensure parental consent is obtained as required by applicable laws.

Consent example text: "Before we begin, please read: This AI-guided chat is a research-based tool to help you reflect on personal skills (like how you handle tasks or stress). It is not a clinical evaluation or therapy. Your participation is voluntary, and you may pause or exit at any time. The chat will ask about your experiences; you can skip any topic you're not comfortable with. Your responses will be used to give you a personal report and to improve the tool. They will be kept confidential (see privacy policy). There are no direct benefits or payments, and minimal risks (you might feel a bit emotional reflecting on challenges, but we'll provide resources if you need support). By continuing, you indicate you understand this and agree to participate. Do you consent to proceed?"

This is in line with ethical guidelines that **informed consent must be obtained with clear explanation of AI involvement and its limits** [80] [81] .

3. Data Minimization and Privacy: We adhere to the principle of **collecting the least amount of personally identifiable information (PII) necessary** [53] : - We do **not** ask for full name, address, or other sensitive PII during the chats. If such details are volunteered, the system will not store them in analysis (and ideally the prompt will instruct the model to not follow up on exact PII). - The conversation data, which by nature might include personal anecdotes, will be stored securely and in de-identified form for analysis. De-identification might involve replacing names or specific places with general tags (e.g., "[HOMETOWN]") in the stored logs. - Only derived **feature data** (scores, linguistic features) may be kept long-term for research, with the raw transcripts either deleted or heavily anonymized after the scoring is done. This addresses data minimization – we keep what's needed for validation and improvement, not the entire identifiable chat longer than necessary. - **Encryption & Access Control:** All data in transit and at rest will be encrypted. Access to raw data will be limited to essential team members under confidentiality agreements. If using cloud LLM services, we'll ensure they have proper data handling policies (or use on-device models to avoid external data flow). - The user will be given the option to **delete their data** after receiving their report if they wish. For instance, a UI button "Delete my conversation data" which will remove their logs from our system (this aligns with emerging AI regulations like Utah's requiring right to deletion [54]). - Privacy policy will clearly state we **do not sell or share** user data to third parties (especially not sensitive mental health info) [50] . The only exception: if user is part of a study with their institution's oversight, even then, data would be aggregated and anonymized. But as a rule, no third-party marketing or unrelated use – which is also mandated by new laws (e.g., Utah HB 452) [50] . - If any data is used for research publications, it will be anonymized quotes at most (with consent) or aggregate statistics; no individual should be identifiable.

4. Misuse Prevention (Scope of Use): We must guard against the tool being used inappropriately, such as in high-stakes decisions or by unauthorized parties: - We will include in our terms of service a **prohibition for use in hiring, academic admissions, employee evaluation, or any decision that significantly affects an individual's rights or opportunities**. The tool is intended for personal development and research, not as a gatekeeping test. This aligns with fairness – these LLM-based scores are provisional and not validated for such purposes. - If an organization expresses interest in using the tool for assessment of candidates or students for selection, we would refuse or require a separate extensive validation and ethical review. At this stage, we explicitly label it "not for selection purposes." In user language: "These results should not be used as a sole basis for any hiring or academic decisions." - **Third-party data access:** We won't allow external parties (like employers or schools) to access an individual's results without explicit consent from the individual for that specific sharing. Even with consent, we'd warn the individual of potential implications. Ideally, any third-party use would involve de-identified aggregate data (e.g., a school gets a report "class average trends" but not who scored what, unless part of a guided intervention program with consent). - **Advertising and conflicts of interest:** The chatbot will not advertise products/services during the session (and our design is to be non-commercial in the chat content) [82] . This avoids any conflict of interest or perceived exploitation of the user's disclosures. If the project is funded by some entity, we will disclose that in info sheets. - We also consider implementing a safeguard that if someone tries to use the tool on someone else's behalf (for example, forcing someone to take it, or a manager asking an employee's results), the design makes that difficult. Since it's a conversation, the person has to actively engage. We rely on consent at the start, and a feature that the account is user-specific.

5. Handling of Acute Risks: While the tool isn't therapy, users might reveal serious issues (e.g., suicidal thoughts, abuse) in conversation. We must have a **crisis management protocol**: - The AI is instructed to recognize keywords or sentiments indicating acute distress or risk (like references to self-harm, hopelessness, abuse). If detected, it should **not continue the normal assessment**. Instead, it will respond with empathy and encourage seeking professional help, and provide resources. For example: "I'm really sorry you're feeling that way. It sounds like you might be going through something very serious. It's important to reach out to a mental health professional or someone you trust. I can provide you with some hotline numbers or resources if you'd like." Then possibly pause the assessment. - We will include in consent that "This is not monitored in real-time by humans, so if you are in crisis, please contact [Hotline]…" thereby also making users aware not to rely on it for emergencies. - If our system has any human oversight, perhaps a flag could alert a human moderator if someone seems at immediate risk (this must be done carefully respecting privacy – perhaps only if user explicitly asks for help or gives clear suicidal intent, and even then we'd need prior consent that we can intervene). This is tricky ethically because of privacy vs. duty to rescue. We might at least ensure the AI strongly urges the user to get help and provides the information. - **Minors protection:** If minors use the system (say high school students with parent consent), we might need protocols if they mention abuse or other situations where mandated reporting could apply if a human were present. Since it's AI, no actual mandated reporter is in the loop to automatically report. Our approach could be to immediately encourage the minor to talk to a safe adult or counselor. We likely state in the minor assent/parent consent that "the AI cannot fulfill mandated reporter duties, so if you reveal something like that, it will encourage you to seek help but cannot itself notify authorities." We might actually restrict certain age groups or ensure a counselor is co-monitoring in a school setting if that's a concern.

6. Transparency and User Understanding: We provide **plain-language explanations** of the tool's capabilities and limits: - At the end of the sessions, when giving the "provisional report," include a section on **"Explanation and Limitations."** For example: "These scores are estimates based on an AI analysis of our conversations. They may not capture every aspect of you, and some factors could have influenced them (like mood or how comfortable you felt). The results are not a medical or definitive assessment, just a reflection for you to consider. You might find some parts accurate and others not – feel free to disagree or discuss them with a mentor/counselor." - Basically a **limitations template** as requested: - "This chat-based assessment cannot measure certain abilities (e.g., processing speed) with high accuracy and does not account for all context. The AI might miss nuances that a human professional would catch. Results should be taken as preliminary and are subject to verification." - "Your personal situation and environment can significantly affect these skills – this is not a judgment of your worth or potential, just a snapshot of certain tendencies." - "If you receive a low score in an area, remember this is not fixed – and it might even be an underestimate. Conversely, a high score is not a guarantee of success, but indicates strength in that area during our chats." - We ensure the user can ask questions to the AI about their report. The AI can then clarify, using the evidence, why it gave a score, or acknowledge when it isn't sure (transparency). E.g., User: "Why do you think my self-efficacy is 60?" AI: "That's a great question. I based it on things you said like 'I often doubt I can do things' in Session 1【hypothetical】. However, you also gave examples of successes. 60 means moderate confidence. This could be different in other situations – it's just what I gathered here." - We will provide contacts or references if they want to learn more or even opt to take a formal assessment with a psychologist for comparison.

7. Governance and Accountability: - Internally, we will maintain documentation of the system's development process, data handling procedures, and an ethical risk assessment. If an IRB or ethics committee is overseeing the research, we'll adhere to that protocol. - We plan to regularly review the system's outputs for any **bias** or unintended harmful suggestions. For example, ensure it's not systematically giving lower scores to a certain demographic or using stereotypes in conversation. This could be part of validation and ongoing QA, possibly involving external auditors or stakeholder feedback (like input from psychologists, educators, and participants). - **User control:** At any point, a user can withdraw. If they do, we stop the sessions and give the option to have data deleted. If a user only completes 3 of 5 sessions and stops, we won't try to "complete the assessment" without them – partial data will either be offered with heavy caution or not at all. - **AI model updates:** If we change the LLM or scoring algorithm, we version the system and ideally re-do at least a subset of validation to ensure consistency. Users should be notified of any major changes that could affect their results if using the system longitudinally. - **Liability:** While we do our utmost to prevent harm, we will have a liability disclaimer (common for such tools) while still taking responsibility seriously. E.g., "The creators of this tool have made efforts to ensure it is helpful and safe. However, we cannot guarantee the accuracy or suitability of the results for any purpose. By using the tool, you acknowledge these results are for personal reflection only." This is legally prudent given unknown unknowns.

In essence, our ethical framework aligns with emerging guidelines like the APA's recommendations for AI in practice (e.g., disclose AI involvement, obtain consent, ensure privacy) [81] [83] and with laws like Utah's which demand disclosure, privacy and an absence of deceptive anthropomorphism in mental health chatbots [49]. By embedding these principles, we aim to protect users from harm, earn their trust, and ensure that the tool is used only in appropriate, constructive ways. Each user's autonomy, dignity, and well-being remain at the center of our design – the technology is a means to facilitate self-discovery, not a determiner of fate.

---

[1] [2] [3] [77] [78] We need to talk about reliability: making better use of test-retest studies for study design and interpretation - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC6536112/

[4] reliability - How is the Spearman-Brown prophecy formula affected by questions of differing difficulties? - Cross Validated
https://stats.stackexchange.com/questions/97085/how-is-the-spearman-brown-prophecy-formula-affected-by-questions-of-differing-di

[5] [6] [57] [69] [70] [72] [73] Reliability and Validity of Measurement – Research Methods in Psychology – 2nd Canadian Edition
https://opentextbc.ca/researchmethods/chapter/reliability-and-validity-of-measurement/

[7] [8] [9] [12] [13] [15] Reliability and Validity of Ecological Momentary Assessment Response Time–Based Measures of Emotional Clarity: Secondary Data Analysis - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC11294766/

[10] [11] [14] Investigating state and trait aspects of resilience using Generalizability theory | Current Psychology | Springer Nature Link
https://link.springer.com/article/10.1007/s12144-023-05072-4

[16] [17] [18] [19] Psychometric Properties of a Chatbot Version of the PHQ-9 With Adults and Older Adults - PubMed
https://pubmed.ncbi.nlm.nih.gov/34713116/

[20] Marcus Crede - Professor of Psychology, Iowa State University
https://scholar.google.com/citations?user=IaVkjxoAAAAJ&hl=en

21 [PDF] What Shall We Do About Grit? A Critical Review
https://dr.lib.iastate.edu/server/api/core/bitstreams/f4a39379-1728-482a-8e3f-9bd7cecba9d8/content

22 23 24 25 26 How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores - PubMed
https://pubmed.ncbi.nlm.nih.gov/36745068/

27 28 29 30 31 Language-Based Detection of Depression with Machine Learning: Systematic Review and Meta- Analysis - PubMed
https://pubmed.ncbi.nlm.nih.gov/41333443/

32 33 34 35 59 60 61 65 66 Structured Interviews
https://www.opm.gov/policy-data-oversight/assessment-and-selection/other-assessment-methods/structured-interviews/

36 63 64 (PDF) Exploring Methods for Developing Behaviorally Anchored Rating Scales for Evaluating Structured Interview P
Methods for Developing BARS
https://www.researchgate.net/publication/
318146611_Exploring_Methods_for_Developing_Behaviorally_Anchored_Rating_Scales_for_Evaluating_Structured_Interview_Performance_Explori

37 38 39 40 41 55 56 58 344260_newintro.pmd
https://motivationalinterviewing.org/sites/default/files/mia-step.pdf

42 43 44 Motivational interviewing quality assurance: A systematic review of assessment tools across research contexts - ScienceDirect
https://www.sciencedirect.com/science/article/abs/pii/S0272735820300970

45 46 47 48 Senator Padilla Introduces Protections from Dangerous AI Therapy Products | California State Senator Steve Padilla
https://sd18.senate.ca.gov/news/senator-padilla-introduces-protections-dangerous-ai-therapy-products

49 50 51 82 Utah Enacts AI Amendments Targeted at Mental Health Chatbots and Generative AI | Healthcare Law Blog
https://www.sheppardhealthlaw.com/2025/05/articles/artificial-intelligence/utah-enacts-ai-amendments-targeted-at-mental-health-chatbots-and-generative-ai/

52 53 54 67 79 7 Essential Guidelines for Building an Ethical AI Chatbot in 2025
https://dialzara.com/blog/7-ethical-guidelines-for-building-trustworthy-ai-chatbots

62 Structured Behavioral Interviews - Human Resources
https://hr.az.gov/structured-behavioral-interviews

68 Behaviorally Anchored Rating Scale: Examples + Guide - AIHR
https://www.aihr.com/blog/behaviorally-anchored-rating-scale/

71 Depression and Anxiety Have Distinct and Overlapping Language ...
https://pmc.ncbi.nlm.nih.gov/articles/PMC10799169/

74 Validity and reliability of Patient-Reported Outcomes Measurement ...
https://pmc.ncbi.nlm.nih.gov/articles/PMC3779528/

75 Known-group validity of patient-reported outcome instruments and ...
https://pmc.ncbi.nlm.nih.gov/articles/PMC5644599/

76 Language sentiment predicts changes in depressive symptoms - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC11441484/

80 Ethical guidance for AI in the professional practice of health service ...
https://www.apa.org/topics/artificial-intelligence-machine-learning/ethical-guidance-ai-professional-practice

81 Artificial intelligence is reshaping how psychologists work
https://www.apaservices.org/practice/news/artificial-intelligence-psychologists-work

83 Ethical considerations for GenAI in mental health care - APA Services
https://www.apaservices.org/practice/ce/expert/ethical-genai-mental-health-care