Can LLMs Reliably Detect Complex Human Emotions such as Curiosity and Anxiety

By

Saachi Gandhi

BS Neuroscience, University of North Carolina, Chapel Hill, 2024

Thesis

Submitted in partial fulfillment of the requirements for the Degree of Master of Science in the Graduate Program of Biotechnology at Brown University

PROVIDENCE, RHODE ISLAND

MAY 2026

This thesis by Saachi Gandhi is accepted in its present form by the Graduate Program in Biotechnology as satisfying the thesis requirements for the degree of Master of Science

Date 4/29/26 Signature: Judson Brewer

Dr. Judson Brewer, Advisor

Date 4/28/26 Signature: _____

Dr. Yong Jong, Reader

Date 4/28/26 Signature: Adam Crego

Dr. Adam Crego, Reader

Approved by the Graduate Council

Date _____ Signature: _____

David P. Lindstrom, Dean of the Graduate School

Acknowledgements

Thank you to everyone in the CAN-DO lab and a big thank you to my mentor Dr. Judson Brewer and my trusty aide, Dr. William Nardi. Also, a huge thank you to Istvan Mezzo from OpenMind AI for helping me with all of my many (many) questions about AI and building a reliable AI chatbot.

Thank you to my committee members, Dr. Adam Crego and Dr. Yong Jong for their time both in and out of the classroom. Your advice and patience will always make Brown a special place of growth for me.

To my mom and brother, thank you for the long nights over the phone discussing how we can be more curious about curiosity. To my supportive partner Lydell, thank you for being my source of light throughout it all.

All of you made this possible, thank you.

Table of Contents

Table of Figures

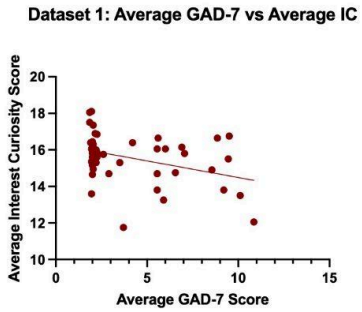

Figure 1a. Baseline data using GPT-4o

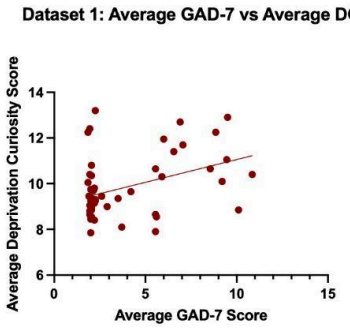

Figure 1b. Baseline data using GPT-4o

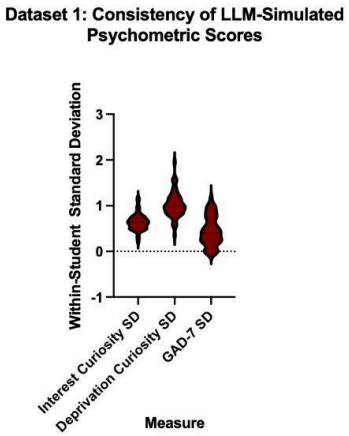

Figure 1c. Baseline data using GPT-4o

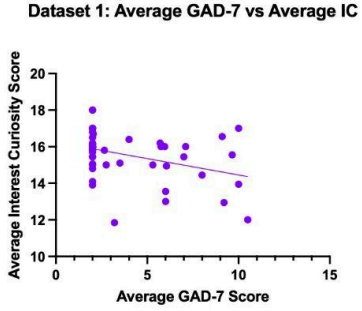

Figure 2a. Baseline data using GPT-4.1

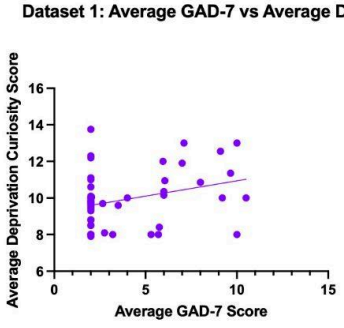

Figure 2b. Baseline data using GPT-4.1

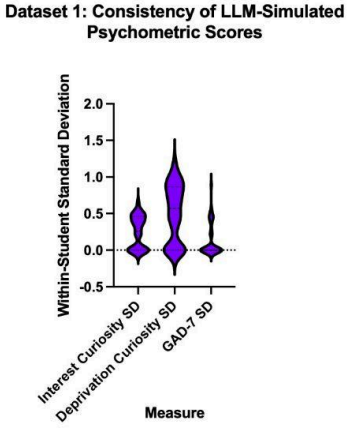

Figure 2c. Baseline data using GPT-4.1

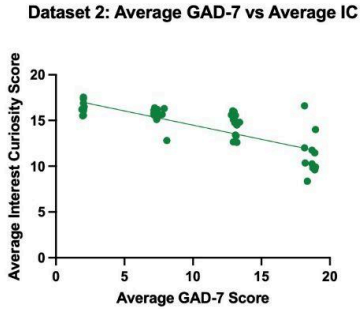

Figure 3a. Baseline + stress data using GPT-4o

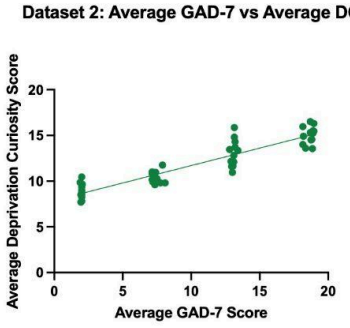

Figure 3b. Baseline + stress data using GPT-4o

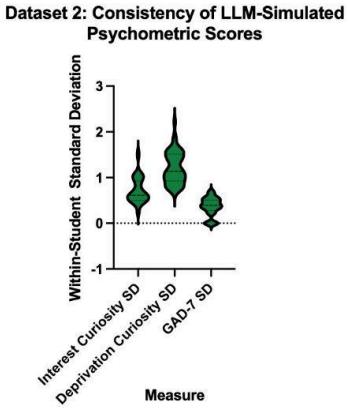

Figure 3c. Baseline + stress data using GPT-4o

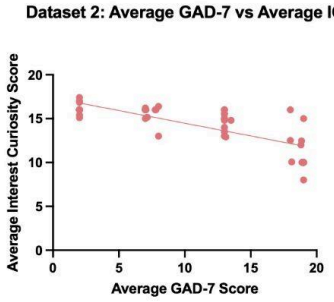

Figure 4a. Baseline + stress data using GPT-4.1

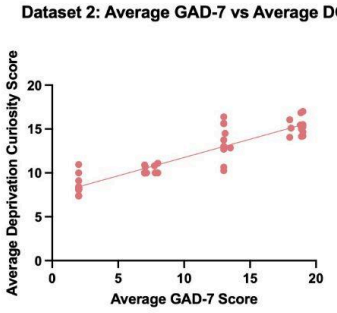

Figure 4b. Baseline + stress data using GPT-4.1

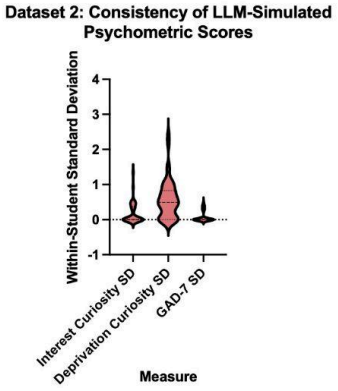

Figure 4c. Baseline + stress data using GPT-4.1

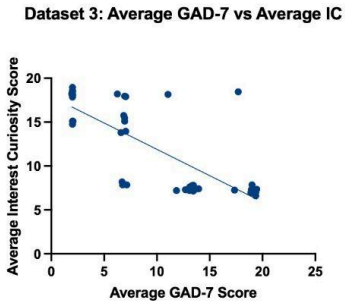

Figure 5a. Baseline + stress + curiosity augmentation data using GPT-4o

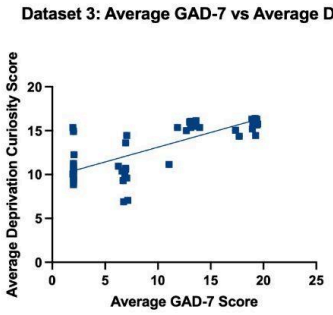

Figure 5b. Baseline + stress + curiosity augmentation data using GPT-4o

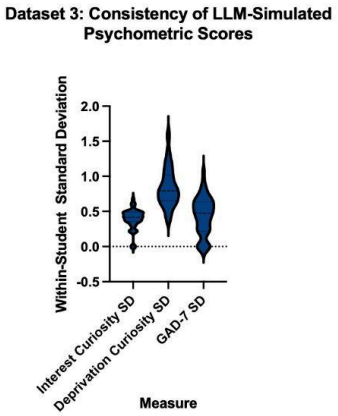

Figure 5c. Baseline + stress + curiosity augmentation data using GPT-4o

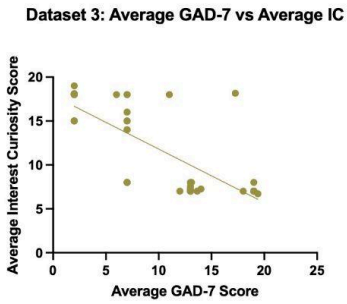

Figure 6a. Baseline + stress + curiosity augmentation data using GPT-4.1

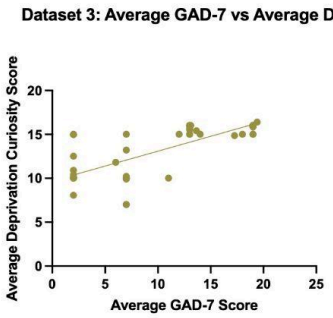

Figure 6b. Baseline + stress + curiosity augmentation data using GPT-4.1

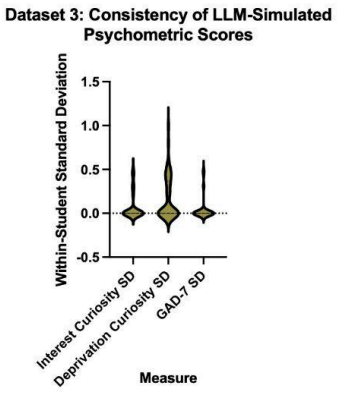

Figure 6c. Baseline + stress + curiosity augmentation data using GPT-4.1

Chapter 1: Introduction

College is a critical transitional period, where students learn how to strike the balance between academic, personal, and financial responsibilities. With the move into adulthood, learning how to adapt and balance daily stressors is critical. However, this phase of life is often also accompanied with several mental health-related problems. In this cognitive tug-of-war, students attempt to harmonize their venture into adulthood and manage their changing mental health. Additionally, as artificial intelligence (AI) systems, particularly large language models (LLMs), become more integrated into daily life, it is important to understand how they interact with human psychology and behavior. This may help inform future mental health assessment and support tools. In this paper, we will 1) examine the intersection of the two main forms of curiosity – interest and deprivation – and their potential correlation with generalized anxiety disorders in college-aged students, 2) understand how we can use artificial intelligence (AI) as a tool to monitor these two primary responses to the unknown and 3) assess the ability of LLMs to consistently detect anxiety and curiosity scores when analyzing complex human emotions and personality traits.

The clinical reality of the mental health burden in college students cannot be understated. According to the 2025 National American College Health Association report, 35.5% of college students in America have been diagnosed with anxiety – including generalized anxiety, social anxiety, panic disorder, or other specific phobias.[1] When further studied, of those diagnosed, 73.6% reported contact with healthcare or a mental health professional within the last 12 months and 46.8% reported anxiety negatively impacting their academic performance.[1] The high incidence of anxiety, likelihood of students seeking help, and resulting impairment in school highlights how standard interventions, while useful, may lag behind the generational advances in

technology that current college students possess. Now, with the rise of AI, students are more likely to seek the counsel of an AI bot before they contact a professional, underlining the need for a tool that can utilize the depths of psychiatric research and provide a means to responsibly answer their questions and measure their cognitive load from home.[2]

The majority of college students already interact with AI tools daily, whether they are streamlining administrative tasks, creating personalized learning plans, or generating ideas.[3] Specifically within psychiatry, AI chatbots can evaluate a user's subjective emotional state, and, by using a wealth of previous literature, they can better identify enduring personality characteristics or particular patterns of behavior reflected in language. We are interested in AI's ability to detect individual levels of the two types of curiosity – interest and deprivation curiosity – and their corresponding levels of generalized anxiety. Interest curiosity (IC) refers to the intrinsic desire to learn new information for the joy of discovery, whereas deprivation curiosity (DC) revolves around the need to resolve knowledge gaps motivated by the unknown or the unresolved.[4]

To understand the potential correlations between levels of generalized anxiety, IC, and DC, we will employ a series of validated tests powered by ChatGPT-4o and ChatGPT-4.1. We are utilizing two ChatGPT agents to measure between-model differences in agentic workflows. Generalized anxiety will be assessed with the GAD-7, and curiosity will be assessed with Litman's Epistemic Curiosity Scale, which distinguishes IC and DC. The chatbot will interact with 50 artificially created students with varying personalities, curiosity types, hobbies, family dynamics, and anxiety levels to calculate their individualized scores. Our use of artificial personas emphasizes a new era of psychology research where we can reduce preventable research errors before involving human participants. Specifically, AI-generated students allow us

to test the structural integrity of the intervention, confirm that survey items are presented as intended, and identify formatting problems, ambiguous phrasing, or implementation errors (e.g., confusing instructions or inconsistent response options) prior to deployment. In this sense, we can “debug” the research design to ensure that it is as safe as possible for real test subjects.

We hypothesize that higher GAD-7 scores will be associated with lower interest curiosity and higher deprivation curiosity, whereas lower GAD-7 scores will show the opposite pattern. Ultimately, the results of this study can help us understand whether LLM-based assessments can reliably distinguish complex human emotions such as curiosity and anxiety and whether interest curiosity may function as a protective factor against generalized anxiety. By navigating the nuances of an AI-powered world, our research takes an important step towards the development of safer, more responsible digital mental health intervention that may define the next generation of healthcare.

Chapter 2: Literature Review

To measure curiosity, researchers typically rely on established questionnaires such as Litman's Epistemic Curiosity Questionnaire, Kashdan's Five Dimensional Curiosity Scale (5DCR), the Curiosity and Exploration Inventory-II (CEI-II), as well as behavioral tasks that assess information seeking and related traits (e.g., joyous exploration, deprivation sensitivity).[4,5,6] The question of how people explore, learn, and respond to uncertainty or new information is frequently sought after, but rarely completely answered.

This is because curiosity is subjective, and developing a reliable metric to quantify it requires a nuanced approach. Curiosity is not a singular state, but rather an approach to dealing with the unknown given personality traits, emotional-motivational states of interest, and the intrinsic pleasure of learning.[7] With these traits in mind, we decided to use Litman's Epistemic Curiosity Questionnaire to train our chatbot because it integrates early theories in curiosity – such as the CEI-II and Berlyne's Curiosity Theory – with modern psychometric measures targeted at understanding the specific psychological "itch" for knowledge that is especially prevalent in college-aged students.[4,8]

2.1 What is Epistemic Curiosity

Litman (2008) discovered that there are two main forms of epistemic curiosity – interest and deprivation. One can think of interest curiosity as the pleasure associated with discovering new ideas.[9] Think of a time when you started learning about your favorite topic – whether it was math, science, English, botany, paleontology or anything that made you revert to the child-like state of asking "Why?" hundreds of times. This is your interest curiosity – the desire to learn for

the sake of learning. Deprivation curiosity differs in that it emphasizes spending time and effort to acquire a specific answer or solution.[9] For example, if you needed to obtain the recipe to make your mom's world-famous quiche, solve a complicated math problem, understand what happens next in your favorite TV show, or resolve any knowledge gap with a specific answer in mind, then your deprivation curiosity activates.

To look at curiosity holistically, one needs to understand how the two types affect different aspects of our personalities. Litman (2008) stated that interest curiosity is correlated with mastery-oriented learning and stimulating positive affect such as diverse exploration and learning something completely new.[9] On the other hand, deprivation curiosity revolves around the reduction of uncertainty, avoiding failure, goal-oriented exploration, success, and acquiring knowledge from an existing knowledge set.[9] Research conducted by Whitecross and Smithson (2023) found similar results where interest curiosity was positively associated with happiness, problem-solving confidence, open-mindedness, and empathic listening skills and negatively associated with distractibility and indecisiveness.[10] Moreover, the researchers found that deprivation curiosity was only positively associated with empathic listening and indecisiveness.[10]

When taken together, these two motivators for understanding the unknown hang in a delicate balance, where situation-specific environments will require more attention from one than the other. It's important to note, however, that curiosity is not the only response to the unknown.

2.2 A Two-Way Street: What Triggers Anxiety and Curiosity

In regards to our target population, our focus more on the knowledge-based aspect of curiosity is to understand *why* people are seeking information – whether that is exploring for the sake of learning (interest-type) or to fulfill a knowledge gap (deprivation-type) – and what its

correlation is with Generalized Anxiety Disorder (GAD). While these two states appear drastically different, curiosity and anxiety can be conceptualized as two opposing responses to uncertainty within related systems. Anxiety leads to withdrawal motivation, where an individual may remove themselves from a potentially dangerous situation.[11] Alternatively, curiosity leads to the approach motivation, where an individual is drawn towards potentially threatening stimuli, especially when the situation is unpredictable.[11] Despite the presentation of the same stimulus, the subjective experience of the perceiving individual will dictate whether they experience anxiety or curiosity. These two responses form a bipolar homeostatic system to respond to the unknown, however, one side may overpower another when dealing with generalized anxiety disorder.

Similar to curiosity, anxiety can stem from various motivations. We chose to measure generalized anxiety disorders (GADs) due to its high prevalence, status as a "trait" anxiety, and its direct link to academic performance and developmental stressors.[1] Whether it's sourced from school, friends, home life, extracurriculars, future job prospects, etc., generalized anxiety is prevalent in every stage of life – especially in college-aged students.[1] To better understand anxiety as a trait, we employed the GAD-7 to analyze the enduring aspects of anxiety while avoiding participant burnout through a shorter, more concise questionnaire.

Finally, even though humans are more versed in understanding complex emotional states, recent models of LLMs have begun displaying an interpretation of these nuances. This is simply because these agents were trained on data supplemented and programmed by humans. However, LLMs do not understand human emotions, but they can detect verbal cues that direct their behavior.

2.3 How do LLMs Infer Traits From Language: Trait Inference and Stable Outputs

To evaluate whether an LLM can meaningfully detect between-group differences on emotional constructs, we must consider what LLMs can and cannot do. Whether it is solving an organic chemistry question, fixing an error in your code, or planning a travel itinerary, LLMs have shown promise for solving complicated tasks. Even so, it lacks the ability to understand emotion. Accordingly, our research question centers on whether LLMs can parse emotional language and accurately distinguish complex human traits such as generalized anxiety and the two forms of epistemic curiosity.

Trait inference from language is a complicated task and LLM outputs have shown to differ in their responses depending on user input. For example, an LLM programmed with the purpose, “You are a college student struggling with anxiety.” will produce different results than, “You are a first-generation college student struggling with generalized anxiety disorder”. Trait inferences and output stability depend on the depth of the prompt.

To further expand on prompt sensitivity, research conducted by Li et al. (2025) emphasized how using highly specific role-playing prompts (ex: You often feel like an outsider looking in) can significantly improve the consistency of an LLM’s response towards a specific trait.[12] If LLMs are to serve as the interim test subjects in psychology research, they must produce responses that are both reliable and interpretable, and their behavior must remain aligned with the intended trait profile across repeated sessions. In our study, this requirement is especially important for epistemic curiosity because for the chatbot to score curiosity accurately, it must be guided towards a consistent curiosity stance so that its answers reflect stable patterns rather than random variation. Simply put, the bot needs to successfully interpret what it means to

be curious. To stimulate curiosity in our chatbot, we used persona injection, which involves guiding the model to become more curious through prompt engineering.[13]

To understand the importance of prompt engineering, Jiang et al. (2023) created personalized LLMs with assigned personality profiles following the Big Five personality model.[14] Their results demonstrated how LLMs can respond in a manner consistent with their designated personality type, shown when the LLMs took the 44-item Big Five Inventory personality test, completed a story writing task, and employed specific linguistic patterns to match their assigned personality in subsequent conversations.[14] Their research ties back into how persona injection can be used to stabilize LLM outputs and through prompt interpretation, adopt specific behavioral tendencies.

Other studies have used the 5DCR to assess the extent of curiosity exhibited by LLMs by using traits such as information seeking, thrill seeking, and social curiosity.[12] It was found that LLMs have the potential to exhibit curiosity similar to that of humans – an important finding when analyzing if LLMs can interpret complex human emotions.

Overall, research in personality profiling and the ability of LLMs to express human emotions such as curiosity, provides experimental support for the future development of using LLMs as test subjects in psychology research.[12] However, the model used will vary the output due to programming schemas beyond customer-interface prompt engineering.

2.4 Agent Differences and Resulting Behavioral Personalities

You may notice that depending on which LLM you use, the model output varies. Ranging from sycophancy to stoicism, LLMs inherently have different tones and jargon regardless of user input. This is not a programming error, but rather an indication that different LLMs may have

different personalities. Heston and Gillette (2025) tested 4 different LLMs – ChatGPT-3.5, Gemini Advanced, Claude 3 Opus, and Grok-regular – against the Big 5 personality tests to assess personality expression and found significant differences in personality profiles depending on the agent used[11]. For example, ChatGPT-3.5 was classified as an ENTJ (Extraverted, Intuitive, Thinking, and Judging) whereas Claude 3 Opus classified as an INTJ (Introverted, Intuitive, Thinking, and Judging).[15] Gemini Advanced and Grok-Regular leaned towards an INFJ (Introverted, Intuitive, Feeling, and Judging), indicating that a need for formal personality evaluations is necessary for LLM integration into mental health settings.[15]

In light of these behavioral and affective contours, we ran our curiosity chatbot on ChatGPT-4o and ChatGPT-4.1. By deploying our chatbots on the same company backend, we can account for personality differences and focus instead on how our chatbots are interacting with the simulated participants.

2.5 Model Selection

ChatGPT-4o was selected to simulate human-like dialogue, analyze complex linguistic patterns, and synthesize large datasets.[13] Prior research suggests this model exhibits the extent of social curiosity comparable to that of a human, allowing future research to build off of an LLM that operates most similarly to a real test subject.[13] Rather than switching to another company backend, we continued the experiment on ChatGPT-4.1 to reduce confounding variables and gain a clearer picture on prompt sensitivity and model capability. Moreover, ChatGPT-4.1 was selected due to its advanced reasoning capabilities and ability to handle large-scale behavioral data.[13] Across both models, evidence of a relatively strong social curiosity supports their use as potential substitutes for humans during early-stage research trials.[13]

2.6 Applications of AI in Healthcare

Accounting for prompt sensitivity and model-specific behavioral profiles is especially important in psychology research because personality expression can influence therapeutic alliance and patient trust in mental health settings.[15] As LLMs are increasingly used in clinical settings for emotional support, cognitive-behavioral therapy, and diagnosis, assessing personality expression in LLMs is crucial for ensuring clinical safety and therapeutic appropriateness.[15] Additionally, LLMs are becoming more embedded in healthcare delivery, and personality profiling should transition from an exploratory exercise to a formal component of responsible AI governance.[15] In mental health contexts where therapeutic rapport and interpersonal nuance affect clinical outcomes, LLM selection and characterization should be guided not only by technical performance, but also by behavioral alignment with patient needs.[15]

Taken together, our work positions curiosity and generalized anxiety as closely related, uncertainty-driven systems that can be separated into approach versus withdrawal depending on how an individual reacts to the unknown. By centering on knowledge-based curiosity (IC and DC), generalized anxiety disorders in a college-aged population, and validated instruments (Litman's Epistemic Curiosity Questionnaire and the GAD-7), we establish a clear psychological framework for interpreting information-seeking behavior. We then extend that framework to LLM-based agents, where prompt sensitivity and model-specific "personality" expression may meaningfully shape the next generation of psychology research where LLMs can successfully mimic human behavior.

Finally, by comparing ChatGPT-4o and ChatGPT-4.1, we can test whether different models can consistently differentiate complex emotional constructs (IC, DC, and GAD) and produce reliable, trait-congruent responses. Ultimately, this approach clarifies the potential

promise and limits of using LLMs as interim test subjects in psychological research, while emphasizing the importance of behavioral alignment, safety, and fairness as these systems move toward broader use in mental health contexts.

Chapter 3: Methods

3.1 Study Overview

This study used simulated college-aged student personas to evaluate whether an AI agent can extract generalized anxiety and epistemic curiosity scores from narrative prompts and reliably extract any patterns of association between generalized anxiety, interest curiosity (IC), and deprivation curiosity (DC). The workflow consisted of 1.) generating standardized persona narratives, 2.) administering validated self-report instruments through our "curiosity chatbot", and 3.) aggregating and analyzing IC, DC, and GAD-7 scores across our experimental conditions.

3.2 Platforms, Models, and Tools Used

The curiosity chatbot was built using OpenAI's Agent Builder and deployed within a Codex-based workflow. The curiosity bot was implemented using two model backends – ChatGPT-4o and ChatGPT-4.1 – to analyze between-model differences, evaluate scoring discrepancies, and reduce any single-model biases (e.g., processing power, complex logic ability, and instruction following). Temperature was set to 0 to make outputs focused, deterministic, and factual by favoring the most likely words. Simulated students were handwritten first, and then fine-tuned using GPT-5.2 to create diverse participants. Experiment runs and repeated trials were conducted in Langfuse, and statistical analyses and visualizations were performed in GraphPad Prism.

3.3 Simulated Participants

Fifty (N = 50) distinct student personas were created to represent a heterogeneous college-aged population with variation in: personality profiles and learning styles, study habits and academic pressures, hobbies and interests, home-life contexts, and baseline anxiety levels.

The students were fine-tuned using ChatGPT-5.2, using profile descriptors easily picked up by an LLM (e.g., "enjoys learning obscure languages,", "loves the thrill of solving complex puzzles"). These cues were included deliberately to test whether the chatbot's natural language understanding could map the student narratives onto stable, trait-relevant responses when the participants took the questionnaires. This method served as a way to check the chatbot's sensitivity to specific wording.

3.4 Questionnaires Used

Generalized anxiety symptoms were measured using the GAD-7, a 7-item self-report measure of symptom frequency over the past two weeks.[16] After each trial, item responses were summed to compute an overall GAD-7 score. Sample items from this measure include:

1. *"Being so restless that it is hard to sit still"*
2. *"Becoming easily annoyed or irritable"*
3. *"Not being able to stop or control worrying"*

Each question was accompanied by a scale ranging from 0–3 to understand how often the participant experienced/aligned with the symptom with 0 = "Not at all", 1 = "Several days", 2= "More than half the days", and 3 = "Nearly every day"[2]. Scores from the seven items were summed and varied from 0-21. Participants were then labeled as minimal anxiety (score 0–4),

mild anxiety (5–9), moderate anxiety (10–14), and severe anxiety (15–21). To reduce speculation, personas were assigned GAD-7 labels outside of the persona prompts so the bot could focus on scoring the curiosity levels first.

Curiosity was assessed using Litman's Epistemic Curiosity Scale.[4] Two scores were calculated, one for interest curiosity (IC) – defined as curiosity driven by the intrinsic enjoyment of learning and discovery – and deprivation curiosity (DC) – curiosity driven by resolving knowledge gaps and reducing uncertainty. After each trial, item responses were summed to compute an overall IC and DC score. Sample items from this questionnaire include:

1. *"I enjoy learning about subjects that are unfamiliar to me."*
2. *"Difficult conceptual problems can keep me awake all night thinking about solutions."*
3. *"I enjoy discussing abstract concepts"*
4. *"I brood for a long time in an attempt to solve some fundamental problem."*

Each item was accompanied by a scale ranging from 1–4 to understand how the participant would describe themselves and how they generally feel. The ten items were then summed with even questions representing interest curiosity and odd questions representing deprivation curiosity. Higher levels of interest curiosity were reflected by a higher score and the same was applied for deprivation curiosity.

3.5 Experimental Design and Prompt Conditions

A repeated-measures design was used in which the same 50 student identities were represented under three progressively detailed persona prompt conditions. Each dataset contained the same set of 50 students, but with different levels of contextual and symptom-specific detail.

1. Dataset 1: Baseline Personas. Brief persona descriptions including ethnicity, hobbies, and broad target anxiety characterizations.

   Example: You are Omar. Identifies as Middle Eastern / African. You are a natural leader who takes charge in group projects. In your free time, you enjoy attending local EDM concerts. You seek a stable career to support your family. You remain remarkably unfazed even during finals week.

2. Dataset 2: Baseline + Stress Augmentation: Expanded descriptions incorporating academic stress cues, study environment, sleep patterns, responses to pressure, and partial GAD-7 style frequency language aligned with the stress symptoms dataset

   Example: You are Omar. You identify as Middle Eastern / African. You are a natural leader who takes charge in group projects. You seek a stable career to support your family. In your free time, you enjoy attending local EDM concerts. Your routine is steady and you can usually handle deadlines without losing sleep. Over the last two weeks, anxiety symptoms have generally not been a problem for you. If worries come up, it is only for several days and you can let go of those worries; you can relax, sit still, and sleep normally.

3. Dataset 3: Baseline + Stress + Curiosity Augmentation: Further enriched descriptions adding more complete GAD-7 frequency language and clearer IC/DC behavioral patterns informed by Litman's curiosity constructs.[4]

   Example: "You are Omar, a college student. Ethnicity: Middle Eastern / North African. Background: Identifies as Middle Eastern / North African. You are a natural leader who takes charge in group projects. In your free time, you enjoy attending local EDM concerts. You seek a stable career to support your family.

You remain remarkably unfazed even during finals week. Your daily routine is fairly steady and school stress feels manageable most of the time. You often seek out unfamiliar topics for fun, and you love to learn something new. When you get stuck on a problem, you are more likely to move on rather than spending hours trying to find the answer. Over the last two weeks, anxiety symptoms have generally not been a problem for you. If worries come up, it is only for several days and you can let them go; you can relax, sit still, and sleep normally.

3.6 Procedure

Each student completed 20 trials per session, across 3 sessions, for each dataset, using each chatbot backend (ChatGPT-4o and ChatGPT-4.1). In each session, the student interacted with the curiosity bot, completed the GAD-7 and Litman's Epistemic Curiosity Scale, and were shown their IC, DC, and GAD-7 scores for that trial. In this study, reliability refers to scoring consistency across repeated trials, agreement with intended GAD-7 severity categories, and consistency across model backends. For example, Omar from Dataset 1, 2, and 3 interacted with the chatbot 20 times per dataset, totaling 120 interactions with both chatbots. Sessions were implemented as repeated trials in Langfuse to quantify run-to-run variability and model sensitivity. Scores from all trials were exported to Prism for analysis. The same procedure was run with the curiosity agent backed by ChatGPT-4o and repeated with ChatGPT-4.1.

3.7 Analysis Tools for Prompt Sensitivity and Reliability

Analyses in Prism focused on descriptive statistics (means and standard deviations) for IC, DC, and GAD-7 across datasets and model backends. Additionally, association testing was conducted between anxiety and curiosity dimensions (IC vs GAD-7 and DC vs GAD-7) within each dataset. Comparisons of average IC, DC, and GAD-7 scores across the three dataset conditions were used to quantify the effects of persona prompt enrichment (baseline vs baseline + stress-augmented vs baseline + stress + curiosity-augmented). Finally, sensitivity checks were run across ChatGPT-4o vs ChatGPT-4.1 to identify model-dependent shifts in scoring or correlation structure.

3.8 Human Subjects Considerations

This study used AI-generated personas rather than human participants. Accordingly, the procedures were intended to validate implementation details (e.g., survey presentation, wording clarity, and scoring stability) and to test whether the pipeline could distinguish between trait-relevant patterns and student narratives. It does not directly measure human emotional reactivity, distress, dropout, or other human-subject risks.

Chapter 4: Results

4.1 Prompt Sensitivity and Scoring Reliability Across Prompt Sets

To test prompt sensitivity, each of the 50 personas were assessed under three prompt conditions that varied in the amount of narrative detail provided: a brief baseline prompt (Dataset 1), an intermediate prompt containing academic stress descriptors and partial GAD-7 frequency language (Dataset 2; "Goldilocks" prompt), and a maximally detailed prompt with full stress and IC/DC behavioral cues (Dataset 3). Across both model backends (GPT‑4o and GPT‑4.1), Dataset 2 produced the most consistent and reliable scoring behavior.

Adding stress-frequency language in Dataset 2 improved alignment between intended and observed GAD-7 severity categories (GPT‑4.1: 50/50 exact matches; GPT‑4o: 50/50 exact matches), whereas Dataset 1 showed the lowest accuracy (GPT‑4.1: 14/50 exact matches; GPT‑4o: 11/50 exact matches). Exact matches refer to the chatbot's ability to score the participant's GAD-7 score correctly depending on the predetermined anxiety level each student was assigned. Curiosity levels were not assigned to see if there is a correlation between set GAD-7 scores and interest/deprivation curiosity language. Although Dataset 3 preserved broad severity coverage and near-perfect label matching (48/50 exact matches for both backends), the additional text did not further strengthen the anxiety-curiosity relationships relative to Dataset 2. By providing a middle ground of structured information (Dataset 2) the chatbot was best supported to produce reliable scoring, consistent with the possibility that overly long prompts increase the risk of LLMs skimming or underweighing key cues.

4.2 Baseline Dataset (Dataset 1): Weak Anxiety-Curiosity Relationships

In Dataset 1, average trait levels clustered toward low anxiety with moderate curiosity for both agent backends. For GPT-4.1, mean interest curiosity (IC) was 15.55 (SD = 1.31), mean deprivation curiosity (DC) was 9.92 (SD = 1.52), and mean GAD-7 was 3.91 (SD = 2.77; median = 2; range = 2-10.5), with most personas classified as minimal anxiety (34/50). For GPT-4o, mean interest curiosity was 15.59 (SD = 1.31), mean deprivation curiosity was 9.85 (SD = 1.40), and mean GAD-7 was 3.94 (SD = 2.76; median = 2.2; range = 1.85-10.85), and observed GAD-7 labels similarly skewed minimal (34/50). In both models, the baseline condition produced the weakest associations between average curiosity and average anxiety (GPT-4o: $R^2 = 0.15$ for GAD-7 vs IC and $R^2 = 0.15$ for GAD-7 vs DC; GPT-4.1: $R^2 = 0.14$ for GAD-7 vs IC and $R^2 = 0.09$ for GAD-7 vs DC), consistent with the limited symptom-frequency information available in Dataset 1 persona narratives (Figures 1a, 1b, 2a, and 2b).

Within-student variability also differed by measure. For GPT-4.1, DC SD was significantly higher than IC SD and GAD-7 SD, and IC SD was higher than GAD-7 SD (all Tukey-adjusted $p < .05$), with many personas showing zero run-to-run variation on GAD-7 (35/50 SD = 0). GPT-4o showed the same ordering (DC SD > IC SD > GAD-7 SD; all Tukey-adjusted $p \leq .0002$), but with substantially larger SD magnitudes overall, indicating greater prompt sensitivity and score instability under the GPT-4o backend in the baseline condition.

4.3 Curiosity and Stress-Augmented Dataset (Dataset 2): "Goldilocks" Prompt With Strongest Anxiety-Curiosity Structure

Dataset 2 (baseline + stress cues) yielded the most reliable scoring across both backends. For GPT-4.1, mean interest curiosity was 14.33 (SD = 2.38), mean deprivation curiosity was 11.97 (SD = 2.89), and mean GAD-7 was 10.50 (SD = 6.32), with perfectly balanced severity labels (12 minimal, 12 mild, 13 moderate, 13 severe) and 50/50 exact matches to intended targets. GPT-4o produced comparable average levels (IC mean = 14.34, SD = 2.47; DC mean = 11.90, SD = 2.59; GAD-7 mean = 10.51, SD = 6.28) and likewise achieved 50/50 exact matches.

Dataset 2 also produced the strongest anxiety-curiosity relationships across the study. For both backends, the relationship between GAD-7 and deprivation curiosity was particularly strong as seen in Figures 3b and 4b (GPT-4o: $R^2 = 0.85$; GPT-4.1: $R^2 = 0.83$), exceeding the corresponding linearity observed in Dataset 1 and Dataset 3 (Figures 1b, 2b, 5b, and 6b). The GAD-7 vs IC relationship was also strong in Dataset 2 as seen in Figures. 3a and 4a (GPT-4o: $R^2 = 0.63$; GPT-4.1: $R^2 = 0.59$). Together, these patterns indicate that the intermediate prompt structure – which included adding explicit stress and symptom-frequency language without overloading the persona narrative – is the most effective method to extract anxiety and curiosity related traits from persona narratives.

In variability analyses, repeated-measures ANOVAs confirmed significant differences in within-student standard deviations across measures for both backends. Under GPT-4.1, DC SD exceeded both IC SD and GAD-7 SD (both adjusted $p < .0001$), while IC SD did not significantly differ from GAD-7 SD (adjusted $p = 0.0536$), and GAD-7 SD showed minimal spread (39/50 SD = 0) (Figure 4c). Under GPT-4o, all three SDs differed (DC SD > IC SD >

GAD-7 SD; all adjusted $p < .0001$), again reflecting greater overall variability under the GPT-4o backend despite stronger accuracy in label matching (Figure 3c).

4.4 Final Prompt With Full Curiosity and Anxiety Cues Dataset (Dataset 3): Improved Over Baseline but Not Beyond Dataset 2

Dataset 3 (baseline + stress + explicit IC/DC behavioral cues) included more personality and behavior characteristics, clearly signaling which students leaned more towards IC versus DC. As a result, students' IC and DC scores were more spread out, as there were more low vs. high curiosity profiles rather than everyone clustering near the middle. For GPT-4.1, mean interest curiosity was 11.29 (SD = 4.93), mean deprivation curiosity was 13.37 (SD = 2.91), and mean GAD-7 was 10.85 (SD = 6.38). For GPT-4o, the corresponding means were similar (IC mean = 11.37, SD = 4.94; DC mean = 13.42, SD = 2.89; GAD-7 mean = 10.89, SD = 6.44). In both backends, label matching remained high (48/50 exact matches), with mismatches concentrated in only two personas.

Dataset 3 strengthened the anxiety-curiosity relationships relative to the baseline prompt but did not exceed the intermediate Dataset 2 prompt. For GPT-4o, R^2 was 0.61 for GAD-7 vs IC (Figure 5a) and 0.56 for GAD-7 vs DC (Figure 5b); for GPT-4.1, R^2 was 0.62 for GAD-7 vs IC (Figure 6a) and 0.55 for GAD-7 vs DC (Figure 6b). This pattern supports the broader prompt-sensitivity conclusion that increased persona detail does not systemically improve model performance. Beyond a moderate level of structured symptom and context cues, additional narrative detail diluted the diagnostic information needed to produce stable psychometric outputs.

Across both model backends, Dataset 1 produced the weakest relationships between average GAD-7 and average curiosity (ChatGPT-4o: $R^2 = 0.15$ for GAD-7 vs IC and $R^2 = 0.15$ for GAD-7 vs DC; ChatGPT-4.1: $R^2 = 0.14$ for GAD-7 vs IC and $R^2 = 0.09$ for GAD-7 vs DC). Dataset 2 produced the strongest anxiety-curiosity relationship, with deprivation curiosity showing the strongest positive relationship with anxiety in both models (ChatGPT-4o: $R^2 = 0.85$; ChatGPT-4.1: $R^2 = 0.83$), and with interest curiosity showing the strongest negative relationship with anxiety (ChatGPT-4o: $R^2 = 0.63$; ChatGPT-4.1: $R^2 = 0.59$). Dataset 3 maintained stronger associations than baseline but did not exceed Dataset 2 (ChatGPT-4o: $R^2 = 0.61$ for GAD-7 vs IC and $R^2 = 0.56$ for GAD-7 vs DC; ChatGPT-4.1: $R^2 = 0.62$ for GAD-7 vs IC and $R^2 = 0.55$ for GAD-7 vs DC).

Within-person SD analyses again showed measure-dependent variability. Under GPT-4o, DC SD was significantly higher than both IC SD and GAD-7 SD (both adjusted $p < .0001$), while IC SD did not differ from GAD-7 SD (adjusted $p = 0.553$) (Figure 5c). Under GPT-4.1, DC SD remained significantly higher than both IC SD (adjusted $p = 0.0194$) and GAD-7 SD (adjusted $p = 0.0015$), while IC SD did not differ from GAD-7 SD (adjusted $p = 0.4128$).(Figure 6c). Across both backends and all datasets, GAD-7 SD tended to be the most tightly clustered, indicating that anxiety scoring was generally more stable run-to-run than deprivation curiosity scoring.

4.5 Comparison of GPT-4o and GPT-4.1 Backends

Across datasets, the overall pattern of anxiety-curiosity associations was consistent between backends (R^2 values were similar for corresponding plots), indicating that the observed prompt-sensitivity effects were not specific to a single model. However, GPT-4.1 generally

exhibited lower within-student variability (smaller SD magnitudes) than GPT‑4o, suggesting a greater scoring stability across repeated trials. Despite this difference in variability, both models displayed the same central result: the intermediate stress-augmented prompt (Dataset 2) provided the best balance of information density and diagnostic specificity, producing the most reliable scoring and the strongest anxiety-curiosity relationship, particularly for deprivation curiosity.

Chapter 5: Discussion

College is a period defined by uncertainty. Academic pressure, shifting social roles, and the daily stressors that come with becoming an adult are all hurdles students must navigate. To examine whether LLMs can capture this cognitive tug-of-war, we tested the ability of an LLM-based assessment in reliably capturing the two forms of epistemic curiosity – interest and deprivation – and their relationship to generalized anxiety. To do so, we evaluated three levels of persona prompting across two model backends (ChatGPT-4o and ChatGPT-4.1) using a brief baseline prompt (Dataset 1), an intermediate "Goldilocks" prompt with academic stress cues and symptom-frequency language (Dataset 2), and a maximally detailed prompt building upon Dataset 2 with additional IC/DC behavioral cues (Dataset 3). By comparing the mean scores across both models, we were able to establish scoring reliability due to both backends scoring participants similarly.

Additionally, we found that deprivation curiosity showed the strongest relationship with generalized anxiety, as seen in Dataset 2 ($R^2 \approx 0.83$-0.85). Across both backends, Dataset 2 produced the most consistent and reliable scoring patterns, highlighting prompt sensitivity as a central factor in LLM-administered psychometrics and showing that more information is not always better.

5.1 Prompt Sensitivity and the "Goldilocks" Prompt

Prompt content strongly shaped both the reliability and strength of anxiety-curiosity relationships. Unlike humans, LLMs do not have innate, stable personalities. Their "psychology" is an interaction between the model's training data and its prompt. Due to this, baseline prompts

appeared to provide insufficient symptom-frequency language and contextual structure, which corresponded with weak anxiety-curiosity relationships and poorer alignment to intended anxiety categories. Dataset 2 added targeted academic stress cues and symptom-frequency language, creating a prompt that was detailed enough to anchor the model without overwhelming it. Dataset 3 introduced additional curiosity-relevant behaviors and more extensive narrative information. While this condition maintained broad dispersion and near-perfect severity matching, it did not outperform Dataset 2 in correlation strength. This finding suggests that longer prompts may dilute or compete with the cues most relevant to scoring, leading to diminishing returns. Regardless of the model used, our present findings demonstrate that prompt structure should be treated as part of the measurement protocol when LLMs are used to administer psychological scales.

5.2 Interpretation of the Anxiety-Curiosity Pattern

Across both agent backends, deprivation curiosity emerged as the dimension most strongly linked with generalized anxiety. As shown in Dataset 2, interest curiosity had a prominent negative relationship with generalized anxiety while deprivation curiosity had a strong positive relationship. These findings hint at a potential correlation between generalized anxiety symptoms and deprivation curiosity. Our research demonstrates a pattern consistent with the conceptual framing that deprivation curiosity reflects an uncertainty-reduction drive – an “itch” to resolve knowledge gaps – which may overlap with worry-based cognitive loops and intolerance of uncertainty.[11] Interest curiosity, in contrast, reflects the intrinsic joy of learning and exploration and is driven by positive anticipation of new information. This drive can boost dopamine, increase ambiguity tolerance, and help shift from fear to proactive exploration.[11]

When taken together, the results support the value of separating IC and DC when evaluating how people respond to uncertainty and suggest that prompt structure can either obscure or reveal these distinctions in LLM-based scoring pipelines.

As mentioned above, our findings suggest that "more words" is not inherently better for LLM-based psychometric administration and that adding more descriptive detail does not provide a clearer relationship between curiosity and anxiety. However, this finding also reveals a critical discrepancy when using LLMs to mimic human behavior: even though a "middle-ground" of information provides a more linear relationship between two fundamental human characteristics, it does not encompass all of the nuances of what it means to be human.

5.3 Reliability and Within-Student Variability

Across datasets and backends, within-student variability differed by measure, with deprivation curiosity consistently showing the greatest run-to-run variability and GAD-7 generally showing the tightest spread. Repeated-measures ANOVAs confirmed significant differences in within-student SD across IC, DC, and GAD-7, and Tukey-adjusted comparisons consistently showed DC SD exceeding the other measures. Across conditions, ChatGPT-4.1 exhibited lower within-student SD magnitudes than ChatGPT-4o, suggesting that the ChatGPT-4.1 workflow was more stable across repeated scoring runs. In addition, GAD-7 SDs tended to be less spread out overall, consistent with anxiety items being more constrained and symptom-frequency based than curiosity items. It's important to note that curiosity items may be more sensitive to narrative interpretation and stochastic sampling. Importantly, despite these differences in variability, both models displayed the same broader prompt-sensitivity pattern and emphasized the association between deprivation curiosity and anxiety.

5.4 Methodological Implications: What AI-Generated Participants Can (and Cannot) Do

These findings underline how synthetic participants can be used responsibly in psychological research. Artificially generated participants can help evaluate whether commonly used scales (such as the GAD‑7 or Litman's Epistemic Curiosity Scale) generalize across diverse profiles, given the longstanding representational gaps in psychological research. Synthetic conversations can also support more ethical AI development and by engaging models with a wider range of identities, backgrounds, and personality profiles, researchers can better surface biases in systems such as Gemini, ChatGPT, and Claude, aiding mitigation efforts and reducing stereotyping in future LLM deployments.

However, at the same time, AI-generated participants cannot capture core human-subject risks such as emotional reactivity to sensitive questions, social desirability effects, dropout, or retraumatization. Subsequent studies must evaluate these risks with appropriate safeguards. For this reason, synthetic participants are best positioned as a piloting tool to "debug" research structure and measurement reliability, rather than as a replacement for human-subject validation.

5.5 Limitations

Several limitations should be considered. First, the use of synthetic personas limits ecological validity and does not allow direct inference about real student experiences. Second, R^2 values summarize association strength but do not establish causality or directionality and may vary with analytic choices. Third, LLM outputs may depend on model configuration and sampling, and results may not generalize to other models beyond ChatGPT‑4o and ChatGPT‑4.1.

Finally, “target range” matching reflects persona construction and may not mirror real world distributions of anxiety in college populations.

5.6 Future Directions

Future work should extend this pipeline to human participants to test whether the observed IC/DC and anxiety patterns replicate under real responding and to evaluate human-subject risks that synthetic trials cannot model. These studies should preregister prompt templates and analysis plans, quantify test-retest reliability explicitly, and conduct bias audits by varying demographic cues while keeping symptom frequency content constant. As LLMs become more integrated into mental health contexts, prompt design and model selection should be treated as formal components of responsible measurement and governance.

Chapter 6: Conclusion

College is a critical transitional period, and students spend much of it learning how to balance academic, personal, and financial responsibilities while adapting to daily stressors that can meaningfully shape mental health. As GenAI – particularly large language models (LLMs) – become increasingly embedded in students' lives, their intersection with psychology raises the question of whether LLM-based tools can responsibly detect and distinguish complex, human traits. In this thesis, we examined 1) the relationship between interest curiosity (IC), deprivation curiosity (DC), and generalized anxiety disorder (GAD) in artificial college-aged students, 2) whether LLMs can be used as a monitoring tool for these two primary responses to the unknown, and 3) the extent to which LLM-based assessments can reliably score these traits across repeated sessions and different model backends.

Our findings highlighted a central theme: prompt design plays a critical role in human-bot interfaces, and adding more detail did not improve scoring performance beyond Dataset 2. We tested our hypotheses using validated instruments (GAD-7 and Litman's Epistemic Curiosity Scale), 50 AI-generated students, and repeated administrations across three prompt conditions. Dataset 2 (baseline + academic stress augmentation) served as a "Goldilocks" prompt across both ChatGPT-4o and ChatGPT-4.1, producing the most consistent scoring and the strongest anxiety-curiosity relationship because it represented the best-performing prompt condition in this study. Baseline prompts (Dataset 1) yielded the weakest relationships and the lowest accuracy in matching intended anxiety categories, reflecting how narratives with a baseline level of detail often fail to convey the symptom-frequency and contextual cues needed for stable scoring. Finally, although the final prompt condition (Dataset 3) contained the most thorough persona narrative – rich with stress-frequency language and clear IC/DC behavioral

patterns – it did not improve the strength of the anxiety-curiosity relationships beyond Dataset 2. This pattern reinforces a practical conclusion about prompt sensitivity: more words are not always better, and excessive narrative detail can dilute or compete with the cues most relevant to psychometric scoring.

Across conditions and backends, deprivation curiosity emerged as the curiosity dimension most strongly associated with generalized anxiety, especially in Dataset 2. This finding is consistent with the conceptual framing presented in our literature review: deprivation curiosity reflects an "itch" to reduce uncertainty and resolve knowledge gaps. It conveys a motivational urge that may overlap with worry-based cognitive patterns. Interest curiosity, in contrast, reflects the intrinsic pleasure of learning and exploration. It increases ambiguity tolerance which helps shift from fear to proactive exploration. Together, these findings emphasize why separating IC and DC is critical when studying how students approach uncertainty. They suggest that LLM-based assessments surface these distinctions, but only when the prompt provides sufficient, but not excessive, diagnostic structure.

At the same time, this work clarifies the promise and limits of AI-generated participants in psychology research. Synthetic personas offer a powerful way to "debug" study design before involving human participants by testing survey flow, administration consistency, scoring stability, and prompt sensitivity. They also provide a method for evaluating whether widely used scales (such as the GAD-7 and Litman's Epistemic Curiosity Scale) behave consistently across a wider range of identities and life contexts than is often represented in traditional research samples. In addition, synthetic conversations can support more ethical AI development by exposing how models may respond differently to demographic and contextual cues, helping

researchers surface bias in systems such as Gemini, ChatGPT, and Claude. These findings may be used to inform mitigation efforts that reduce stereotyping in future deployments.

However, artificial participants cannot substitute for human-subject research when the goal is to evaluate real-world clinical safety. AI-generated participants cannot capture core human risks such as emotional reactivity to sensitive questions, social desirability effects, dropout, or retraumatization. These outcomes must be examined in subsequent studies with appropriate safeguards. Future work should therefore extend this pipeline to human participants, preregister prompt templates and analysis plans, quantify test-retest reliability, and conduct systematic bias audits by varying demographic cues while holding symptom-frequency content constant.

As AI and LLMs become more integrated into students' help-seeking behavior, building tools that are reliable, interpretable, fair, and clinically responsible is essential. Our research takes a step in that direction by showing that LLM-based psychometric scoring can be coherent and meaningful, but only when the narrative/prompt has been engineered with the same care as the psychological instruments it seeks to administer.

Overall, this study indicates that LLM-based psychometric scoring is highly sensitive to how personas are represented and that an intermediate prompt structure (Dataset 2) provides the most reliable extraction of anxiety-curiosity relationships across both ChatGPT-4o and ChatGPT-4.1. While synthetic participants can support generalizability checks and bias surfacing, they cannot substitute for human-subject testing of emotional and ethical risks. Finally, our research shows that LLM-based psychometric scoring is highly sensitive to prompt structure, and that an intermediate prompt condition produced the most reliable scoring pattern across both model backends.

References

1. *SPRING 2025 Reference Group Executive Summary*. (n.d.). https://www.acha.org/wp-content/uploads/NCHA-IIIb_SPRING_2025_REFERENCE_GROUP_INSTITUTIONAL_EXECUTIVE_SUMMARY.pdf
2. *One in eight adolescents and young adults use AI chatbots for mental health advice*. (2025, November 12). School of Public Health | Brown University. https://sph.brown.edu/news/2025-11-18/teens-ai-chatbots
3. Legatt, D. A. (2025, September 18). 90% Of College Students Use AI: Higher Ed Needs AI Fluency Support Now. *Forbes*. https://www.forbes.com/sites/avivalegatt/2025/09/18/90-of-college-students-use-ai-higher-ed-needs-ai-fluency-support-now/
4. Litman, J. A. (2008). Epistemic Curiosity Questionnaire. *PsycTESTS Dataset*. https://doi.org/10.1037/t11264-000
5. Kashdan, T. B., Disabato, D. J., Goodman, F. R., & McKnight, P. E. (2020). The Five-Dimensional Curiosity Scale Revised (5DCR): Briefer subscales while separating overt and covert social curiosity. *Personality and Individual Differences*, *157*, 109836. https://doi.org/10.1016/j.paid.2020.109836
6. Kashdan, T. B., Gallagher, M. W., Silvia, P. J., Winterstein, B. P., Breen, W. E., Terhar, D., & Steger, M. F. (2009). The curiosity and exploration inventory-II: Development, factor structure, and psychometrics. *Journal of Research in Personality*, *43*(6), 987–998. https://doi.org/10.1016/j.jrp.2009.04.011

7. Litman, J. A., & Spielberger, C. D. (2003). Measuring Epistemic Curiosity and Its Diversive and Specific Components. *Journal of Personality Assessment*, *80*(1), 75–86. https://doi.org/10.1207/s15327752jpa8001_16
8. Berlyne, D. E. (1951). A theory of human curiosity [Review of *A theory of human curiosity*]. *British Journal of Psychology*, *45*, 180–191. https://psycnet.apa.org/record/1955-03567-001
9. Litman, J. A. (2008). Interest and deprivation factors of epistemic curiosity. *Personality and Individual Differences*, *44*(7), 1585–1595. https://doi.org/10.1016/j.paid.2008.01.014
10. Whitecross, W. M., & Smithson, M. (2023). Curiously different: Interest-curiosity and deprivation-curiosity may have distinct benefits and drawbacks. *Personality and Individual Differences*, *213*, 112310. https://doi.org/10.1016/j.paid.2023.112310
11. Hilmerich, C., Hofmann, M. J., & Briesemeister, B. B. (2024). Anxiety and curiosity in hierarchical models of neural emotion processing—A mini review. *Frontiers in Human Neuroscience*, *18*. https://doi.org/10.3389/fnhum.2024.1384020
12. Li, Y., Huang, Y., Wang, H., Zhang, X., Zou, J., & Sun, L. (2024, June 25). *Quantifying AI Psychology: A Psychometrics Benchmark for Large Language Models*. ArXiv.org. https://doi.org/10.48550/arXiv.2406.17675
13. Wang, H., Jiang, S., Chen, Y., Wang, Y., & Xiao, Y. (2025). *Why Did Apple Fall To The Ground: Evaluating Curiosity In Large Language Model*. ArXiv.org. https://arxiv.org/abs/2510.20635
14. Jiang, H., Zhang, X., Cao, X., Breazeal, C., Kabbara, J., & Roy, D. (2024, February 26). *PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits*. ArXiv.org. https://doi.org/10.48550/arXiv.2305.02547

15. Heston, T. F., & Gillette, J. (2025). Large Language Models Demonstrate Distinct Personality Profiles. *Cureus*. https://doi.org/10.7759/cureus.84706

16. Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. Archives of Internal Medicine, 166(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092