JURNAL INFOTEL
VOL. 18, NO. 1, FEBRUARY 2026, PP. 42–58. DOI:10.20895/INFOTEL.V18I1.1456

RESEARCH ARTICLE

LLM-Based Interview Bot for Student Big Five Assessment and Career Recommendation

Sang Dara Parameswari[1], Muharman Lubis[2], Sinung Suakanto[3,*] and Jan M. Pawlowski[4]

[1,2,3]System Information Study Program, Telkom University, Bandung 40257, Indonesia
[4]Business Information Systems, Ruhr West University of Applied Sciences, Germany

*Corresponding email: sinung@telkomuniversity.ac.id

Received: September 30, 2025; Revised: March 17, 2026; Accepted: April 23, 2026.

Abstract: The development of Artificial Intelligence (AI) and Natural Language Processing (NLP) offers new opportunities to make psychological assessments more interactive and meaningful. However, personality tests such as the International Personality Item Pool – Big Five Factor Markers (IPIP-BFM-50) still rely on static self-report questionnaires, which may limit engagement and contextual interpretation. This study proposes an InterviewBot-based Big Five Personality system (IB-B5P) that combines rule-based IPIP scoring with Large Language Model (LLM)-driven conversational assessment using GPT-3.5 Turbo. The system generates both quantitative personality scores and qualitative narrative profiles. Evaluation results show moderate to strong correlations (r = 0.31–0.71) between IB-B5P and IPIP scores, with Openness and Extraversion showing statistically significant relationships. These findings suggest that the hybrid rule–LLM approach can approximate IPIP tendencies while providing richer context-aware interpretations. The novelty of this study lies in integrating LLM-based conversational intelligence with a standardized psychometric framework, with potential applications in career guidance, educational counseling, and digital psychological assessment in higher education.

Keywords: Big Five Personality, Career Recommendations, Interview bot, IPIP-BFM, Large Language Model

1 Introduction

Large Language Models (LLMs) have rapidly transformed higher education by enabling intelligent conversation agents and learning support systems [1]. As AI adoption rises

 ISSN: 2085-3688; e-ISSN: 2460-0997

globally—from 20 percent in 2017 to 72 percent in 2024 [2]—LLMs offer the potential to address persistent challenges in career guidance: subjective assessments, counselor shortages, and the difficulty of personalizing support for many students [3]. Consequently, higher education institutions urgently need innovative systems that deliver objective, consistent personality assessments and recommend career paths tailored to student profiles.

Several previous studies have highlighted LLMs' ability to express personality through generated text. Jiang et al. [4] showed, through their Persona LLM study, that LLMs can display consistent personality traits when guided by instructions based on the Big Five Personality Traits. Even the model's personality test results proved quite stable and recognizable to human evaluators. Another study by Jiang G et al. [5] proposed the Machine Personality Inventory (MPI) as an instrument for measuring personality traits in pre-trained models, while introducing personality prompting techniques capable of inducing specific traits; however, consistency remains a challenge in long-term interactions. Wang et al. [6] also found that LLMs exhibit high internal reliability and convergent validity when asked to mimic real personality profiles based on the Big Five framework, although predictions of external variables, such as organizational behavior, are not yet optimal. Another approach developed by Zheng et al. [7] is the Language Model Linguistic Personality Assessment (LMLPA), which uses linguistic features to quantitatively assess model personality and has been shown to effectively distinguish the Big Five personality traits. Additionally, Maharjan et al. [8] demonstrated that LLM embedding representations can be used to predict personality traits with fairly high psychometric validity, while Li et al. [9] highlighted the importance of assessment methods, showing that the forced-choice approach is more effective than the Likert scale in minimizing model response bias. Finally, Goldberg [10] emphasized that the Big Five Personality Traits framework remains the most widely used personality theory in psychology, making it a relevant basis for AI-based assessment systems.

Conventional personality assessments such as the International Personality Item Pool – Big Five Factor Markers (IPIP-BFM-50) have been widely used and validated in psychological research. However, they require respondents to answer a fixed set of Likert-scale items, which can be monotonous, time-consuming, and limited in scope. These instruments measure only predefined aspects of personality, without allowing respondents to elaborate on their unique experiences or demonstrate behavioral nuances in natural communication. Moreover, interpreting IPIP scores still requires significant effort on the part of career counselors to translate numerical values into meaningful career guidance, creating inefficiencies in practice. In contrast, an InterviewBot approach leveraging LLMs can transform this process by asking open-ended, natural questions, interpreting free-text responses, and extracting not only personality traits but also additional indicators such as competencies or communication style. The InterviewBot can also combine and generalize conventional items into fewer, more engaging prompts, making the assessment process more effective in gathering rich information while reducing participant fatigue. This highlights a critical gap between the rigid, questionnaire-based IPIP method and the more dynamic, conversational assessment approach that modern LLMs can support. In our preliminary development, we have explored the concept of an InterviewBot that poses interview-style questions and has the human user provide narrative responses [11]. These responses can then be processed using Natural Language Processing (NLP) and machine learning techniques to score and interpret them [12]. In this study, we extend that approach by employing LLMs as the pri-

mary engine for question generation, response interpretation, and profile summarization, thereby advancing the capabilities of automated personality assessment systems.

Compared with traditional psychometric questionnaires, Large Language Models (LLMs) offer several advantages for assessing personality. Static questionnaires use fixed response options for each item. In contrast, a conversational LLM elicits more open-ended, richly descriptive text from users, better capturing personal experiences and behaviors [13]. This conversational approach allows for a more flexible, engaging, and detailed assessment. It can reveal psychological content often missed by standard questionnaires [14]. LLMs also generate human and natural-sounding narrative summaries and career suggestions, which benefit both counselors and students. Integrating LLMs with psychometric methods enables more efficient and meaningful personality assessments [15].

In terms of state-of-the-art, there is a growing trend of integrating AI and natural language processing into psychological and educational tools. For example, conversational agents have been adopted in mental health contexts to deliver scalable support [16], while adaptive testing systems have been used to shorten assessment length without compromising validity [17]. Recent advancements in multimodal LLMs also hold promise for analyzing diverse inputs, such as text and speech, to produce more holistic assessments [18]. However, few studies have attempted to directly integrate LLM-driven conversational assessments with established psychometric frameworks such as the Big Five, especially in the context of career guidance for students. This positions the proposed research at the intersection of psychology, education, and AI system design. Unlike prior studies that primarily evaluate personality traits in LLM behavior, this study develops a practical InterviewBot system that integrates conversational LLM interaction with the validated IPIP-BFM-50 instrument. This integration enables both quantitative personality scoring and qualitative narrative interpretation for career guidance in higher education.

This research aims to design and develop an InterviewBot system using LLMs for automated Big Five personality assessment in higher education. The system will provide engaging, open-ended interactions that help students better understand their own personalities and identify suitable career paths. These interactions will yield both quantitative trait scores and qualitative narrative profiles, offering actionable guidance for students. Career path recommendations will align with students' personalities. IPIP will serve as a validated benchmark to evaluate the system's reliability and validity. This approach improves interactivity and efficiency while aligning with established psychometric standards.

2 Research Method

This study proposes InterviewBot (IB-B5P), a novel system for personality assessment that uses Large Language Models (LLMs) within the Big Five framework. Unlike existing approaches, our system enables automated, interactive assessment via conversational AI. To guide its development, we adopt the Design Science Research (DSR) methodology, a structured approach to designing, building, and evaluating systems [19]. The DSR framework ensures scientific rigor and practical relevance by integrating analysis and problem identification, design, and system evaluation. Figure 1 illustrates the detailed stages of the research process in this study.

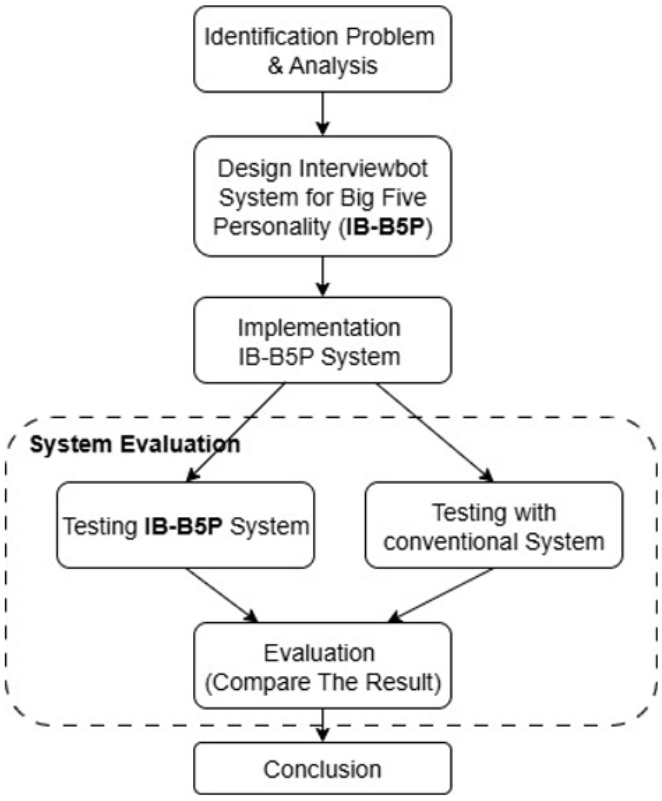

Figure 1: Research method.

2.1 Identification Problem & Analysis

In this phase, the study identified several key issues in the current practice of personality assessment in higher education. Conventional methods such as the International Personality Item Pool – Big Five Factor Markers (IPIP-BFM-50) rely on structured self-report questionnaires, where respondents must answer a fixed set of 50 items with Likert-scale responses. To better understand the role of the International Personality Item Pool (IPIP) in this study, it is important to briefly review the attributes and traits it measures. The IPIP is designed to operationalize the Big Five Personality Traits (OCEAN), which consist of five major dimensions: Openness to Experience, reflecting creativity, curiosity, and willingness to engage with new ideas; Conscientiousness, indicating organization, responsibility, and goal-directed behavior; Extraversion, representing sociability, assertiveness, and energy in interpersonal interactions; Agreeableness, associated with empathy, trust, and cooperativeness; and Neuroticism (sometimes referred to as Emotional Stability in the reversed form), which reflects the tendency to experience negative emotions such as anxiety, stress, or mood instability. Each trait is assessed by multiple items in the IPIP-BFM-50 instrument, allowing quantitative evaluation of personality across these five dimensions (Table 1). The early phases of problem identification and business requirement elicitation are essential, as they ensure that the solutions developed are contextually relevant and aligned with organizational needs, thereby supporting their successful application in the business world [20].

This structured measurement provides a strong foundation for personality assessment and will serve as the benchmark for comparison with the proposed InterviewBot system. This instrument is widely used because it is open, validated, and highly reliable in personality psychology research [21, 22]. Respondents provided answers using a Likert scale (1 = strongly disagree with 5 = strongly agree). Positively keyed items were calculated directly according to the scale values, while negatively keyed items were reversed for consistent interpretation. The total score per dimension was obtained by summing all items, then dividing by the number of items to produce an average representing the individual's personality traits. This average value was then used as ground truth to compare the results of the LLM-based interview bot assessment [23, 24].

Table 1: Example of IPIP scoring

No	Item Question (Example)	Keyed	Respon Likert	Score
1	'I like trying new things.'	+	Very Accurate (5)	5
2	'I often put off important tasks.'	−	Moderately Accurate (4)	2
3	'I get along easily with other people.'	+	Moderately Accurate (4)	4
4	'I rarely help other people.'	−	Very Inaccurate (1)	5
5	'I feel anxious in many situations.'	+	Neither (3)	3

The calculations for Table 1 are as follows:

1. Item (+) keyed, scores follow the Likert scale (1–5).
2. Item (–) keyed, scores are reversed (e.g., Very Accurate = 1, Very Inaccurate = 5).
3. Total dimension score = sum of all item scores. Eq. (1) shows the final dimension score.

$$\text{Final Dimension Score} = \frac{\text{Average total score}}{\text{Number of items}} \tag{1}$$

As an example of the calculation results, the scores for the five IPIP items given by respondents were 5, 2, 4, 5, and 3. When added together, the total score obtained was 19. Next, the total score was divided by the number of items (5) to obtain an average dimension score of 3.8. This average value then represents the respondents' tendencies towards the personality dimension being measured [10, 25].

While scientifically validated, this method is often perceived by students as rigid, repetitive, and monotonous. It captures only direct responses to predefined questions and does not allow deeper exploration or elaboration of the respondent's unique characteristics. Furthermore, interpreting scores requires additional effort from career counselors, which can reduce the efficiency with which timely and personalized guidance is delivered. In higher education, career guidance services are important for student readiness but remain constrained by limited counselors, limited-service scale, and the subjectivity of assessment [26]. Although the IPIP has certain limitations in terms of interactivity, it remains a validated and widely recognized instrument. Therefore, in this study, IPIP will serve as a benchmark for comparing the proposed system's results.

To address these limitations, a system analysis was conducted by mapping the user perspective, psychological insights, and available technological opportunities. From the user side, both students and counselors expect an assessment tool that is not only accurate but also engaging, interactive, and capable of producing meaningful narrative insights beyond raw numerical scores. From the psychological framework, the Big Five (OCEAN) remains the dominant and widely adopted model for measuring personality. On the technological side, recent advancements in Large Language Models (LLMs) enable the generation of interview-style questions, the processing of open-ended responses, and the production of human-like summaries of personality profiles. Unlike conventional IPIP, the InterviewBot approach allows the extraction of additional information, such as indicators of competencies, communication styles, and other relevant behavioral attributes, from the interviewee's elaborated answers. Moreover, LLM-driven question design can combine or generalize multiple aspects of conventional items into fewer, more natural questions, making the assessment process more efficient while still comprehensive.

This analysis highlights a potential solution: an InterviewBot system that integrates LLMs with structured scoring mechanisms. Such an approach offers not only efficiency and engagement but also the flexibility to extract richer data from participants, ultimately enhancing the value of personality assessment for higher education. At this stage, we use IPIP as a validated method, ensuring that the evaluation of the InterviewBot system can be rigorously compared against an established standard.

2.2 Design System

In this stage, we introduce the InterviewBot for Big Five Personality (IB-B5P) as the core system design. The IB-B5P framework combines the Big Five personality model with Large Language Models (LLMs) to deliver interactive personality assessments through natural language conversations. The system generates interview-style questions, captures open-ended responses, transforms them into structured trait scores, and provides both narrative summaries and career recommendations. Figure 2 illustrates the overall architecture and workflow, showing the main components and their interactions.

The proposed InterviewBot system, referred to as IB-B5P, is designed to automate the personality assessment process through structured interactions between participants and

an LLM-powered interviewer bot. As illustrated in Figure 2, the process begins with the Question Generator, which formulates interview-style questions aligned with the Big Five Personality Traits framework. These questions are delivered to the participants who provide text-based responses during the interview session.

The collected responses are then processed by the system in two sequential steps. First, the Answer Reception module stores the participant's input as text. Second, the Answer Conversion module transforms these responses into quantifiable scores based on the Big Five dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. This scoring process utilizes prompt engineering strategies within the LLM to map open-ended answers onto structured personality scales.

Following the scoring phase, the system moves to Profile Generation, where the LLM creates a comprehensive personality profile with both numerical trait scores and narrative summaries providing context for each participant. Finally, the results are presented, enabling participants to gain concise, interpretable insights into their personality profiles.

The system development in this research focuses not merely on traditional development, but on designing an LLM-based framework that integrates prompt engineering strategies and a structured data processing pipeline [27]. At this stage, the system uses a Large Language Model (OpenAI GPT-3.5 Turbo) via an API management module to generate adaptive interview questions and process student responses. To ensure more consistent, easily processed model outputs, a structured, template-based prompt engineering strategy is used, requesting results in JSON format with elements such as title, percentage, and the reason for each mapped category. The LLM-generated outputs are then combined with the IPIP score calculated through the rule-based scoring engine, enabling the system to present a quantitative and narrative Big Five-based personality assessment.

2.3 Prompt Engineering Strategy

To ensure reliable, easy-to-understand results, the proposed system uses a clear prompt engineering approach when working with the Large Language Model (LLM). Prompt engineering helps the model create questions about personality and understand user answers using the Big Five personality model [28]. With simple instructions, the model provides structured replies that the system can automatically use, turning the usual paper personality tests into a more natural conversation.

The prompt design follows an instruction-based structure consisting of role definition, task specification, contextual information, and output constraints. The model is instructed to act as a psychological assessment assistant, generating behavioral questions and analyzing user responses. The generated outputs follow a structured JSON format that includes attributes such as the scenario description, narrative question, personality trait identifier, and keying indicator. This structured format enables the system to automatically parse the generated outputs and integrate them into the personality scoring module [29].

Since users provide responses in natural language rather than selecting predefined options, an additional prompt is used to classify each response into a numerical score. Specifically, the model assesses the semantic relationship between the question scenario and the user's narrative response, assigning a score on a five-point Likert scale. Next, the scores are aggregated across the five personality dimensions, with reverse scoring applied to negatively keyed items. In this way, the mechanism ensures compatibility with established psychometric assessment approaches while leveraging the flexibility of conversational in-

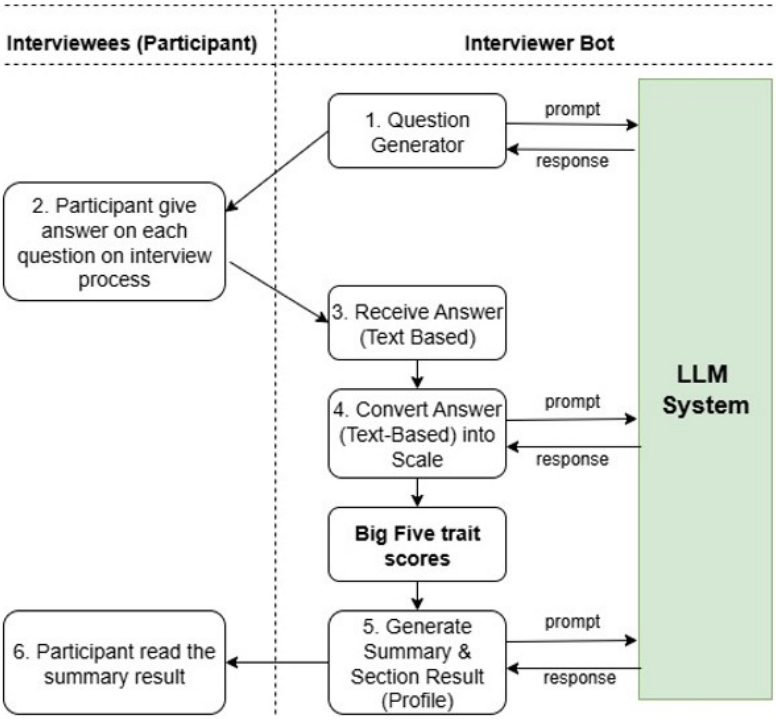

Figure 2: Proposed system (IB-B5P) architecture.

teraction enabled by LLMs [30]. To further clarify this methodology, the prompt engineering components used during the interaction with the language model are summarized in Table 2.

Table 2: Prompt engineering structure used in the proposed system

Component	Description	Example
Role Instruction	Defines the role of the language model in the assessment process	"You are psychological assessment assistant based on the Big Five personality framework"
Task Instruction	Specifies the task performed by the model	"Analyze the user's response and determine its alignment with the personality indicator"
Context Information	Additional contextual information used to personalize prompts	User academic major and preferred language
Output Constraint	Defines the required output format for system processing	Return a score from 1-5 representing the personality indicator

Table 2 summarizes the key elements of the prompt engineering framework applied in the proposed system. The prompt architecture comprises role specification, task delineation, contextual augmentation, and output formatting constraints. Each component directs the language model during the psychometric evaluation process. Role specification articulates the model's function as a psychological assessment tool. Task delineation details operations required of the model, such as analyzing respondent inputs or constructing personality-diagnostic questions. Contextual augmentation tailors the generated content to the respondent's demographic or psychographic profile. Output formatting constraints require the model to deliver responses in a machine-readable format for automated processing. This configuration enables the language model to consistently generate unambiguous outputs supporting the personality-scoring algorithm.

3 Results

3.1 Implementation Result

This part describes the application developed to support the personality assessment process using Big Five Personality and advanced analysis based on Large Language Model (LLM). The application is designed to provide an interactive assessment experience. Users begin by entering initial data and then complete questionnaire items and open-ended questions. The results are presented with graphical visualizations and narrative descriptions, along with quantitative scores for each dimension of the Big Five Personality Traits. Thus, this application serves as a psychometric measurement instrument and a tool for student self-reflection and career planning, powered by the latest artificial intelligence technology.

The front-page form, as shown in Figure 3, serves as the starting point for the assessment, where users fill in basic information before beginning the test. This screen displays the assessment title and the five main personality dimensions to be measured, namely Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Intel-

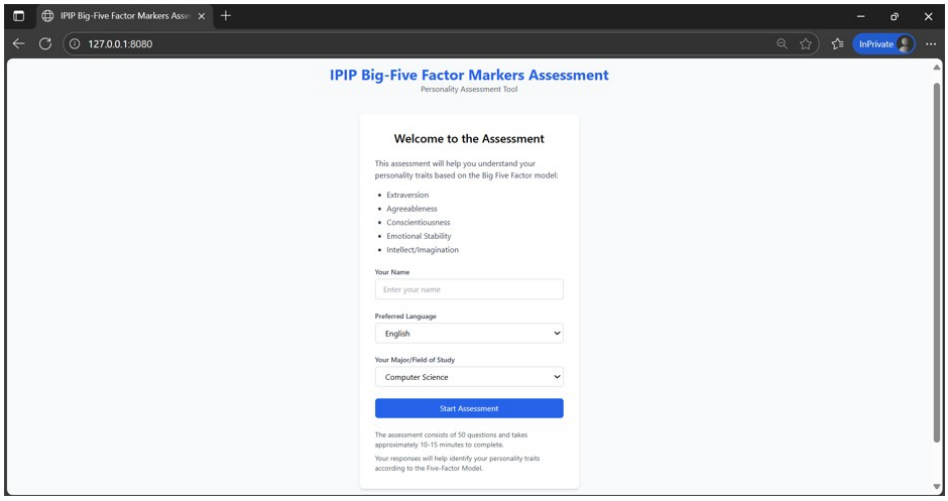

Figure 3: The system front page form.

lect/Imagination. Additionally, there are several input fields, such as name, preferred language, and study program, that aim to contextualize the assessment results based on the user's background. The Start Assessment button starts the IPIP-BFM-50 test. With this design, the front page not only functions as a user-friendly initial interface but also ensures that the user's basic data is properly recorded to support the interpretation of the assessment results.

Figure 4 shows the interview form used in the proposed system. This interface enables participants to respond to questions generated by the InterviewBot. The system presents each question in the main area, along with a relevant description or scenario that reflects one of the Big Five Personality Traits. Participants provide narrative or experience-based answers in the supplied text field. To help users navigate, the system organizes all items by their respective personality dimensions using a Question Navigator, letting participants track progress and revisit earlier questions as needed. Additional options allow participants to proceed to the next question or cancel the session, ensuring flexibility in the interview process. Overall, this form serves as an interactive medium, allowing participants to engage in interviews and supply input for subsequent personality analysis.

The result page, as shown in Figure 5, is the final stage of the personality evaluation process, based on five main dimensions: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Intellect/Imagination. At this stage, the system displays personality scores in an easy-to-read graph, accompanied by a narrative description of the participant's psychological tendencies. Each dimension is explained separately to highlight behavioral characteristics, social interaction patterns, and potential related to academic and professional contexts. In addition, the system provides personal recommendations on career direction, learning style, team role, communication patterns, and social compatibility. This additional information is obtained through LLM response analysis, so the results displayed are more adaptive and tailored to each individual's profile. By combining quantitative data from IPIP scores with qualitative analysis using an LLM, the results page

serves as a medium for self-reflection and a practical guide for participants in planning their self-development and career direction.

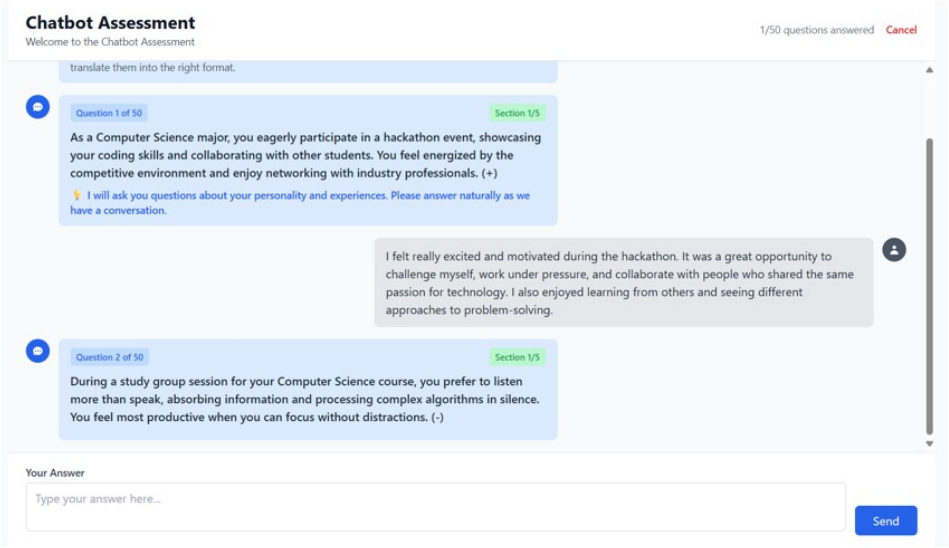

Figure 4: Assessment page.

3.2 Evaluation

The evaluation involved 40 undergraduate students who participated in both the IPIP-BFM-50 questionnaire and the IB-B5P interview-based assessment. Each participant completed the conventional IPIP personality test and the interview interaction with the proposed system. In the evaluation stage, the study compares the results from the proposed IB-B5P system (Method A) with those from the conventional personality assessment instrument, IPIP-BFM-50 (Method B). Specifically, the comparison is carried out across the five personality dimensions of the Big Five framework: Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N). For each respondent, the percentage scores generated by both methods are recorded, and the difference (Δ) between them is calculated. Notably, a larger Δ value indicates a greater divergence between the proposed system and the established IPIP assessment in identifying specific traits, whereas a smaller Δ suggests that the IB-B5P system produces more consistent results with the validated conventional method. To further assess consistency, each respondent was tested using both methods to examine whether the results remain stable across approaches.

An example of raw evaluation data from several respondents is presented in Table 3, which compares Method A and Method B across the five traits. This tabular representation provides an initial overview of how closely the IB-B5P system aligns with the reference standard and highlights areas where differences are more pronounced.

To evaluate the proposed IB-B5P system, two statistical tests were conducted to compare its results with the standard IPIP scores, which served as the reference tool. As shown in Table 4, the comparison was intended to assess the connection and any potential dif-

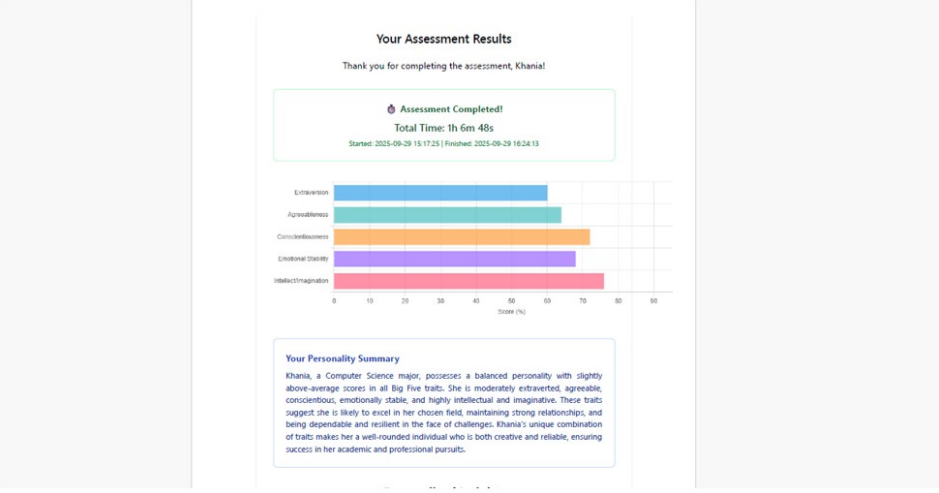

Figure 5: Result page.

Table 3: Example of evaluation result from respondents

No	Name	O			C			E			A			N		
		A	B	Δ	A	B	Δ	A	B	Δ	A	B	Δ	A	B	Δ
1	Respondent-1	59	84	-25	93	94	-1	58	58	0	71	78	-7	65	64	1
2	Respondent-2	59	54	5	62	60	2	46	56	-10	46	64	-18	48	62	-14
3	Respondent-3	40	40	0	41	60	-19	41	52	-11	60	68	-8	35	60	-25
4	Respondent-4	76	74	2	60	56	4	80	75	5	70	57	13	56	68	-12
5	Respondent-5	45	54	-9	88	70	18	68	82	-14	84	80	4	80	74	6

ferences between the two systems for the five personality traits. The Pearson correlation measures how strongly the IPIP and IB-B5P results are linked, indicating whether both systems yield similar high and low trait scores. The Wilcoxon signed-rank test checks if the differences between the two tools are meaningful. Together, these tests help show how closely the IB-B5P matches the established IPIP system.

Table 4: Comparative analysis of personality trait scores between IB-B5P and IPIP

Trait	**Mean (IPIP)**	**Mean (IB-B5P)**	**Mean Diff.**	**Pearson r**	**Interpretation (r)**	**Wilcoxon p**	**Significance**
Openness	54.65	62	7.35	0.662	Moderate–strong positive correlation	0.002	Significant
Conscientiousness	72.4	71.03	−1.38	0.559	Moderate positive correlation	0.428	Not significant
Extraversion	50.78	58.05	7.28	0.708	Strong positive correlation	0.006	Significant
Agreeableness	62.58	67.28	4.7	0.312	Weak correlation	0.054	Borderline
Neuroticism	58.8	61.88	3.08	0.479	Moderate correlation	0.27	Not significant

4 Discussion

Table 4 shows that the proposed IB-B5P system and the IPIP instrument generate consistent results across the five personality traits. Correlation coefficients (r = 0.312-0.708) indicate

that IB-B5P matches IPIP patterns, especially for Extraversion (r = 0.708) and Openness (r = 0.662), which show strong positive correlations. Conscientiousness and Neuroticism have moderate correlations, while Agreeableness shows a weaker yet positive relationship. Overall, IB-B5P generally mirrors the trait rankings found by IPIP, but the alignment strength varies by trait.

The Wilcoxon signed-rank test results indicate that most differences between the two instruments are not statistically significant, except for Openness and Extraversion, where significant differences were observed ($p < 0.01$). This implies that, while the two systems are directionally consistent, IB-B5P may yield slightly higher or lower absolute scores for certain traits. Nevertheless, the overall pattern supports the conclusion that IB-B5P yields outcomes comparable to IPIP in assessing personality tendencies. Hence, the proposed system demonstrates acceptable validity as an alternative approach, capable of approximating IPIP trait measurements while maintaining similar interpretive trends, as discussed in the following section.

In interpreting the results, it is important to note that the observed mean differences between IB-B5P and IPIP scores remain relatively small when considered within the 0–100 scale. Such variations are expected in psychological assessments, as even standardized instruments like IPIP do not always yield identical scores when administered multiple times to the same respondent. Instead, consistency is typically reflected in maintaining similar tendencies, for instance, whether a respondent consistently scores in the lower, middle, or higher range of a trait. This characteristic reflects the inherent variability in self-reported measures rather than a flaw in the instrument. To support this notion, a repeated-measures test was conducted using IPIP itself (see Table 5), which shows that individual scores fluctuate slightly across trials while maintaining stable overall personality tendencies.

Although numerical differences may exist between the two measurement systems, it is unnecessary to expect identical or near-identical scores across instruments. What is more important is the directional consistency - whether the scores from both systems fall within the same general range of low, medium, or high tendencies. As shown in Tables 3 and 4, the IB-B5P system may yield slightly different mean scores than the IPIP benchmark; however, the overall correlation and rank-order consistency indicate that both instruments capture similar personality patterns. This suggests that the proposed interview-based IB-B5P system can effectively approximate the IPIP assessment and serve as a reliable alternative for personality trait evaluation, particularly when a structured interview format is more suitable than a self-report questionnaire.

Table 5: Stability of IPIP personality trait scores in repeated measurements

No	Name	O			C			E			A			N		
		T1	T2	Δ	T1	T2	Δ	T1	T2	Δ	T1	T2	Δ	T1	T2	Δ
1	Respondent-1	28	59	31	26	22	4	74	70	4	45	45	0	89	89	0
2	Respondent-2	59	34	25	64	76	12	50	41	9	72	56	16	80	66	14
3	Respondent-3	3	11	8	92	87	5	78	78	0	76	82	6	66	74	8
4	Respondent-4	65	70	5	80	92	12	21	19	2	40	25	15	74	91	17
5	Respondent-5	46	59	13	80	67	13	84	74	10	74	89	15	36	66	30
6	Respondent-6	34	59	25	67	22	45	34	41	7	30	10	20	16	26	10

From the empirical evaluation, the comparison between IB-B5P and IPIP (Table 3) demonstrates that the proposed system produces results that are directionally consistent

with the standard instrument. Despite small mean differences across traits, the moderate to strong positive correlations (r = 0.56–0.71 for most traits) indicate that both systems capture similar personality tendencies. In addition, the Wilcoxon signed-rank test showed that most differences were statistically insignificant, suggesting that the IB-B5P scores are not meaningfully different from IPIP at the group level. Therefore, the interview-based IB-B5P can be considered a valid approximation of IPIP in identifying high, medium, and low personality tendencies. The results of developing an interview bot system based on the IPIP-BFM-50 and a Large Language Model (LLM) show that a hybrid approach combining a rule engine and generative AI can support student personality assessment. This aligns with previous studies that emphasize the importance of structured prompting to improve the stability of language model output [31].

In addition to generating personality profiles, the system provides career recommendations based on Big Five assessment results. After identifying dominant traits, it interprets these findings in relation to job preferences. For instance, high Extraversion suggests a fit for roles that involve communication and teamwork, while high Conscientiousness indicates a fit for structured, meticulous work. High Openness aligns with creative or research fields, and high Agreeableness with collaborative or service-oriented professions. The system then suggests career options aligned with students' personality characteristics.

From a practical perspective, the system can present assessment results in quantitative graphs and interpretative narratives. The results page not only provides information about personality scores but also offers additional insights, such as career recommendations, communication styles, and social compatibility. This feature makes the system more useful than conventional methods that only present numerical scores. Thus, the developed application has the potential to serve as a tool for self-reflection and support career guidance services in higher education.

Overall, this study contributes to two main aspects. First, it presents an integration framework between standard psychometric instruments (IPIP-BFM-50) and an LLM that enables richer, more adaptive assessments. Second, it provides practical implications for career guidance services in higher education through an interactive, flexible, and easy-to-use web-based system. Moving forward, the research can be further developed by conducting more comprehensive validity and reliability tests, expanding the sample size, and comparing the system's results with those of other psychological assessment methods.

This study acknowledges several limitations, including the relatively small sample of 40 undergraduate students used in the evaluation, which may limit the generalizability of these findings to a broader, more heterogeneous population. Furthermore, the personality assessment process relied solely on text-based responses, raising the possibility that individual variations in communication style and written language proficiency may have influenced the quality and consistency of the collected data. In addition, the career recommendations generated by the system remain broad in scope, as they stem from general interpretations of the Big Five personality dimensions, rather than from empirically validated job profiles or detailed, specific job criteria. Given these limitations, it is hoped that future research will expand the participant pool by including a larger, more diverse sample. It is also recommended that subsequent studies integrate additional psychological assessment methods to complement the existing framework and use more comprehensive, structured job datasets to enhance the accuracy and practical application of the career recommendation component.

5 Conclusion

This study presents the design, implementation, and evaluation of an InterviewBot system (IB-B5P) that integrates the International Personality Item Pool – Big Five Factor Markers (IPIP-BFM-50) with Large Language Model (LLM) technology for personality assessment in higher education. The results demonstrate that combining a rule-based scoring mechanism with generative AI enables a balanced approach that produces both quantitative and qualitative outputs, maintaining psychometric rigor while enhancing user engagement. The evaluation results indicate that the IB-B5P system achieves moderate to strong correlations (r = 0.31–0.71) with the standard IPIP assessment, particularly for the traits of Openness and Extraversion, for which statistically significant relationships were observed. These findings suggest that the proposed system can effectively replicate the personality tendencies measured by IPIP, thereby validating its potential as a complementary or alternative assessment tool. Moreover, structured prompt engineering and JSON-formatted outputs improve the consistency, interpretability, and traceability of LLM-generated results.

From a practical standpoint, IB-B5P extends beyond traditional questionnaire-based approaches by providing interactive, interview-style assessments that generate narrative interpretations, career recommendations, and behavioral insights. This makes the system particularly useful for career guidance and self-reflection, offering students a more personalized and engaging assessment experience. In conclusion, this research advances AI-assisted psychometric assessment by demonstrating how LLMs can be integrated with standardized psychological instruments to create adaptive, interpretable, and efficient personality assessment systems. Future work will focus on expanding the participant base, performing comprehensive validity and reliability tests, and exploring integration with multimodal inputs or other psychological frameworks to enhance the robustness and applicability of the InterviewBot system in educational and organizational contexts.

References

[1] P. Wang, H. Zou, H. Chen, T. Sun, Z. Xiao, and F. L. Oswald, "Personality structured interview for large language model simulation in personality research," *arXiv preprint arXiv:2502.12109*, 2025.

[2] McKinsey & Company, "The state of ai in early 2024: Gen ai adoption spikes and starts to generate value." https://www.mckinsey.com, 2024.

[3] J. B. Monreal and T. Palaoag, "Use of artificial intelligence in career guidance: Perspectives of secondary guidance counselor," *International Journal of Information and Education Technology*, vol. 14, no. 2, 2024.

[4] H. Jiang, X. Zhang, X. Cao, C. Breazeal, D. Roy, and J. Kabbara, "Personallm: Investigating the ability of large language models to express personality traits," *arXiv preprint arXiv:2305.02547*, 2024.

[5] G. Jiang, M. Xu, S.-C. Zhu, W. Han, C. Zhang, and Y. Zhu, "Evaluating and inducing personality in pre-trained language models," *arXiv preprint arXiv:2206.07550*.

[6] Y. Wang, J. Zhao, D. S. Ones, L. He, and X. Xu, "Evaluating the ability of large language models to emulate personality," *Scientific Reports*, vol. 15, no. 1, 2025.

[7] X. Zheng *et al.*, "Lmlpa: Language model linguistic personality assessment," *Computational Linguistics*, 2025.

[8] J. Maharjan, R. Jin, J. Zhu, and D. Kenne, "Psychometric evaluation of large language model embeddings for personality trait prediction," *Journal of Medical Internet Research*, vol. 27, 2025.

[9] X. Li, H. Shi, Z. Yu, Y. Tu, and C. Zheng, "Decoding llm personality measurement: Forced-choice vs. likert," *arXiv preprint arXiv:2401.XXXXX*.

[10] L. R. Goldberg, "The development of markers for the big-five factor structure," *Psychological Assessment*, vol. 4, no. 1, pp. 26–42, 1992.

[11] S. Suakanto, J. Siswanto, T. F. Kusumasari, I. R. Prasetyo, and M. Hardiyanti, "Interview bot for improving human resource management," in *Proc. IEEE Int. Conf. Information Systems and Software Technologies (ICISS)*, 2021.

[12] J. Siswanto, S. Suakanto, M. Andriani, M. Hardiyanti, and T. F. Kusumasari, "Interview bot development with natural language processing and machine learning," *International Journal of Technology*, vol. 13, no. 2, 2022.

[13] H. Zou, P. Wang, Z. Yan, T. Sun, and Z. Xiao, "Can llm self-report? evaluating the validity of self-report scales in measuring personality design in llm-based chatbots," *arXiv preprint arXiv:2412.00207*, 2025.

[14] L. Ke, S. Tong, P. Cheng, and K. Peng, "Exploring the frontiers of llms in psychological applications: A comprehensive review," *Artificial Intelligence Review*, vol. 58, no. 10, 2025.

[15] H. Ye, J. Jin, Y. Xie, X. Zhang, and G. Song, "Large language model psychometrics: A systematic review of evaluation, validation, and enhancement," *arXiv preprint arXiv:2505.08245*, 2026.

[16] J. Torous *et al.*, "The growing field of digital psychiatry," *World Psychiatry*, vol. 20, no. 3, pp. 318–335, 2021.

[17] L. T. Car *et al.*, "Conversational agents in health care," *Journal of Medical Internet Research*, vol. 22, no. 8, 2020.

[18] M. S. Yiğiter and N. Doğan, "Comparison of different computerized adaptive testing approaches," *Journal of Measurement and Evaluation in Education and Psychology*, vol. 14, no. 4, pp. 396–412, 2023.

[19] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.

[20] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying crisp-dm process model," in *Procedia Computer Science*, vol. 181, pp. 526–534, 2021.

[21] A. Sorokovikova, N. Fedorova, S. Rezagholi, and I. P. Yamshchikov, "Llms simulate big five personality traits: Further evidence," *arXiv preprint arXiv:2402.01765*, 2024.

[22] W. Lenhard and A. Lenhard, "Improvement of norm score quality via regression-based continuous norming," *Educational and Psychological Measurement*, vol. 81, no. 2, pp. 229–261, 2021.

[23] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis)contents: A survey of dataset development and use in machine learning research," *Patterns*, vol. 2, no. 11, 2021.

[24] K. M. Shum, M. Ptaszynski, and F. Masui, "Big five personality trait prediction based on user comments," *Information*, vol. 16, no. 5, 2025.

[25] R. S. Wulandari, P. D. Kusuma, and C. Setianingsih, "Sistem pemetaan faktor kepribadian big five sebagai rekomendasi pemilihan pekerjaan dengan algoritma c4.5," *e-Proceeding of Engineering*, vol. 8, 2021.

[26] A. M. Al-Zahrani and T. M. Alasmari, "Exploring the impact of artificial intelligence on higher education: The dynamics of ethical, social, and educational implications," *Humanities and Social Sciences Communications*, vol. 11, no. 1, 2024.

[27] M. Rahman *et al.*, "Artificial intelligence in career counseling: A test case with resumai," *arXiv preprint arXiv:2308.14301*, 2023.

[28] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, 2023.

[29] Y. Zhou *et al.*, "Large language models are human-level prompt engineers," *arXiv preprint arXiv:2211.01910*, 2023.

[30] L. Qin *et al.*, "Large language models meet nlp: A survey," *Frontiers of Computer Science*, vol. 20, no. 11, p. 2011361, 2026.

[31] S. Han, D. Zhang, Y. Sui, M. Zhou, and D. Zhang, "Evaluating and enhancing structural understanding capabilities of large language models on tables via input designs," in *Proc. ACM Int. Conf. Web Search and Data Mining (WSDM)*, 2024.