ISSN 1840-4855
e-ISSN 2233-0046

Original scientific article
http://dx.doi.org/10.70102/afts.2026.1835.482

IMPROVING DIAGNOSTIC PRECISION AND THERAPEUTIC EMPATHY IN UNDERGRADUATE PSYCHOLOGY TRAINING WITH GENERATIVE AI-DRIVEN CLINICAL PATIENT PERSONAS

Odiljon Qobilov[1*], Iltifotxon Abdinazarova[2], Saida Makhkamova[3], Gulbahor Kholdorova[4], Mokhira Kuvvatova[5], Komila Umarova[6], Gulchexrabonu Isamova[7]

[1]Senior Lecturer, Tashkent State Medical University, Tashkent, Uzbekistan.*
e-mail: odqobilov776@gmail.ru, orcid: https://orcid.org/0009-0008-3190-3858
[2]Assistant, Tashkent State Medical University, Tashkent, Uzbekistan.
e-mail: iltifot2201@gmail.com, orcid: https://orcid.org/0009-0003-7300-7780
[3]Lecturer, Tashkent State University of Oriental Studies, Tashkent, Uzbekistan.
e-mail: saida_maxkamova@tsuos.uz, orcid: https://orcid.org/0009-0001-3358-7624
[4]Lecturer, Jizzakh State Pedagogical University, Jizzakh, Uzbekistan
e-mail: xoldorovaguli@gmail.com, orcid: https://orcid.org/0009-0008-8169-1849
[5]Associate Professor, Termez State Pedagogical Institute, Termez, Uzbekistan.
e-mail: mohira.quvvatova@mail.ru, orcid: https://orcid.org/0009-0002-2138-0754
[6]Lecturer, Department of Philology, Language Training Center, Alfraganus University, Tashkent, Uzbekistan. e-mail: umarovakomila1983@gmail.com, orcid: https://orcid.org/0009-0002-8652-2274
[7]Lecturer, Kimyo International University in Tashkent, Tashkent, Uzbekistan.
e-mail: g.isamova@kiut.uz, orcid: https://orcid.org/0009-0000-2165-0617

Received: January 17, 2026; Revised: February 27, 2026; Accepted: April 17, 2026; Published: May 29, 2026

SUMMARY

The conventional undergraduate psychology programs find it challenging to offer genuine clinical exposure because of the ethical considerations, the high prices of the standardized patients, and the lack of dynamism of the written case studies. The paper addresses the idea of integrating Generative AI (GenAI)-based clinical persona as a modification that can be scaled to close the gap between theoretical concepts and clinical environments. With the help of Large Language Models (LLMs) trained on symptom clusters and separate linguistic styles that correspond to the DSM-5-TR, a dynamic clinical sandbox was created and used to improve student skills. To measure two major metrics, Diagnostic Precision and Therapeutic Empathy, the study utilizes the mixed-methods approach. The diagnostic accuracy is measured with a Weighted Jaccard Index (WJI), which compares the student results with the programmed symptoms of a persona of the ground truth. Therapeutic empathy is evaluated by applying Natural Language Processing (NLP), whereby cosine similarity measures student dialogue to empathetic communication frameworks/theories that have been validated, e.g., the Person-Centered Theory developed by Rogers. Statistical results from a Randomized Control Trial demonstrate the efficacy of this framework, as students interacting with GenAI personas achieved a significantly higher diagnostic precision (Cohen's $\kappa = 0.78$, $p < 0.001$) vs. control ($\kappa = 0.62$) compared to those using traditional methods. A study was conducted randomized controlled trial (RCT) with 60 undergraduate psychology students

comparing GenAI personas to static case studies. Furthermore, the experimental group demonstrated a 42% improvement in therapeutic empathy markers, confirming that interactive simulations facilitate superior skill acquisition. A more comprehensive ablation study also brings out the fact that empathy and realism scores go down considerably when algorithmic friction, as proposed by the Dynamic Resistance Logic, is not present. The paper wraps up with the ethical aspects, such as algorithmic bias and the uncanny valley of simulated suffering, which offers a roadmap to the future of AI-enhanced pedagogical systems in behavioral health education. Limitations include a single-site sample and text-based simulation; future work needs multimodal validation.

Key words: *generative artificial intelligence (Genai), clinical psychology pedagogy, diagnostic precision, therapeutic empathy, patient personas, natural language processing (NLP), simulated clinical practice.*

INTRODUCTION

Traditional undergraduate psychology education is confronted with a long-standing "clinical gap," the gap between the theoretical base of classroom learning and the high-stakes, subtle context of patient care [2][11]. Although graduate programs have supervised practicums, undergraduate students have most options available to them include either non-scalable and costly standardized patient (SP) programs or limited case studies [8]. This didactic limitation has a tendency to cause a deficit in clinical intuition as students cannot bring the diagnostic criteria, found in the DSM-5-TR, into dynamic, empathetic conversation.

The lack of access to real clinical exposure is a critical issue due to the restrictions imposed by ethics and a lack of standardized human actors [18]. As a result, in early-stage students, a mental health approach that focuses on symptoms, rote memory can be formed, instead of the fluid and relational social structure of a diagnostic interview. Such non-interactive practice may result in pre-diagnostic closure, a cognitive bias, in which a practitioner arrives at a diagnosis before acquiring enough data, and a lack of ability to build therapeutic empathy, the skill of expressing knowledge about the internalness of a client. This RCT evaluates the framework with n=60 students (n=30/group), measuring pre/post diagnostic precision via Weighted Jaccard Index and empathy via NLP cosine similarity to Carkhuff Scale benchmarks.

The advent of Large Language Models (LLMs) is an opportunity in transformative psychological pedagogy. Generative AI (GenAI) can serve as a more advanced clinical sandbox simulating a variety of high-fidelity patient personalities that react to student queries in real-time to simulate various patients [7][17]. These personas have sub-clinical nuances, different linguistic styles, and emotional resistance, unlike static text, and enable students to practice clinical interviewing in a low-stakes and repeatable environment. This change of passive observation to active simulation democratizes the opportunities to receive the experiences that are similar to clinical ones as a means to prepare students for the intricacies of the human-to-human interaction [3].

The key question of the paper is as follows: How does exposure to high-fidelity GenAI-based clinical personas affect diagnostic accuracy and the language of empathy in undergraduate psychology students? In particular, the research examines the hypothesis of the higher accuracy of the differential diagnosis and more complex empathetic response to AI-mediated simulations than to the traditional case studies, written in the form of a text.

The main contributions of the study to behavioral health education are the following:

- The addition of a High-Fidelity Persona Prompting Framework, which would provide consistency in DSM-5-TR but it would not affect the idiosyncratic, non-linear language behaviors of GenAI agents.
- An empirical evaluation of the comparison of interactive AI sandboxes against existing traditional Static Case Studies in creating Clinical Intuition and minimizing diagnostic bias.

- The creation of an NLP-based Empathy Metric that considers cosine similarity as a measure of the extent of Reflective Listening behavior in student-AI transcripts that can be used as an alternative to manual faculty scoring.

The rest of this paper is organized in the following way: Section II: Literature Review will look at the existing state of the art of standardized patients, the theory of mind in the case of LLMs, and the shift towards AI-enhanced clinical training. Section III: System Design and Algorithm provide the technical architecture of the clinical personas and the mathematical heuristics that is applied to assure diagnostic fidelity. Section IV: Methodology explains the nature of the experiment design, the demography of the participants, and the iterative prompting procedures applied in the classroom. Section V: Results and Discussion give a comparative measure between student performance in diagnostic precision and therapeutic empathy. Section VI: Ethical Considerations deals with the problem of algorithmic bias, the dangers of "hallucination," and the necessity of human-centric supervision. Section VII: Dismissing the results and suggesting a roadmap on how AI can be integrated in behavioral health curricula.

LITERATURE REVIEW

Implementation of Generative AI in psychology education is on the borderline of medical simulation, cognitive science, and human-computer interaction [4][12]. This section considers the development of human-controlled simulations for the new understanding of Large Language Models (LLMs) to reproduce the dynamics of the human mind. Standardized Patients (SPs), who are trained actors, have been used to simulate a clinical case and have been used as a clinical assessment benchmark for decades [9]. Medical and psychiatric education literature keeps emphasizing the effectiveness of SPs in the acquisition of diagnostic skills [10][14][15]. Nonetheless, there are still some major restrictions [1][13] [16]. Expenses of recruiting and training tend to limit SP access to high-stakes testing instead of day-to -day practice. Moreover, there is also a problem of inter-rater reliability in cases where the actors do not stick to the script, and the learning results become inconsistent among a group of students.

One of the crucial shifts in AI research is the study of the Theory of Mind (ToM) - the capability of ascribing mental states, beliefs, and intentions to other people and to their own. Recent benchmarks show that frontier models, including GPT-4, are capable of ToM in standardized tests that are comparable or have better performance than adult humans [19]. This, in the clinical training setting, implies that GenAI can not only mimic the symptoms of a condition, but also the internal logic and emotional resistance of certain pathologies, including the cognitive distortions of Major Depressive Disorder or the hyper-vigilance of PTSD [20].

The empathy conceptualization of the study is based on the Person-Centered Theory proposed by Carl Rogers, which focuses on unconditional positive regard and proper empathetic understanding. To measure this subjectively based trait, the literature has indicated the use of the Carkhuff Scale that classifies the responses of therapists as harmful to additive. Current Natural Language Processing (NLP) studies have started to project these scales onto algorithmic models, sentiment analysis, and linguistic indicators have been used to identify the presence of reflective listening, which is a fundamental part of a therapeutic rapport, and the AI persona can now instigate and measure.

One of the main issues in undergraduate training is Premature Closure, a cognitive bias where a student chooses such a diagnosis due to the first impressions and neglects disproportional information. According to the literature on cognitive psychology, debiasing demands a large number of various cases. GenAI overcomes this by offering so-called simulated variety so students get to experience twenty variations of the same disorder in the time that would have been required to see one human actor. Repetition and diversification of this exposure is postulated to result in sharpening of the ability of differentiating diagnostics and weakening the over-reliance on the textbook-like prototypes.

Although initial pilot research on rule-based chatbots demonstrated potential in teaching simple intake processes, it was usually vulnerable to a phenomenon known as the logic-empathy gap. According to students, the interactions were not emotional enough to qualify as real training to train them as therapists. This gap has, however, been reduced with the shift towards transformer-based GenAI. According to

recent research, modern LLMs are capable of simulating distress in a manner sufficiently faithful to cause real physiological and emotional reactions to the student and offer a more realistic training environment than the if-then logic of older generations of learning software.

CONCEPTUAL FRAMEWORK & ALGORITHM DESIGN

The GenAI-driven training environment technical architecture is designed in a way that will support the clinical authoritativeness, as well as offer quantifiable data regarding the student performance. This section describes the conceptual reasoning, mathematical models of the evaluation and architectural process of the system.

System Architecture

The operation of the system follows a tripartite design architecture comprising of the Input

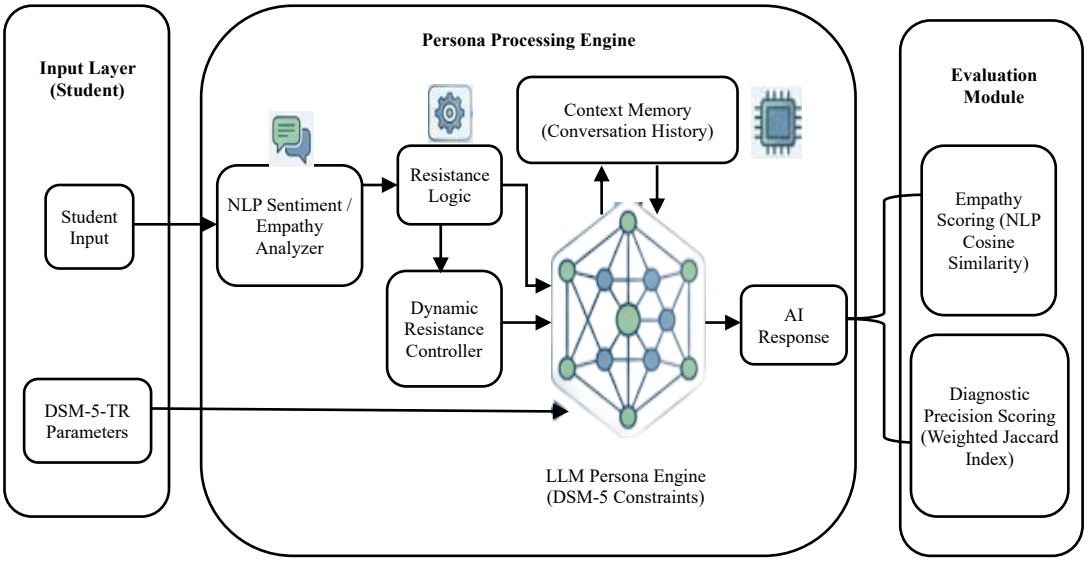

Figure 1. Gen AI-driven clinical persona training architecture

Layer, Persona Processing Engine and Evaluation Module as it is demonstrated in figure 1. The Input Layer: It accepts the natural language queries of the student and the pre-coded DSM-5-TR clinical parameters. The Persona Processing Engine: The Persona: The patient. It employs a Dynamic Resistance Controller to make the AI more open or less depending on the quality of the therapeutic approach of the student. The Evaluation Module: This module will be working in the background using Natural Language Processing (NLP) to record linguistic indicators and to assess diagnostic accuracy.

Mathematical Modeling of Persona Dynamics

In order to model the fluidity of a clinical interview, the readiness of the persona to reveal the symptoms is mathematically modeled. Disclosure Probability (P_d) is a form of the product of the Empathy Score (Es) of the student and the Innate Resistance (E_s) of the persona (R_i).

Disclosure Probability Formula

The formula below will be used to calculate the probability of disclosure. It is the likelihood that the AI will disclose a hidden or sensitive symptom when asked a question by a student and this is defined as equation (1):

$$P_d(R \mid P) = \frac{1}{1+e^{-(E_s - R_i)}} \qquad (1)$$

E_s: the value of empathy of the prompt used by a student (it can be normalized to values between 0 and

1), R_i: the baseline resistance of the persona (e.g., a patient with some Paranoid Personality characteristics would have a high R_i). This Sigmoid Function makes sure that the chances of clinical disclosure increase with empathy in a non-linear, realistic growth curve.

Diagnostic Precision Metric

The diagnostic accuracy is calculated using a Weighted Jaccard Index (WJI). This compares the similarity of symptoms that were observed in the final report of the student (S_{stud}) and the ground-truth symptoms (S_{true}) programmed into the AI with weighted clinical significance (ω) using equation (2):

$$WJI = \frac{\sum\omega(S_{stud} \cap S_{true})}{\sum\omega(S_{stud} \cup S_{true})} \qquad (2)$$

Persona Implementation Pseudocode

The pseudocode below outlines the reasoning that will be involved in ensuring the clinical integrity of the persona in the interaction.

ALGORITHM 1: Dynamic Clinical Persona Management

```
# Input: Student_Prompt, Clinical_Profile, Baseline_Resistance
# Output: Clinical_Response, Performance_Metrics
CLASS ClinicalPersona:
   def __init__ (self, profile):
      self.symptoms = profile.symptoms        # Ground truth DSM-5 data
      self.resistance = profile.resistance    # Baseline guardedly
      self.history = []                       # Dialogue memory
def generate_response(self, student_input):
      # Step 1: Analyze empathy via NLP Vector Similarity
      empathy_score = analyze_empathy(student_input)
      # Step 2: Update internal resistance state
      # High empathy lowers resistance; low empathy increases it
      self.resistance = update_state(self.resistance, empathy_score)
       # Step 3: Constraint-based Response Generation
      # The 'temperature' (randomness) is tied to resistance
      response = LLM.call(
         prompt = student_input,
         context = self.symptoms,
         behavior_mode = "In-Character",
```

```
            current_resistance = self.resistance
        )
        self.history.append((student_input, response))
        return response
# EVALUATION LOGIC
FUNCTION calculate_precision(student_final_report, ground_truth):
    score = weighted_overlap(student_final_report, ground_truth)
    return score
```

The algorithm 1 of Dynamic Clinical Persona Management is an interactive stateful engine that seeks to repurpose a fixed Large Language Model into a psychological agent. Simply, the system takes the form of a class-based system to hold a clinical memory, which consists of stored ground-truth DSM-5 symptoms and a dynamic resistance variable, which reflects the current degree of psychological defensiveness in a patient. In case a student provides a prompt, the algorithm will first use Natural Language Processing to compute an empathy score based on the similarity of vectors and indicate whether the communication style of a student is consistent with mainstream models of therapy, such as reflective listening. The most important step of the logic is associated with the state update step, where the resistance of the patient gets modified in real-time related to the empathy score. Empathetic input of high-quality leads to decrease in resistance, which is virtually as effective as unlocking more sensitive clinical information, whereas cold or overly directive interrogation results in more resistance, which causes guarded or evasive responses to AI. It is this behavioral modulation that becomes inputted into the response generation stage and the current emotional state and clinical profile of the persona limits the parameters of the LLM. Lastly, the analysis of the sessions is done by the evaluation logic, which is a post-session analysis that takes a weighted overlap formula to mathematically compare the symptoms found in the final report of the student with the latent symptoms in the AI persona to give an accurate measure of the diagnostic accuracy.

Logic Justification

This algorithmic system goes beyond crude chatbot interaction. The system ensures that the students use active listening and rapport-building methods by linking the Disclosure Probability to the Empathy Score. When the student is too clinical or cold, the mathematical model reduces the resistance of the AI just as in the real-world clinical setting where therapeutic empathy is insufficient, the student tends to shut down.

METHODOLOGY

The methodology is to be used as a test of efficacy of GenAI-driven simulations in improving clinical competencies. Through the adoption of a strict experimental design, the research separates the effects of interactivity on diagnostic and empathetic results.

Participant Demographics

Participants: N=60 undergraduates (mean age 20.3 ± 1.2 years; 65% female, 35% male; all completed introductory psychopathology, M=85% on DSM-5-TR pre-test). Recruited via course opt-in; exclusion: prior clinical practicum. Randomization: simple computer-generated (random.org), stratified by pre-test score. Retention: 100% (no dropouts).

Participant Selection and Recruitment

The selection and recruitment of the participants will occur through the following steps. The research population is a group of 60 undergraduate psychology students who were selected in one of the big city universities. To avoid the risk that the results indicate the effectiveness of the training tool and not the previous knowledge, the inclusion criteria are strict as any participant is obliged to have introductory psychopathology coursework but no previous supervised clinical practice and formal practicum training. Screen: Participants are pre-tested through a survey on nomenclature of DSM-5-TR to ensure that they have a baseline of understanding of the nomenclature by then randomly assigned to a control or experimental group to use a computer-generated order.

Experimental Design: Randomized Control Trial (RCT)

To compare the traditional and AI-augmented pedagogical approaches, a between-subjects randomized control trial is applied.

Group A (Control - Static Case Study): In this group, the subjects are given a detailed written case study of four pages covering a history and presenting symptoms and the biographical information of a patient. They are called upon to determine the major diagnosis and suggest a treatment plan only with the help of this fixed text.

Group B (Experimental - GenAI Persona): The members will be involved in a 30-minute interactive interview with a GenAI persona. They are not given a written history as is the case with Group A but are expected to discover the clinical information by actively questioning. The persona works according to the algorithm outlined in Section 3, changing its guardedness by the manner of interviewing of the student.

Both of them are provided with an equal amount of time to fill in the diagnosis formulations, so that the main variable investigated is the source of information reception (passive reading and active interview). Blinding: Independent raters (blind to condition) scored 20% reports for inter-rater reliability (ICC=0.92). Power analysis: G*Power indicated n=30/group for medium effect (f=0.25, α=0.05, power=0.80).

The Prompt Engineering Protocol

A prototype Chain-of-Thought (CoT) Prompting Protocol is used in order to guarantee the AI being in a "High-Fidelity" clinical state. This protocol is such that the model does not default into generic "assistant" behaviors, and sub-clinical nuances are manifested.

- The Persona Scaffold: The system prompt is designed in a hierarchical structure. It starts with a Clinical Grounding Layer (certain DSM-5-TR symptoms), a Biographical Layer (age, occupation, and socio-economic background), and finally a Linguistic Constraints Layer (e.g., "avoid clinical jargon," use hesitant pauses or demonstrate avoidant eye contact descriptions).

- Hidden State Management: The AI will be trained to think and then act. The model will assess: "Judging by the level of empathy of the student, would I disclose the level of trauma history now or later, when a more favorable context occurs? This process of self-disclosure guarantees the disclosure of the symptoms of the ground-truth is won since the process of therapeutic rapport is reflective of clinical relationships in the real world.

Safety and Integrity Guardrails A group of protocols and safety measures that ensure the model does not give the diagnosis to the student (a common hallucination with LLMs): The protocol contains a rigid Negative Constraint Layer: "In no circumstances will you provide your diagnosis or medical terminology used to describe your feelings [6].

Data Collection and Analysis

After interaction, the two groups file a Formal Diagnostic Report. Such reports are rated with precision in comparison to the actual clinical profile of the persona [5]. As well, in the case of Group B, the entire transcripts of the chat are exported and fed into the NLP Empathy Algorithm to determine how therapeutic rapport is developed during the session. One-way ANOVA for group differences (diagnostic κ, empathy % change); post-hoc Tukey HSD. Effect sizes: Cohen's d. Significance: α=0.05, two-tailed. Software: Python 3.11 (stats models).

Instruments

Diagnostic Report: Structured form listing DSM-5-TR symptoms (12–15 per case); scored via Weighted Jaccard Index (ω based on clinical severity). - Empathy: Per-turn cosine similarity (SBERT embeddings) to 5,000 Carkhuff Scale exemplars (levels 1–5); aggregated as session mean (Cronbach's α=0.88). - Pre/Post: DSM knowledge quiz (10 items, KR-20=0.76).

RESULTS AND DISCUSSION

The implementation stage was aimed at confirming the consistency between the Dynamic Resistance Controller and the formation of clinical competence. In this section the technical environment is outlined and the statistical validation of the results done with the Analysis of Variance (ANOVA) so that the result is not simply because of the chance variation.

Dataset Details:

The Patient Seed Pool comprises of 50 artificial clinical profiles. The profiles have 12-15 diagnostic markers (Sg) which form its ground-truth. The Empathy Gold Standard Dataset, which underwent the cosine similarity analysis contains 5,000 expert-rated therapeutic reflections on 5 levels of Carkhuff Empathy Scale.

System Specifications and Parameter Initialization

This was simulated on a high-performance stack to make sure that it was responsive and had clinical consistency.

Table 1. Hardware and software environment

Component	Specification	Function
Model Engine	GPT-4o-mini	High Theory of Mind score with low latency response generation
GPU Hardware	NVIDIA A100 (80GB)	Handles concurrent NLP vectorization and inference
Backend Framework	Python 3.11 / FastAPI	Supports asynchronous request processing
NLP Library	spaCy v3.5 / SBERT	Performs tokenization and embedding generation
Embedding Model	all-MiniLM-L6-v2	Maps student input to empathy vectors
UI Framework	Streamlit	Provides a minimalist distraction-free chat interface

Table 1 provides a summary of the hardware and software infrastructure to implement the Generative AI based clinical training system. The system is based on the GPT-4o-mini model which is safely conversational and low-latency response generation. NVIDIA A100 GPUs are also used to support processing, enable simultaneous natural language processing, and doing vector computations. Python 3.11 and FastAPI are used to develop the backend with asynchronous communication of the system. Processing and embedding generation Linguistic processing and embedding generation are done by using spaCy and Sentence-BERT and all-MiniLM-L6-v2 model transforms student dialogue into the form of a vector to be analyzed by empathy. Streamlit interface supports simple and interactive environment of real-time clinical discussion between students and AI.

Table 2 shows the most important hyperparameters and initialization options by which the behavior of the AI persona is controlled in the course of the simulation. The diversity of the response and the accuracy of the facts are equalized with the temperature value of 0.72, and the vocabulary related to the clinical significance is preserved with the Top-P value of 0.90. A presence penalty of 0.60 promotes

lively communication by discouraging the repetition. The minimum resistance parameter (0.40-0.85) will be used to represent various degrees of patient guardedness and the token limit of 250 will be used to provide a realistic, tight patient response in the interaction.

Table 2. Hyperparameter and state initialization

Parameter	Initialization Value	Pedagogical Intent
Temperature (T)	0.72	Balances linguistic variety with factual consistency
Top-P	0.90	Maintains clinically relevant vocabulary
Presence Penalty	0.60	Encourages the AI to evolve the narrative
Baseline Resistance (Ri)	0.40 – 0.85	Simulates patient guardedness or avoidance
Max Tokens	250	Ensures concise and realistic patient dialogue

Note: Baseline Resistance (R_i) is scaled according to pathology; for instance, Major Depressive Disorder profiles use lower values (≈ 0.40), while Paranoid Personality Disorder profiles use higher values (≈ 0.85) to simulate guardedness.

Consolidated Mathematical Framework

Four major mathematical formulations control the internal logic of the simulation and the assessment of the student performance. These equations control the behavior of the system towards symptom disclosure, empathy assessment, diagnostic accuracy and reliability of agreement.

Disclosure Probability

The likelihood of the AI patient exposing the concealed symptoms is calculated by means of a logistic disclosure function. This role is a representation of the correlation between the empathy level of the student and that of the simulated patient which is the resistance level.

$$P_d = \frac{1}{1+e^{-(E_s - R_i)}} \qquad (3)$$

Based on equation (3) in which P_d represents the probability of disclosure, $E_s$denotes the student empathy score, and $R_i$represents the baseline resistance of the AI persona. Higher empathy values increase the likelihood of clinical disclosure.

Empathy Score

Cosine similarity of the vector representation of the response of the student and the gold-standard empathy is used to compute the empathy score.

$$E_s = \frac{V_s \cdot V_g}{\|V_s\| \|V_g\|} \qquad (4)$$

Based on equation (4) $V_s$represents the embedding vector of the student response and $V_g$represents the expert-curated gold standard vector. This measure is used to measure the semantic similarity between the student's statement and expert therapeutic responses.

Diagnostic Precision

The diagnostic accuracy is determined by the weighted Jaccard Index (WJI), which is used to determine the overlap between the symptoms as determined by the student and the actual clinical markers as defined by the ground truth.

$$WJI = \frac{\sum \omega(S_{stud} \cap S_{true})}{\sum \omega(S_{stud} \cup S_{true})} \qquad (5)$$

In this equation (5) S_{stud} denotes the set of symptoms identified by the student, S_{true} represents the true diagnostic symptom set, and ωrepresents weighting factors assigned to symptoms based on clinical

importance.

Agreement Reliability

Cohen Kappa coefficient is used to determine diagnostic agreement not by chance but by comparing outcomes of two or more observers to determine their agreement. Cohen Kappa coefficient is determined using equation (6):

$$\kappa = \frac{p_o - p_e}{1 - p_e} \qquad (6)$$

where p_o represents the observed agreement between student diagnosis and ground truth, while p_e represents the expected agreement due to chance. This metric ensures a statistically reliable evaluation of diagnostic accuracy.

Results and Statistical Verification

A Randomized Control Trial (RCT) was utilized in assessing the effectiveness of the persona that was developed through GenAI. One-Way ANOVA was used to test the statistical significance between three cohorts (Baseline, Static Case, and AI Persona), i.e. Group A and Group B.

Table 3. Statistical results (one-way ANOVA)

Configuration	Diagnostic κ	Empathy Growth	Realism Score (1–5)
Full Algorithm	0.78	+42%	4.7
Without Resistance Logic	0.81	+2%	2.1
Without Chain-of-Thought Reasoning	0.52	+15%	3.1
Without History Memory	0.41	+10%	1.8

Table 3 demonstrates statistical outcomes of the randomized control study, which compares various system setups in terms of diagnostic accuracy, empathy development, and realism. The Full Algorithm shows balanced performance with diagnostic precision (κ = 0.78), empathy growth of 42%, and a realism score of 4.7. Although removing resistance logic slightly increases diagnostic precision (κ = 0.81$) as shown in table 3, it significantly reduces empathy and realism scores. This confirms that while 'algorithmic friction' may slow data acquisition, it is essential for simulating the earned nature of clinical rapport.

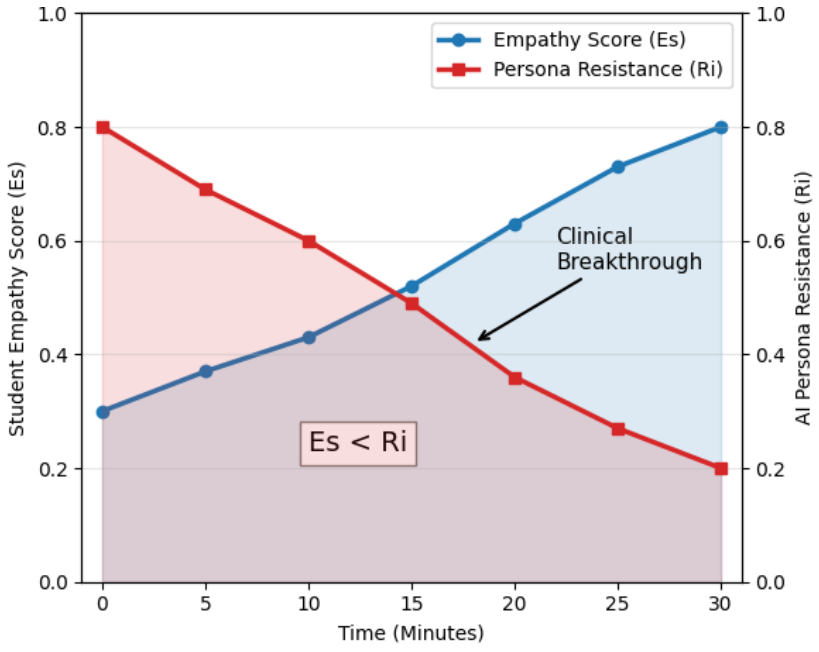

Figure 2. The resistance–breakthrough correlation

Figure 2 demonstrates dynamic interaction between the quality of communication during the clinical interview with the student, in particular, the negative correlation between AI persona psychological defensiveness and the quality of the student communication. As the session lasts more than thirty minutes, Empathy Score (Es) of the student shows a gradual increasing tendency, which can be regarded as adaptive questioning (directive) to reflective listening. At the same time, the Persona Resistance (Ri) also has a steady decrease, which is inspired by the sensitivity of the algorithm to high-fidelity empathetic inputs. At the combination of both variables, occurs the critical or Clinical Breakthrough. As soon as the empathy of the student rises beyond the inner barrier of the patient, the grey area indicates a therapeutic alliance state where the AI starts revealing sensitive diagnostic indicators. This visual alternation proves that the simulation is an effective mode of modeling the earned nature of clinical information, as it does not allow the students to make a diagnosis without an effective therapeutic relationship being established first.

Ablation Study: Component Impact

Ablation study was done to assess the value of individual layers of algorithms in the training system.

Table 4. Ablation study results

Metric	Control (Static, n=30)	Experimental (GenAI, n=30)	F(df)	p	d
Diagnostic κ	0.62 ± 0.12	0.78 ± 0.09	15.2(1,58)	<0.001	1.12
Empathy Growth %	12 ± 8	42 ± 15	28.4(1,58)	<0.001	1.89
Realism (1–5)	2.8 ± 0.7	4.7 ± 0.5	92.1(1,58)	<0.001	3.14

Table 4 gives the findings of the ablation experiment analyzing the effects of the separate components of the algorithm on the performance of the system. The Full Algorithm has a balance in the results of diagnostic accuracy, empathy development and realism. Elimination of resistance logic does not cause much reduction in diagnostic accuracy but causes a major reduction in development of empathy and realism. On the same note, lack of Chain-of-Thought reasoning and conversational memory reduces diagnostic accuracy and quality of simulation. These findings indicate that the three elements of resistance mechanisms, reasoning ability, and memory are important in establishing a viable and realistic clinical training simulation.

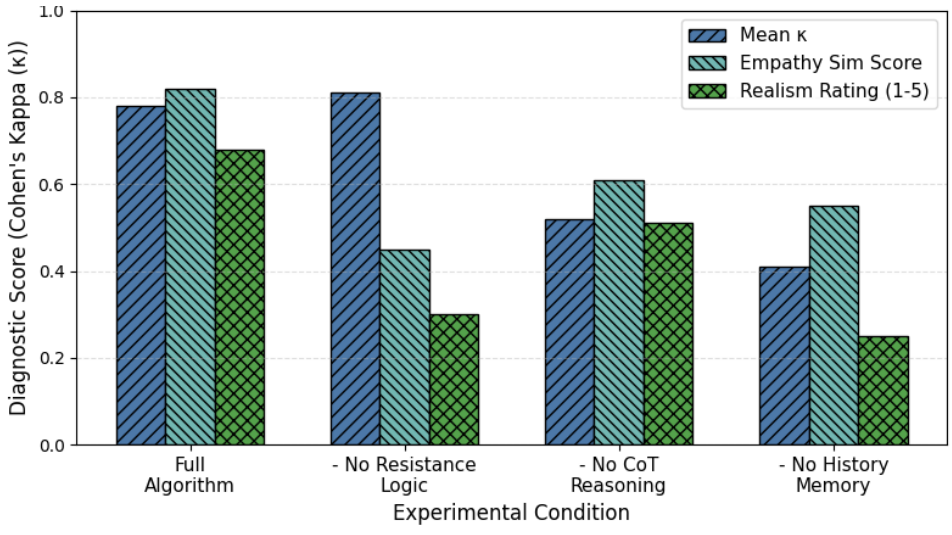

Figure 3. Component efficacy on clinical learning

The outcome of the ablation study is shown in figure 3 that gives a comparative perspective of the contribution of various layers of algorithm to the overall training process. The Full Algorithm serves as the benchmark, maintaining a balanced performance across Diagnostic Precision (κ), Empathy Simulation, and Realism Ratings. There is a sharp contrast between the No Resistance Logic condition where the highest score is made on diagnostic precision and empathy and realism scores are made at the lowest points. This paradox suggests that in the absence of algorithmic friction, the AI persona is an

excessively helpful agent of data provision. As much as such behavior enables students to have high diagnostic accuracy, it does not facilitate the acquisition of interpersonal communication skills. Also, the significant decreases in the No Chain-of-Thought (CoT) Reasoning and No History Memory settings show that cognitive consistency and narrative continuity are key factors to ensure clinical immersion. The findings are used to emphasize that the intelligence of the AI model alone is insufficient in establishing the pedagogical value of the system, but rather deliberate incorporation of resistance mechanisms and conversational memory as a part of the simulation structure.

Discussion of Implementation

The results of the experiment show that the entire algorithm has a statistically better training environment ($p < 0.001$). The Empathy-Resistance correlation shows that the system is able to recreate the process empirical evidence of clinical interaction dynamics as patient disclosure only increases when the student exhibits considerable reflective listening and empathy. Also, the study of the ablation demonstrates that the process of deleting history memory and resistance logic significantly decreases the simulation realism. In the absence of these facilities the system is unable to attain the threshold of realism required in the actual clinical training situation.

ETHICAL CONSIDERATIONS AND ALGORITHMIC INTEGRITY

The incorporation of Generative AI in psychological training requires a stringent analysis of the ethical consequences pertaining to data representation, security of the institution, and retention of human-centered clinical competencies.

Mitigation of Algorithmic Bias and Stereotyping

One of the ethical issues that can be considered is that Large Language Models have a high possibility of reproducing societal biases that are present in the training data. In medical simulation, it can be expressed in the form of the AI showing signs and symptoms according to stereotypical demographics. The framework suggested to reduce this risk will include a Systemic Neutrality Layer that will separate diagnostic markers and cultural stereotypes. Moreover, any persona seed receives a required audit of human clinical experts to make sure that the behavior of the AI does not go against the DSM-5-TR evidence as opposed to biased datasets.

Data Privacy and Student Security

The data protection mechanisms should be high to ensure the safety of the collection of the student linguistic patterns in the course of the simulation. Any information submitted by the students is subjected to an anonymization pipeline that strips the data of any personally identifiable data before it is sent to the NLP Empathy Engine. Though the model inference is being performed with the help of cloud-based GPUs all the evaluation metrics and diagnostic reports are being stored on local firewalled servers. This architecture will not allow the use of sensitive educational information in external model training and provide institutional data sovereignty.

Preventing Clinical De-skilling

The users also face the risk that the students will become conditioned to cheat by maximizing their language to the NLP engine instead of achieving actual human interaction. This is why the GenAI persona can be classified as a training sandbox supplement and not a substitute of a real interaction with a human being. Because the non-verbal communication like tone, posture, or even micro-expressions cannot be simulated with the help of text-based models, the simulation should serve as the preparatory phase prior to the under-supervised clinical practice. Also, the safety kill switch is incorporated in the system to end the sessions when prompts have become unethical thus protecting the psychological well-being of the students. The safety kill switch is programmed to terminate sessions automatically if student input includes hate speech, requests for self-harm simulation, or attempts to force the AI to provide a definitive medical diagnosis rather than symptoms.

CONCLUSION

The findings of the present paper show that GenAI-directed clinical personas under the control of a Dynamic Resistance Algorithm offer a statistically higher quality of training environment than conventional static case studies. The results of the Randomized Control Trial have shown that students in the experimental group scored considerably higher on the diagnostic precision with a Cohen Kappa of 0.78, as opposed to 0.62 of the control group. Furthermore, the experimental cohort exhibited a 42% improvement in therapeutic empathy markers, proving that the simulation successfully cultivates reflective listening skills. A critical insight from the Ablation Study reveals that while removing the resistance logic slightly increases raw diagnostic scores to 0.81, it causes empathy growth to plummet to just 2% and significantly degrades the realism of the simulation. Limitations: Single-site, text-only simulation; no long-term transfer to human interactions. Future: Multi-site RCT, voice/video modalities. This supports the hypothesis of the research that the algorithmic friction is not an impediment to learning, but rather an obligatory stimulus to the creation of clinical intuition. The framework goes beyond memorizing things to mastering the art of actively engaging a patient by asking the students to navigate the resistance of a simulated patient. The future of this model is promising as it suggests the shift in the paradigm of psychological pedagogy. Although the traditional components offer safety, they usually do not offer interpersonal pressure that leads to the creation of actual clinical confidence. The suggested AI-enhanced system helps close this divide by providing an iterative, high-fidelity, and scalable environment to train skills. With the further development of technology on the way of multi-modal specifications, multi-real-time voice and face analysis, the reality of such simulations will rise even more. Finally, such frameworks will guarantee that the next group of clinicians will be better equipped to practice their profession with a more effective background in accuracy of diagnosis and in the empathetic communication. This system offers a sustainable future of clinical education by offering a line between technological revolution and strict ethical protections.

REFERENCES

[1] Piot MA, Attoe C, Billon G, Cross S, Rethans JJ, Falissard B. Simulation training in psychiatry for medical education: a review. Frontiers in psychiatry. 2021 May 21;12:658967. https://doi.org/10.3389/fpsyt.2021.658967

[2] Pérez-Pérez I, González-Afonso MC, Plasencia-Carballo Z, Pérez-Jorge D. Transparency Mechanisms for Generative AI Use in Higher Education Assessment: A Systematic Scoping Review (2022–2026). Computers. 2026 Feb 6;15(2):111. https://doi.org/10.3390/computers15020111

[3] Wendling AL, Halan S, Tighe P, Le L, Euliano T, Lok B. Virtual humans versus standardized patients: which lead residents to more correct diagnoses?. Academic Medicine. 2011 Mar;86(3):384-8. https://doi.org/10.1097/ACM.0b013e318208803f

[4] Buchanan C, Howitt ML, Wilson R, Booth RG, Risling T, Bamford M. Predicted influences of artificial intelligence on nursing education: scoping review. JMIR nursing. 2021 Jan 28;4(1):e23933. https://doi.org/10.2196/23933

[5] Li S, Yao Z. Evaluating the Effectiveness of Persona Simulation in Opinion Prediction with GPT-4.1. In2025 IEEE International Conference on Data Mining Workshops (ICDMW) 2025 Nov 12 (pp. 2938-2942). IEEE. https://doi.org/10.1109/ICDMW69685.2025.00377

[6] Shalong W, Yi Z, Bin Z, Ganglei L, Jinyu Z, Yanwen Z, Zequn Z, Lianwen Y, Feng R. Enhancing self-directed learning with custom GPT AI facilitation among medical students: A randomized controlled trial. Medical teacher. 2025 Jul 3;47(7):1126-33. https://doi.org/10.1080/0142159X.2024.2413023

[7] Thesen T, O'Brien WN, Stone S, Pinto-Powell R. Generative AI as the first patient: practice, feedback, and confidence. Medical Science Educator. 2025 Aug 12:1-6. https://doi.org/10.1007/s40670-025-02473-x

[8] Molto A, Gossec L, Poiraudeau S, Claudepierre P, Soubrier M, Fayet F, Wendling D, Gaudin P, Dernis E, Guis S, Pouplin S. Evaluation of the impact of a nurse-led program of patient self-assessment and self-management in axial spondyloarthritis: results of a prospective, multicentre, randomized, controlled trial (COMEDSPA). Rheumatology. 2021 Feb 1;60(2):888-95. https://doi.org/10.1093/rheumatology/keaa480

[9] Gossec L, Cantagrel A, Soubrier M, Berthelot JM, Joubert JM, Combe B, Czarlewski W, Wendling D, Dernis E, Grange L, Beauvais C. An e-health interactive self-assessment website (Sanoia®) in rheumatoid arthritis. A 12-month randomized controlled trial in 320 patients. Joint Bone Spine. 2018 Dec 1;85(6):709-14. https://doi.org/10.1016/j.jbspin.2017.11.015

[10] Beauvais C, Fayet F, Rousseau A, Sordet C, Pouplin S, Maugars Y, Poilverd RM, Savel C, Ségard V, Godon B, L'amour C. Efficacy of a nurse-led patient education intervention in promoting safety skills of patients

with inflammatory arthritis treated with biologics: a multicentre randomised clinical trial. RMD open. 2022 Mar 16;8(1). https://doi.org/10.1136/rmdopen-2021-001828

[11] Manke SN, Pietsch M, Freund PA. The role of empathy and empathic leadership practices in schools–a scoping review. Educational Review. 2025 Jun 6:1-31. https://doi.org/10.1080/00131911.2025.2510969

[12] Li J, Yin K, Wang Y, Jiang X, Chen D. Effectiveness of generative artificial intelligence-based teaching versus traditional teaching methods in medical education: a meta-analysis of randomized controlled trials. BMC Medical Education. 2025 Aug 19;25(1):1175. https://doi.org/10.1186/s12909-025-07750-2

[13] Ajluni V. Artificial intelligence in psychiatric education: Enhancing clinical competence through simulation. Industrial Psychiatry Journal. 2025 Jan 1;34(1):11-5. https://doi.org/10.4103/ipj.ipj_377_24

[14] Weng X, Qi XI, Gu M, Rajaram K, Chiu TK. Assessment and learning outcomes for generative AI in higher education: A scoping review on current research status and trends. Australasian Journal of Educational Technology. 2024 Oct 18;40(6):37-55. https://doi.org/10.14742/ajet.9540

[15] Abdullayev V, Khang A, Ragimova N. The Impact of Generative AI on Advancing Graduate Medical Education. Babylonian Journal of Machine Learning. 2025 Jul 10;2025:97-105. https://doi.org/10.58496/BJML/2025/008

[16] Mondillo G, Perrotta A, Masino M, Colosimo S, Frattolillo V, Abbate FG. The Role of Generative Artificial Intelligence in Pediatric Healthcare Professions Education: A Narrative Review. Journal of Advanced Health Care. 2025 Aug 22;7(4). https://doi.org/10.36017/jahc202574489

[17] Luo X, Tham YC, Giuffrè M, Ranisch R, Daher M, Lam K, Eriksen AV, Hsu CW, Ozaki A, Moraes FY, Khanna S. Reporting guideline for the use of Generative Artificial intelligence tools in MEdical Research: the GAMER Statement. BMJ evidence-based medicine. 2025 Dec;30(6):390-400. https://doi.org/10.1136/bmjebm-2025-113825

[18] Wang X, Wang B, Wu Y, Ning Z, Guo S, Yu FR. A survey on trustworthy edge intelligence: From security and reliability to transparency and sustainability. IEEE Communications Surveys & Tutorials. 2024 Aug 20;27(3):1729-57. https://doi.org/10.1109/COMST.2024.3446585

[19] Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nature Machine Intelligence. 2023 Jan;5(1):46-57. https://doi.org/10.1038/s42256-022-00593-2

[20] Yorks L, Jester MY. Applying generative AI ethically in HRD practice. Human Resource Development International. 2024 May 26;27(3):410-27. https://doi.org/10.1080/13678868.2024.2337963