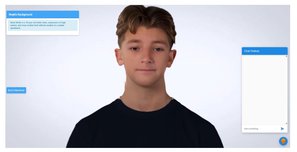
Bloch-Elkouby Lab for Suicide Prevention and Psychotherapy Research

Ferkauf
Ferkauf Graduate School of Psychology

Acceptability and Feasibility of Virtual Patient with and without ChatGPT



Jimmy Pengyu Chen[1,2], Anna Stefania Prekas[1,2,3], Hanjiang Xu[1,2,4], Yongxue Wei[1,2,4], Lilian Guo[1,2,4], Madison Wakely[1,2], Sarah Bloch-Elkouby[1,2]

1. Icahn School of Medicine at Mount Sinai; 2. Ferkauf Graduate School of Psychology, Yeshiva University; 3. Adelphi University; 4. Teachers College, Columbia University

Background

The evolution of chatbot engines has created an unprecedented opportunity to develop virtual patients (VPs) capable of **engaging in free-flowing conversations with clinicians in training** while **maintaining fidelity** to their original clinical design. Over the past 60 years, chatbot systems have progressed from rule-based pattern-matching models to intent-recognition and dialogue-management frameworks, and more recently to large-scale neural networks powering conversational AI (He et al., 2025). However, improvements in chatbot capability do not necessarily translate into greater acceptability, satisfaction, intention to use, or training effectiveness.

Several studies have established the feasibility of a VP based clinical training tool, even with limited conversational capabilities. Dupuy et al. (2020) presented a VP with a profile of major depressive disorder to medical students, who conducted a psychiatric interview by selecting on of the two pre-written sentences. Medical students reacted with a welcoming attitude and enjoyed the interaction. Bloch-Elkouby et al. (2025) tested the effectiveness of an emotional awareness training that feature virtual patient interactions powered by Google DialogFlow (DF; Google Could, n.d.), an intent matching system, and established that clinicians' who received the training built stronger bond and achieve greater reduction of suicide related symptoms with their clients.

As LLM capabilities increased substantially over the past few years, the excitement in technological advancement can lure researchers to attempt to implement VPs based on LLMs. However, there are a few cautions against using them: As the LLM are difficult to control, they might create unintended confounds to the research, which renders a VP based on intent matching system like DF much more attractive.

	Pros	Cons
LLM	Free, natural conversation	Hallucination
	Less development time if the model performs well	Low fidelity creates confounds in research
		Difficult to improve models
		Paraphrasing, not reproducing
DF	Pre-scripted responses, output based on matches	Prone to mismatch and no match
	High fidelity, suitable for research	Time consuming to develop

Our lab initially elected to develop a DF powered VP and sacrificed the natural conversation flow for high fidelity. Nevertheless, the DF powered VP have substantial shortcomings. In Bloch-Elkouby et al. (2026), observer-based ratings identified that roughly 15% of interactions in an average interview transcript between the clinician and the VP contains (one clinician question + one VP response) were disrupted by failures to match the clinician input with a pre-existing intent to produce an appropriate output, and that roughly 11% of interactions were disrupted by inappropriate responses, i.e., intent mismatches.

To address the technical disruptions and strike a balance between fidelity and conversational flow, our lab subsequently connected the DF system to ChatGPT, which is designed to produce responses only when there is an intent no match or mismatch.

The Current Study

The current study tested the two generations of VPs on clinical or counseling psychology doctoral students **(DF only: *n*=69; DF+ChatGPT: *n*=32)** and compared the acceptability and feasibility of each VP. Students were asked to conduct a diagnostic interview with the VP over Zoom and completed the post-VP assessments.

Hypothesis: The DF+ChatGPT VP will produce substantially less observer-based and subjective disruptions and be perceived as more enjoyable and valuable.

Method

Acceptability

- Enjoyment: 5-point Likert scale, self reported.
- Perceived value of the VP: 5-point Likert scale, self reported.

Feasibility

- Subjective reports of technical disruptions:
  - the number of inappropriate VP responses
  - the degree of disruption caused by language issues, 7-point Likert scale
  - the degree of disruption caused by technical issues, 7-point Likert scale
- Objective counts of disruptions during the interaction
  - No match
  - Inappropriate responses coded by human raters (DF only) and ChatGPT (DF+ChatGPT)

Results

Outcome	GPT+DF Mean	DF Mean	t	df	p
Total interactions	102.13	102.49	−.037	99	.971
Objective inappropriate	5.3750	11.7681	−4.130	99	**<.001**
Objective inappropriate (%)	4.6002	11.1021	−6.732	99	**<.001**
No-match	6.38	15.30	−3.853	99	**<.001**
No-match (%)	7.9685	14.5033	−3.090	99	**.003**
Subjective inappropriate	8.48	11.40	−2.346	91	**.021**
Language issue disruptiveness	3.53	4.14	−2.343	99	**.021**
Technical issue disruptiveness	2.17	1.94	.523	91	.602
Enjoyment	2.81	2.46	1.673	99	.097
Training tool value	2.84	2.67	.810	99	.420

Discussion and Additional Insights

As expected, integrating ChatGPT into the DF based VP was associated with improved conversational feasibility. Compared with DF alone, the hybrid system produced fewer objectively coded inappropriate and no-match responses, fewer participant-reported inappropriate responses, and less disruptive language issues, despite similar total interaction counts. However, the groups did not differ significantly in perceived technical disruption, enjoyment, or training value. Thus, the reduction in conversational disruptions was not accompanied by clear evidence of improved acceptability. This discrepancy between the acceptability and feasibility might be explained the threshold of the number of disruptions to create a good experience. As indicated in the results, technical disruptions still account for roughly 12-13% of all interactions. The statistical decrease of disruptions might not have resulted in the improvement of experience.

Additionally, researchers should use high caution when developing a conversational agent with involvement of any commercial LLM, which is known to update on the backend or limit the intellectual level of the same model after substantial usage from the same account without informing the users. The DF+GPT participants suffer significant fluctuations with the number of inappropriate responses and no matches, clearly indicated by the time they completed the interview, while the study could not identify any way to investigate why. Furthermore, as commercial models update, the original prompt and setup might no longer work well. Constant updates and examinations are required for these models to ensure integrity.

References

Bloch-Elkouby, S., Johnson, B., Chen, J. P., El Hayek, R., Jones, A., Eid, T., de Siqueira, A. G., Apter, Y., Yao, H., Lok, B., Foster, A., & Galynker, I. (2025). Can training clinicians in emotional awareness using virtual patient interactions improve the therapeutic alliance with psychiatric outpatients and reduce suicidal outcomes? *Psychiatry Research, 356*, 116846. https://doi.org/10.1016/j.psychres.2025.116846

Bloch-Elkouby, S., Johnson, B. N., Chen, J. P., Shi, R., Goncearenco, I., Khaikin, S., Wei, Y., Xu, H., Gomes De Siqueira, A., Yao, H., Lok, B., Cohen, L., Apter-Levi, Y., Shimon Raz, O., Williams, M. T., & Galynker, I. (2026). *Race, bias, and suicide risk assessment: An experimental study using a male adolescent virtual patient* [Manuscript submitted for publication]. Icahn School of Medicine at Mount Sinai.

Dupuy, L., Micoulaud-Franchi, J.-A., Cassoudesalle, H., Ballot, O., Dehail, P., Aouizerate, B., Cuny, E., De Sevin, E., & Philip, P. (2020). Evaluation of a virtual agent to train medical students conducting psychiatric interviews for diagnosing major depressive disorders. *Journal of Affective Disorders, 263*, 1–8. https://doi.org/10.1016/j.jad.2019.11.117

Google Cloud. (n.d.). *Dialogflow ES documentation*. Retrieved October 13, 2026, from https://docs.cloud.google.com/dialogflow/es/docs

He, R., Cao, J., & Tan, T. (2025). Generative artificial intelligence: A historical perspective. *National Science Review, 12*(5), Article nwaf050. https://doi.org/10.1093/nsr/nwaf050