The Relationship Between a Decrease in Technical Disruptions and an Increase in Therapeutic Alliance

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

1. Teachers College, Columbia University, New York, NY, USA 2. Ferkauf Graduate School of Psychology, Yeshiva University, New York, NY, USA
3. Icahn School of Medicine at Mount Sinai, New York, NY, USA

Background

- AI-powered virtual patients (VPs) offer scalable, low-risk practice for clinical training in suicide risk assessment.
- Technical errors still produce nonsensical or off-topic VP responses, which can break immersion and make it harder for trainees to build a therapeutic alliance.
- Prior work from our lab found no association between technical disruptions and working alliance or ruptures (Chen et al., 2025), suggesting minimal impact under current technology.
- Adding ChatGPT to a DialogFlow VP offered a natural test: if errors matter, a VP that makes fewer of them should yield a stronger alliance.

Present Study

Aim: Determine whether a systematic decrease in technical disruptions is associated with an increase in therapeutic alliance.

Hypotheses: Compared with the DialogFlow VP, the ChatGPT-enhanced VP would show:

H1 Fewer objective and perceived technical disruptions

H2 Higher Rupture Resolution Rating System (3RS) working together scores

H3 Lower rupture impact scores

H4 Fewer technical disruptions would predict better alliance outcomes

Method

Participants

- N = 101 doctoral student clinicians
  - DialogFlow VP (previous study): n = 69
  - DialogFlow + ChatGPT VP (ongoing pilot): n = 32

Procedure

- Diagnostic interview via Zoom with a teenage VP presenting with suicidal ideation, followed by post-session ratings

Measures

- **Objective disruption:** % of VP responses coded as No Match (input did not match any DialogFlow intent, triggering a fallback reply) or Nonsensical (input matched to the wrong intent, producing an irrelevant reply)
- **Perceived disruption:** single-item ratings (0 to 6) of platform disruption (e.g., Zoom, audio) and VP language disruption
- **Alliance and rupture (3RS):** working together (5 items, 1 to 7, α = .83); impact of internal, confrontation, and withdrawal ruptures (1 to 7)

Analyses (R)

- Welch's t tests with Cohen's d; simultaneous OLS regression (collinearity checked via VIF) with best subset selection by BIC; bootstrap indirect effect (5,000 resamples)

Results

29.6% → 17.3%
Objective disruption rate
p < .001

4.47 → 3.59
Internal rupture impact
p = .005

3.89 → 3.99
Working together
p = .639

H1 Far fewer disruptions with ChatGPT

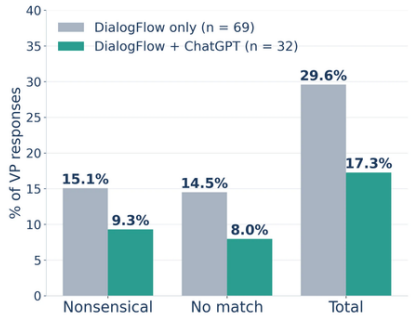


Figure 1. Objective disruptions by VP condition

Perceived language:
lower, t(48.0) = −2.12, p = .040
Perceived technical:
no difference, p = .591
H1 largely supported

H2, H3 Rupture impact eased; working together unchanged

Outcome	DialogFlow M	DialogFlow + ChatGPT M	p
Working together	3.89	3.99	.639
Internal rupture impact	4.47	3.59	**.005**
Confrontation rupture impact	2.91	2.29	.070
Withdrawal rupture impact	2.75	2.26	.151

Table 1. Alliance outcomes by VP condition. Welch's t tests; n (DialogFlow / ChatGPT): working together 69 / 32; internal 68 / 32; confrontation 64 / 31; withdrawal 63 / 31. Bold values indicate p < .05.

H2 not supported; H3 partially supported (internal rupture impact only)

H4 Perceived language disruption predicted alliance

Predictor	Working together		Internal rupture impact	
	b	*p*	*b*	*p*
Objective disruption rate	0.013	.254	0.000	.983
Perceived platform disruption	-0.057	.294	0.044	.536
Perceived language disruption	**-0.238**	**.012**	**0.258**	**.034**
VP condition (ChatGPT = 1)	0.200	.447	**-0.779**	**.024**
Model Fit				
R^2 (adjusted R^2)	.107 (.066)		.165 (.126)	
Odds Ratio	2.61, p = .041		4.28, p = .003	
Best subset model (BIC)	Perceived language disruption only		Language + VP condition	

Table 2. Simultaneous regression of alliance outcomes on disruption measures and VP condition. n = 92 participants with complete data (63 DialogFlow, 29 ChatGPT). All VIFs < 1.5. Bolded values indicate p < .05. Best subset selection compared all 16 predictor combinations by BIC.

H4 partially supported

Perceived language disruption, but not the objective error rate, predicted both alliance outcomes.

The lower internal rupture impact in the ChatGPT condition was not explained by fewer objective errors (indirect effect ab = −0.10, 95% bootstrap CI [−0.48, 0.32], n = 100).

Key takeaway

The ChatGPT-enhanced VP made fewer errors and elicited less negative internal reactions, but alliance tracked how disruptive trainees perceived the VP's language, not how many errors actually occurred.

Future direction

Randomized comparison of VP versions; coding schemes that capture ChatGPT-specific problems such as inauthentic responses.

Discussion

- The ChatGPT-enhanced VP was associated with about 40% fewer disruptions and less negative internal reactions, yet working alliance did not improve, echoing our earlier findings (Chen et al., 2025).
- Perceived language disruption predicted both alliance outcomes: how natural the VP sounds may matter more than how often it fails.
- Limitations:
  - nonrandomized cohorts; smaller ChatGPT sample; perceived disruption and alliance were both self-reported at the same time (shared method variance); multiple tests without correction.

References

Chen, J., Wei, J., Datta, S., Shi, R., & Bloch-Elkouby, S. (2025, June). *How much do inappropriate responses impede clinicians' engagement and therapeutic alliance with a virtual patient?* [Poster presentation].

Eubanks, C. F., Lubitz, J., Muran, J. C., & Safran, J. D. (2019). *Rupture Resolution Rating System (3RS): Development and validation.* Psychotherapy Research, 29(3), 306–319. https://doi.org/10.1080/10503307.2018.1552034