Toward Anxiety-Reducing Conversational AI Agents for Breast Cancer Care

Neda Ghafouri, Jane S. Gibson, Chaithanya Renduchintala, Yu Tian, Aritra Dutta, and Amrit Singh Bedi

University of Central Florida, Orlando, FL, USA
neda.ghafouri@ucf.edu, Jane.Gibson@ucf.edu, Chait.Rendu@ucf.edu, yu.tian2@ucf.edu, Aritra.Dutta@ucf.edu, amritbedi@ucf.edu

Abstract. Breast cancer care depends on patient–oncologist communication that can address medical concerns while also helping patients manage anxiety. Building conversational agents for this setting is difficult because real consultations are private and available dialogue corpora rarely represent emotional concerns across the care pathway. We present a multi-agent framework that converts patient narratives accessed with permission into evidence-constrained longitudinal conversations and compares support agents through a simulated expressed-anxiety proxy. Quality control retained 275 of 284 generated sessions (96.8%), yielding 1,650 support-agent responses with 74.5% scheduled-fact coverage and 99.1% correct prior-session references. The support agent was adapted using conversations from 21 training patients. The base and adapted agents were then evaluated on 14 unseen patients under matched scenarios, evidence, seeds, and dialogue length. Of 224 matched pairs across two seeds, 218 met the eligibility criteria. On the secondary final-turn measure, the adapted agent had lower proxy anxiety in 109 pairs and the base agent in 77; 32 were tied. Supportive prompting reduced the mean trajectory proxy from 7.774 to 7.296, whereas uniformly rewriting training responses into a fixed supportive structure increased repetition and worsened simulated outcomes.

Keywords: Anxiety Reduction · Breast Cancer · Multi-Agent Simulation · Conversational agent · Synthetic Medical Dialogue · Supportive Oncology

1 Introduction

Anxiety can persist across diagnosis, treatment, and follow-up in breast cancer, particularly when patients face uncertainty or fear of recurrence [12, 6]. Clinician communication can affect anxiety after consultation [10], but opportunities for support during and between visits are limited. Conversational agents may provide a controlled setting for studying supportive communication, provided that they remain grounded in the patient's history and are evaluated on more than how empathetic their responses sound. The development of such an agent faces two obstacles.

First, suitable training data are not readily available. Real patient–clinician consultations contain sensitive medical and personal information and are rarely available for model development. Note-conditioned methods such as NoteChat and SynDial generate medically grounded conversations, but provide limited information about patients' emotional experiences over time [16, 5]. General emotional-support datasets capture distress but lack cancer-specific medical grounding [11]. Public patient narratives contain both medical events and personal concerns, but they must be organized into time-ordered sessions and restricted to information available at each stage. Synthetic generation introduces further risks, including unsupported details, generic reassurance, and repeated response templates.

Second, supportive language is not itself a patient outcome. Existing medical simulators often emphasize diagnostic accuracy or patient realism [14, 8], while general emotional-support datasets may track distress or emotion change without the longitudinal medical constraints of breast-cancer care [11, 19]. Our question is therefore narrower: under matched simulated conditions, does a support agent's communication change the simulated patient's expressed-anxiety trajectory? This requires paired interactions and a fixed outcome definition, while treating the resulting score only as a simulation-based evaluation measure.

To address the data challenge, we convert public breast-cancer patient narratives into longitudinal support sessions and restrict each session to patient-specific evidence available at that stage. To address the evaluation challenge, we compare support-agent conditions through matched six-round interactions and track the simulated patient's expressed-anxiety trajectory. The resulting framework supports pre-deployment research on supportive medical communication; it is not an autonomous oncology service. We then use the framework to examine how support-focused adaptation and the extent of conversation rewriting affect response diversity and simulated an expressed-anxiety trajectories. The complete framework is shown in Fig. 1. The main contributions of our work are:

- **Longitudinal lived-experience dialogue construction.** We convert public breast-cancer narratives into temporally ordered support sessions with patient-specific evidence constraints, producing 275 quality-controlled sessions spanning diagnosis, treatment, and follow-up.
- **Outcome-oriented paired simulation.** We compare doctor models on matched fixtures using changes in a fixed simulated patient's turn-level expressed-anxiety proxy rather than response-level empathy alone.
- **Evidence on training-data design.** We conduct a controlled analysis of how the composition and structural diversity of synthetic support conversations affect the adapted agent's response patterns and simulated expressed-anxiety trajectory.

2 Related Work

Emotional support and cancer chatbots. ESConv annotates support strategies and changes in seeker emotion across a conversation [11], while FEEL uses

multiple language-model judges to assess support quality [19]. SocialSim generates persona-conditioned emotional-support dialogues [3], and evaluation with difficult seekers shows that apparent effectiveness can depend on how readily the simulated user accepts support [18]. Cancer-focused trials also suggest that digital counseling and educational chatbots may reduce distress in specific settings [1, 9]. These studies motivate emotion-aware evaluation, while work with difficult seekers cautions that simulated outcomes may depend on how readily the user model accepts support.

Synthetic medical dialogue and patient simulation. NoteChat and SynDial construct medical conversations from clinical documents using multi-agent generation or iterative refinement [16, 5]. PatientSim constructs and evaluates persona-grounded simulated patients, whereas AgentClinic uses simulated encounters to evaluate sequential diagnostic decisions [8, 14]. These systems primarily emphasize source fidelity, patient realism, or diagnostic performance. Model-based simulation and evaluation also raise validity concerns: patient responses may depend on how readily the simulator accepts support, while support-quality judgments may vary with the model-based evaluation procedure [19, 18]. Matched conditions and a condition-blind observer support controlled comparison but do not clinically validate the resulting proxy. Our work differs by jointly studying breast-cancer lived-experience narratives, stage-specific evidence constraints, continuity across multiple sessions, and matched evaluation of simulated an expressed-anxiety trajectories.

3 Proposed Framework

Our goal is to construct longitudinal support conversations that remain faithful to a patient's experience while allowing us to study whether different support strategies change the simulated patient's expressed anxiety. This requires controlling both the information available to the agents and the quality of the generated conversations. We therefore build the framework in three stages: constructing temporally grounded dialogue sessions, filtering the resulting conversations, and using an anxiety-based proxy to select and evaluate support strategies.

Longitudinal dialogue construction. We first convert each patient narrative into a structured profile containing supported clinical facts, concerns, and recurring experiences. These are ordered along the patient's care journey and used to construct session fixtures that contain the current focus, evidence available at that stage, a summary of prior sessions, and an initial simulated anxiety state. At each session, the support agent receives the current fixture, dialogue history, and retrieved supportive-care guidance, while the patient simulator receives its profile, state, and conversation history. Importantly, information from later stages of the patient's journey is withheld until the corresponding session. This temporal restriction prevents future events from leaking into earlier conversations and ensures that patient-specific claims can be traced to the source narrative. Grounding the dialogue in this way controls what information the agents can use, but it does not by itself guarantee that a generated conversation

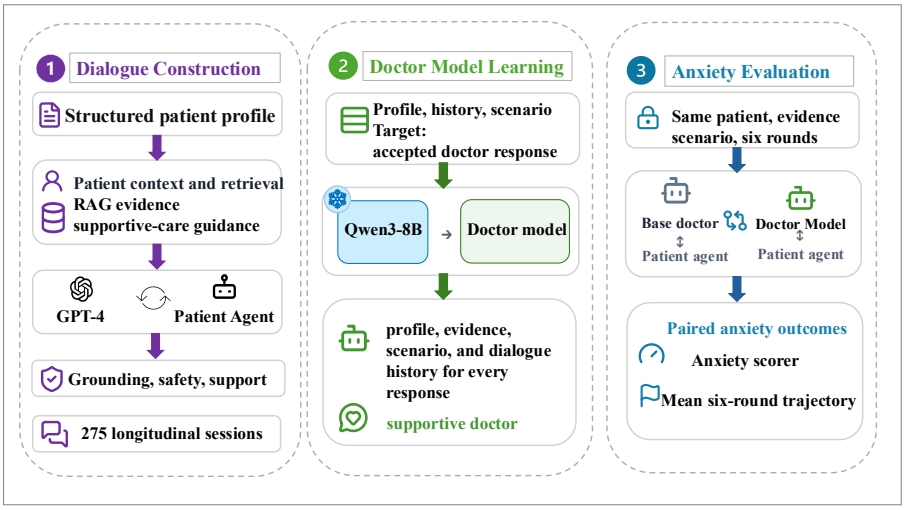

Fig. 1. Overview of the proposed framework. Public breast-cancer narratives are converted into structured patient profiles and temporally constrained support sessions. Quality-controlled doctor responses are used for model adaptation. The base and adapted doctor models are then compared through matched interactions with the same patient simulator, and a fixed, condition-blind observer evaluates their simulated expressed-anxiety trajectories.

is suitable for training. Responses may still be incomplete, unsafe, inconsistent with earlier sessions, or overly repetitive. We therefore apply a second stage of conversation-level quality control.

Conversation quality control. Each generated conversation is checked for grounding, safety, completeness, continuity, and repetition. Let $q_m(c) \in \{0,1\}$ indicate whether conversation c satisfies check m. We retain

$$\mathcal{D}_Q = \left\{ c : \mathcal{Q}(c) = \prod_{m=1}^{M} q_m(c) = 1 \right\}, \tag{1}$$

and reject or regenerate conversations that fail any check. We also monitor recurring openings and response patterns to ensure that quality filtering does not collapse support-agent responses into a small set of templates. These checks determine whether a conversation is sufficiently grounded and coherent to be considered further, but they do not tell us whether the interaction had a favorable effect on the simulated patient. To make that distinction, we separately measure the simulated patient's expressed anxiety. This allows two responses that are both safe and grounded to be distinguished by how the subsequent patient responses evolve.

Expressed-anxiety proxy and conversation selection. HADS-A is a seven-item self-report instrument covering symptoms over the preceding week [13, 15].

We do not administer HADS-A. Instead, a fixed GPT-5.5 observer applies a HADS-A-inspired rubric to each generated patient response. The patient simulator generates the dialogue but never evaluates itself. Each item is scored three times, and the turn-level score is obtained by taking the median score for each item and then summing across the seven items:

$$\widehat{A}_t = \sum_{\ell=1}^{7} \mathrm{median}_{k=1}^{3}\, a_{t,\ell}^{(k)}, \qquad B(c) = \frac{1}{6}\sum_{t=1}^{6} \widehat{A}_t. \tag{2}$$

Here, $B(c)$ summarizes the expressed-anxiety trajectory over the six patient turns. It is a simulation-based proxy rather than a clinical HADS-A score. We use this proxy during conversation selection by comparing each candidate conversation c with a reference conversation r generated under the same patient, evidence, initial state, random seed, and dialogue length. After requiring replicated observer scores to satisfy the stability criterion $\mathcal{R}$, the retained outcome set is

$$\mathcal{D}_O = \{c \in \mathcal{D}_Q : \mathcal{R}(c) = \mathcal{R}(r) = 1,\ B(c) \le B(r)\}. \tag{3}$$

Thus, the training corpus contains conversations that satisfy the grounding and quality requirements while not producing a worse simulated anxiety trajectory than their matched references. Because this outcome proxy is also used during conversation selection, a fair comparison of the resulting adapted agent requires controlling the interaction conditions at evaluation time. We therefore compare the base and adapted agents through matched simulations.

Matched evaluation. For each patient and session, the base and adapted support agents interact with the same patient simulator using the same evidence, initial state, random seed, and conversation length. A pair is included only when both six-round conversations are complete and their replicated observer scores satisfy the stability criterion. For patient i, session s, and seed z, we define the gain

$$\begin{aligned} G_{i,s,z} &= B(c_{i,s,z}^{\mathrm{base}}) - B(c_{i,s,z}^{\mathrm{adapted}}), \\ \overline{G}_i &= \frac{1}{|\mathcal{E}_i|}\sum_{(s,z)\in\mathcal{E}_i} G_{i,s,z}, \qquad \overline{G} = \frac{1}{N_p}\sum_{i=1}^{N_p}\overline{G}_i, \end{aligned} \tag{4}$$

where $\mathcal{E}_i$ contains the eligible session–seed pairs for patient i. A positive gain therefore indicates a lower expressed-anxiety proxy for the adapted agent. We first average sessions and seeds within each patient and estimate confidence intervals by resampling patients. Mean six-turn gain is the primary endpoint, while final-turn gain and win–tie–loss counts are reported as secondary comparisons. The GPT-5.5 observer remains fixed and blinded to agent condition throughout evaluation.

4 Experiments

Data and patient-level split. We use 35 breast-cancer narratives from the Health Experiences Canada Breast Cancer in Women module with permission

from the data provider [2]. Before adaptation, patients were divided by identity into 21 training patients and 14 evaluation patients; no separate development-patient cohort was used. Every derived profile, evidence packet, session fixture, and dialogue retained its patient's assignment, and no evaluation-patient material entered adaptation. Scenario entry priors follow HADS-A distributions reported for breast-cancer cohorts before surgery, during treatment, and in survivorship [4, 7, 17, 12]. They initialize simulated states only and are not HEC participant labels or clinical HADS-A measurements.

Evaluation protocol. Base and adapted agents interact with the same patient simulator for six rounds under matched profile, evidence, scenario, entry state, seed, and response limit (200 new tokens). Across the 14 held-out evaluation patients, 112 unique patient–session fixtures were each run with two random seeds, yielding 224 matched base–adapted pairs and 448 individual conversations. Of these, 218 pairs satisfied the prespecified completeness and replicated-score stability criteria and were included in the analysis.

The fixed observer is blinded to agent condition and scores each patient response three times. Patient simulators produce dialogue only; they do not assign anxiety scores. GPT-5.5 is the sole prespecified evaluator for the primary and secondary comparisons and remains fixed when the patient simulator changes. Mean six-turn trajectory gain is the primary endpoint; final-turn gain is secondary. Session and seed gains are averaged within patient before computing the patient-weighted mean, and 95% confidence intervals use patient-level bootstrap resampling. Win–tie–loss counts are descriptive. Because GPT-5.5-based scoring also influenced conversation selection, the evaluation is an internally consistent simulation measure rather than independent clinical validation.

Implementation. Dialogues were generated with GPT-4, and Qwen2.5-14B-Instruct acted as the patient simulator. Qwen3-8B was the primary base support agent and was adapted with response-only LoRA ($r = 8$, $\alpha = 16$, dropout 0.05) under 4-bit quantization. Training used AdamW, a learning rate of 5×10^{-5}, effective batch size eight, five epochs, and a 2,048-token context. Gemma-2-2B used the same response-only adaptation protocol. Training and inference used one NVIDIA B200 GPU.

4.1 Results and Discussion

Conversation quality. Quality control retained 275 of 284 sessions (96.8%), yielding 1,650 support-agent turns across 275 six-turn conversations (Table 1). No retained session triggered the deterministic grounding filter; because failures were regenerated or removed, this reflects the admission gate rather than an independent factuality estimate. Dialogues covered 74.5% of scheduled facts and obtained source-to-dialogue ROUGE-1/2/L scores of 0.364/0.097/0.163, indicating moderate lexical overlap with the supporting evidence. Prior-session references were correct in 99.1% of longitudinal sessions. Support-agent turns were less repetitive than the source narratives (Self-BLEU-4: 0.544 versus 0.577), while

Table 1. Intrinsic evaluation of the generated conversation. Arrows mark the preferred direction (↑ higher is better, ↓ lower is better); extractiveness carries no arrow because moderate values are desirable.

Axis	Metric	Result
Factuality	Unsupported clinical claims (deterministic audit) ↓	0
	Scheduled-fact coverage in doctor turns ↑	74.5%
Extractiveness	ROUGE-1/2/L F1, source facts → dialogue	0.364 / 0.097 / 0.163
Diversity	Self-BLEU-4, doctor / patient / all ↓	0.544 / 0.442 / 0.516
	Self-BLEU-4, source narratives (reference) ↓	0.577
Continuity	Prior-session reference rate ↑	99.1%
Anxiety rendering	Sessions with robust HADS-inspired measurements ↑	275/275
Measurement agreement	Item-score vs. linguistic observer, mean $\lvert\Delta\rvert$ ↓	0.373

patient turns were more diverse than support-agent turns (0.442 versus 0.544), providing no evidence of corpus-wide template collapse.

Prompting and training-data design. With matched fixtures, the mean trajectory proxy decreased from 7.774 under detached instructions to 7.526 under the standard prompt and 7.296 under supportive instructions (Table 2). The support-focused corpus produced the largest selection-stage gain (+1.417), whereas grounded-varied and warmth-weighted corpora remained near the base agent (−0.050 and −0.324). Uniform rewriting produced a −0.863 gain (95% CI $[-1.195, -0.562]$), together with fewer question-ending turns (45.2% to 17.6%), higher Self-BLEU-4 (0.442 to 0.595), and one recurring template in 13.2% of turns.

Evaluation and robustness. The final evaluation included 218 matched pairs from 14 unseen patients across two seeds. For the primary six-turn trajectory endpoint, the patient-weighted gain was −0.042, indicating near parity. For the secondary final-turn endpoint, the adapted agent had the lower proxy score in 109 pairs, the base agent in 77, and 32 were tied.

Gemma-2-2B, evaluated on 10 non-training patients and 92 pairs, showed no clear adaptation advantage (−0.175, 95% CI $[-0.536, 0.174]$). Its base and adapted arms moved from a mean entry prior of 10.09 to final proxy scores of 1.68 and 1.84, respectively, with 47/92 pairs (51.1%) indistinguishable, suggesting a floor effect that limited separation. On the same pairs, the internal and external observers produced opposite base–adapted gains (+0.163 and −0.152) and agreed on only 45/92 outcomes (48.9%), showing sensitivity to evaluator choice. We therefore use one prespecified, condition-blind observer consistently while treating the resulting effects as observer-dependent simulation outcomes. At the turn level, question-ending support turns were followed by larger proxy decreases than other turns (−1.629 versus −0.646), whereas response length showed little association with next-turn change ($\rho = 0.079$; 1,980 turns).

Table 2. Effects of prompting and training-corpus construction. Panel (a) reports mean trajectory scores (lower is calmer); panel (b) reports the gain, base minus adapted, where positive favours the adapted agent.

Condition	What was changed	Result
(a) Prompt comparison: mean trajectory proxy score ↓		
Detached prompt	Accurate and safe responses, without an instruction to provide emotional support	7.774
Standard prompt	Default doctor instructions used by the base model	7.526
Supportive prompt	Recognize the concern, provide bounded reassurance, and offer one manageable action	**7.296**
(b) Selection-stage base–adapted proxy gain ↑		
Support-focused	Emotional recognition, bounded reassurance, one action, and gentle closure	**+1.417**
Grounded varied	Patient-specific evidence, natural wording, varied responses, and follow-up	−0.050
Warmth-weighted	The same grounded corpus, with warmer and more reassuring responses sampled more often	−0.324
Heavily rewritten	Doctor responses rewritten to follow a similar validation–action–closure structure	−0.863
Language changes after heavy rewriting		
Turns ending with a question	Base doctor → heavily rewritten doctor	45.2% → 17.6%
Self-BLEU-4	Base doctor → heavily rewritten doctor; higher means more repetition	0.442 → 0.595
Most frequent template	Percentage of heavily rewritten doctor turns using the same recurring template	13.2%

Table 3. Cross-model robustness and turn level sensitivity. Positive proxy gain favors adaptation; more negative next-turn change is better.

Analysis	Scope	Result
Second support model: Gemma-2-2B		
Base–adapted final-turn proxy gain	10 patients, 92 pairs	−0.175
Patient-bootstrap 95% CI	10 patients	$[-0.536, 0.174]$
Indistinguishable final states	External observer	47/92 (51.1%)
Entry prior → final proxy	Base / adapted	10.09→1.68 / 1.84
Evaluator and turn-level sensitivity		
Base–adapted gain, internal / external	Same 92 pairs	+0.163 / −0.152
Evaluator outcome agreement	Same 92 pairs	45/92 (48.9%)
Next-turn change: question / other	7 patients, 3 seeds	−1.629 / −0.646
Length vs. next-turn change	1,980 turns	$\rho = 0.079$

5 Conclusions

We built a multi-agent framework that converts public breast-cancer stories into grounded, longitudinal support conversations. Instead of judging agents by how supportive their wording appears, we selected training data and compared models using the simulated patient's measured anxiety. Training on conversations that recognized the patient's concern, offered bounded reassurance, and closed with one manageable step produced the largest gain. When every response was rewritten into the same structure, the agent repeated itself and outcomes declined. Warmth helped when it addressed the concern, and stopped once it became a template. Future work should include independent human and clinical review of naturalness, faithfulness, safety, and support quality, with inter-rater

agreement analysis. Prospective validation is needed before these trajectories can be linked to real patient outcomes.

Acknowledgment. This material is based upon work supported by the Florida Department of Health.

References

1. Akdogan, O., Yesilbas, E., Baskurt, K., Ozdemir, N., Yazici, O., Oksuzoglu, B., Uner, A., Uyar, G.C., Malkoc, N.A., Ozet, A., Sutcuoglu, O.: Effect of a ChatGPT-based digital counseling intervention on anxiety and depression in patients with cancer: A prospective, randomized trial. European Journal of Cancer **221**, 115408 (2025)
2. Canadian Health Experiences Group: Breast cancer in women module. Health Experiences Canada, St. Mary's Research Centre (2024), https://healthexperiences.ca/breast-cancer/, accessed: June 2026. Research led by S. Law, D. Stern, and I. Ormel, in collaboration with the DIPEx International network.
3. Chen, Z., Cao, Y., Bi, G., Wu, J., Zhou, J., Xiao, X., Chen, S., Wang, H., Huang, M.: SocialSim: Towards socialized simulation of emotional support conversation. Proceedings of the AAAI Conference on Artificial Intelligence **39**(2), 1274–1282 (2025). https://doi.org/10.1609/aaai.v39i2.32116
4. Civilotti, C., Botto, R., Acquadro Maran, D., Bianciotto, B., De Leonardis, B., Stanizzo, M.R.: Anxiety and depression in women newly diagnosed with breast cancer and waiting for surgery: Prevalence and associations with socio-demographic variables. Medicina **60**(7), 1112 (2024). https://doi.org/10.3390/medicina60071112
5. Das, T., Albassam, D., Sun, J.: Synthetic patient-physician dialogue generation from clinical notes using LLM. arXiv preprint arXiv:2408.06285 (2024)
6. Jiang, Y., Li, H., Xiong, Y., Zheng, X., Liu, Y., Zhou, J., Ye, Z.: Association between fear of cancer recurrence and emotional distress in breast cancer: a latent profile and moderation analysis. Frontiers in Psychiatry **16**, 1521555 (2025). https://doi.org/10.3389/fpsyt.2025.1521555
7. Kim, J., Cho, J., Lee, S.K., Choi, E.K., Kim, I.R., Lee, J.E., Kim, S.W., Nam, S.J.: Surgical impact on anxiety of patients with breast cancer: 12-month follow-up prospective longitudinal study. Annals of Surgical Treatment and Research **98**(5), 215–223 (2020). https://doi.org/10.4174/astr.2020.98.5.215
8. Kyung, D., Chung, H., Bae, S., Kim, J., Sohn, J.H., Kim, T., Kim, S.K., Choi, E.: PatientSim: A persona-driven simulator for realistic doctor-patient interactions. arXiv preprint arXiv:2505.17818 (2025)
9. Lee, J., Byun, H.K., Kim, Y.T., Shin, J., Kim, Y.B.: A study on breast cancer patient care using chatbot and video education for radiation therapy: A randomized controlled trial. International Journal of Radiation Oncology, Biology, Physics (2025)
10. Liénard, A., Merckaert, I., Libert, Y., Delvaux, N., Marchal, S., Boniver, J., Etienne, A.M., Klastersky, J., Reynaert, C., Scalliet, P., Slachmuylder, J.L., Razavi, D.: Factors that influence cancer patients' anxiety following a medical consultation: impact of a communication skills training programme for physicians. Annals of Oncology **17**(9), 1450–1458 (2006). https://doi.org/10.1093/annonc/mdl142

11. Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., Jiang, Y., Huang, M.: Towards emotional support dialog systems. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP). pp. 3469–3483 (2021)
12. Lopes, C., Lopes-Conceição, L., Fontes, F., Ferreira, A., Pereira, S., Lunet, N., Araújo, N.: Prevalence and persistence of anxiety and depression over five years since breast cancer diagnosis—the NEON-BC prospective study. Cancers **16**(5), 946 (2024). https://doi.org/10.3390/cancers16050946
13. Rodgers, J., Martin, C.R., Morse, R.C., Kendell, K., Verrill, M.: An investigation into the psychometric properties of the hospital anxiety and depression scale in patients with breast cancer. Health and Quality of Life Outcomes **3**, 41 (2005). https://doi.org/10.1186/1477-7525-3-41
14. Schmidgall, S., Ziaei, R., Harris, C., Kim, J.W., Reis, E., Jopling, J., Moor, M.: AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. arXiv preprint arXiv:2405.07960 (2024)
15. Vodermaier, A., Millman, R.D.: Accuracy of the hospital anxiety and depression scale as a screening tool in cancer patients: a systematic review and meta-analysis. Supportive Care in Cancer **19**, 1899–1908 (2011). https://doi.org/10.1007/s00520-011-1251-4
16. Wang, J., Yao, Z., Yang, Z., Zhou, H., Li, R., Wang, X., Xu, Y., Yu, H.: NoteChat: A dataset of synthetic patient-physician conversations conditioned on clinical notes. arXiv preprint arXiv:2310.15959 (2023)
17. Wang, S.J., Feng, X., Zhang, W., Sun, G.Y., Fang, H., Jing, H., Tang, Y., Li, T., Qi, S.N., Song, Y.W., Zhang, W.W., Lu, N.N., Tang, Y., Liu, Y.P., Chen, B., Liu, X., Li, Y.X., Zhai, Y.R., Wang, S.L.: Depression and anxiety in patients with breast cancer receiving radiotherapy: a longitudinal study. Therapeutic Advances in Medical Oncology **17** (2025). https://doi.org/10.1177/17588359251345930
18. Yang, J., Li, Y., Chen, G., Fan, R., Bai, X., He, T.: When seekers are hard to help: Evaluating emotional support dialogue systems in worst-case interactions. arXiv preprint arXiv:2605.28228 (2026)
19. Zhang, H., Chen, Y., Wang, M., Feng, S.: FEEL: A framework for evaluating emotional support capability with large language models. arXiv preprint arXiv:2403.15699 (2024)