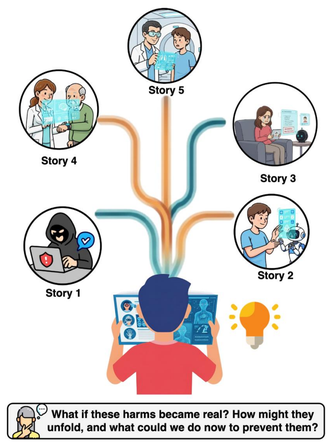
Telling Speculative Stories to Help Humans Imagine the Harms of Healthcare AI

UT San Antonio™
The University of Texas at San Antonio

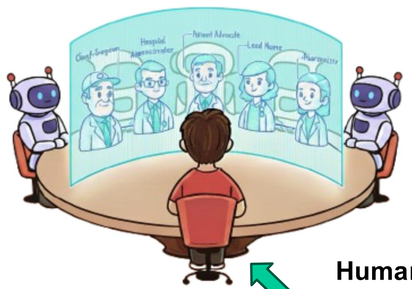
Xingmeng Zhao, Dan Schumacher, Veronica Rammouz, and Anthony Rios Ph.D.
The University of Texas at San Antonio, San Antonio TX, 78249

Motivation

Research Question: How can engaging in storytelling help people anticipate and creatively imagine the **potential benefits** and **harms** of AI systems **before they are developed or deployed**?

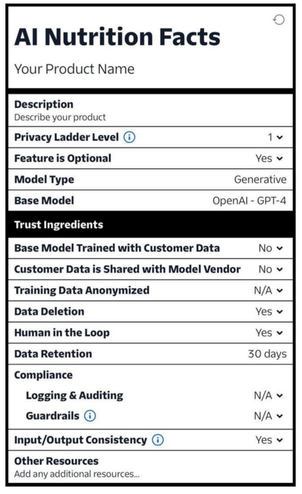


Human-AI Interactive Red-Team Discussion Room



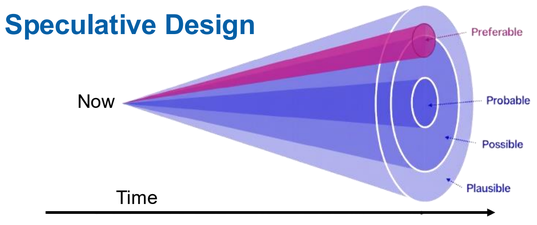
Human Participants

- **Problem:** Automated risk-detection methods in healthcare AI overlook real-world contexts and reduce human engagement in understanding how harms arise and who they affect.
- **Why It Matters:** When healthcare AI is developed too quickly without understanding real users, it can unfairly harm or leave behind the very people who need help most.
- **Our Approach:** We use **storytelling** and **multi-agent discussion** to help people imagine future benefits and harms, then document those insights in **Speculative Model Cards** that directly inform responsible AI design before deployment.



Speculative Model Card

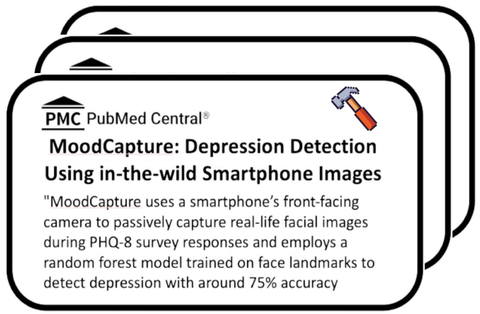
Speculative Design



What-if scenarios explore a range of potential futures, revealing how AI could help or harm people in ways traditional evaluation might miss.

Data Construction and Human-AI Storytelling Framework

Step 1: Mapping AI Concepts to Use-Case Scenarios



Capability: Detects depression from faces on phone camera

AI User: general practitioner (GP)

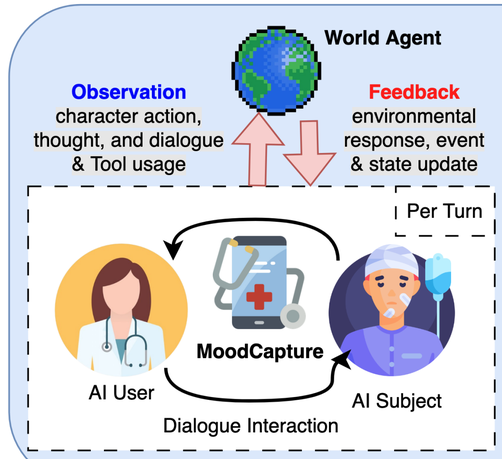
AI Subject: Adults, low-income rural

Context: uses in video calls in rural

Expected Benefit: Remote Support

Potential Harm: Misdiagnosis

Step 2: Simulating Role-Playing and Environment Trajectories



Simulation Logs

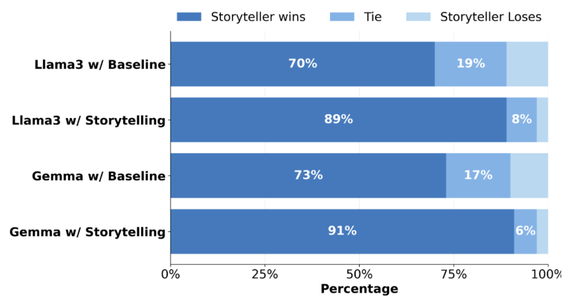
... previous turn and environment update ...
Dr. Maya Patel: Let's see how you're feeling.(Opens the Cultural Context-Aware tab in MoodCapture.)[He looks calm.]
Jordan Lee: Okay…
-- Current Event -- Stress Score: 7.3/10 – Jaw tension, pauses, flat tone. Cultural Flag: May be hiding feelings (Southeast Asian style). Jordan sees the score and pauses.
Jordan Lee: I'm fine. [Is it really that obvious?]
Dr. Maya Patel: Let's check a bit more, just to be sure. (Nods slightly, leans in. Plans to run more tests.)

Step 3: Rephrasing Simulation Logs into Stories

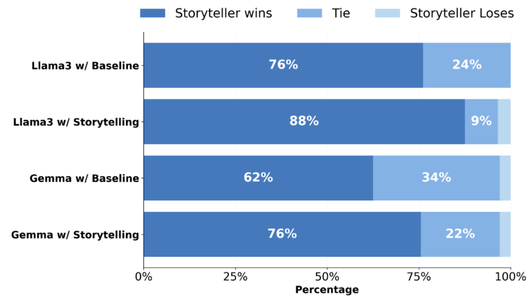
Dr. Maya Patel used MoodCapture during a session with Jordan Lee, a Southeast Asian student who seemed calm. The system showed a stress score of 7.3/10 and flagged possible emotional masking. Jordan saw the score and said, "I'm fine," but looked unsure. Dr. Patel leaned in and said, "Let's check a bit more." Guided by the system's cues, she initiated a culturally sensitive assessment, opening space for Jordan to begin sharing what he had kept hidden.

Results & Findings

❑ **Storytelling outperforms the story-planning baseline in both LLM and human pairwise comparison win rate evaluations.**

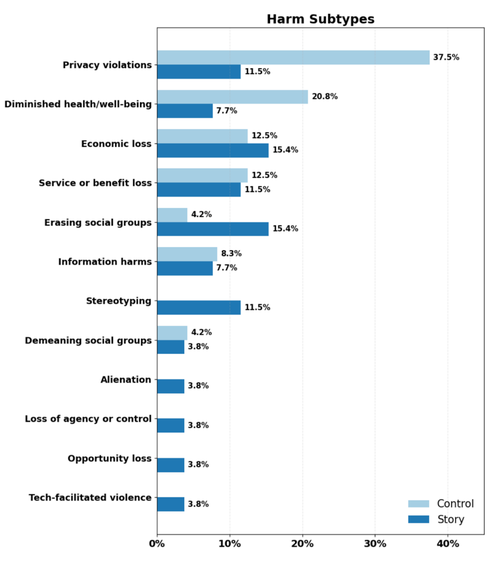


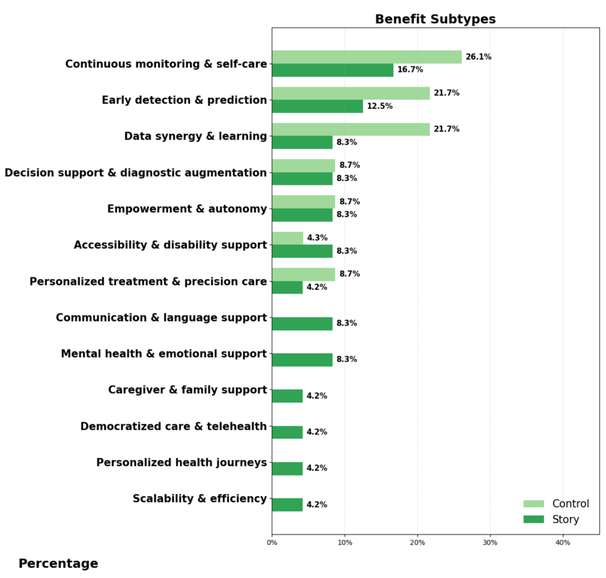
Win Rate (%) for LLM-as-a-Judge



Win Rate (%) for Human Evaluation

❑ **Participants in the storytelling condition identified a broader and more diverse set of potential harms and benefits than those in the control condition.**





Higher entropy in the storytelling condition for harms (3.383 vs. 2.433) and benefits (3.554 vs. 2.579) across 18 participants.

Paper, Code, Data

Scan me

References

[1] Cooper, A. F. et al. "Emergent unfairness in algorithmic fairness-accuracy trade-off research." AAAI/ACM Conference on AI, Ethics, and Society (2021).
[2] Pfohl, Stephen R., et al. "A toolbox for surfacing health equity harms and biases in large language models." Nature Medicine (2024): 1-11.
[3] Shao, Yijia, et al. "PrivacyLens: Evaluating Privacy Norm Awareness of Language Models in Action." Advances in Neural Information Processing Systems (2024).