Speculative Model Risk in Healthcare AI: Using Storytelling to Surface Unintended Harms

Xingmeng Zhao, Dan Schumacher, Veronica Rammouz, and **Anthony Rios**
The University of Texas at San Antonio
{xingmeng.zhao, anthony.rios}@utsa.edu

Abstract

Artificial intelligence (AI) is rapidly transforming healthcare, enabling fast development of tools like stress monitors, wellness trackers, and mental health chatbots. However, rapid and low-barrier development can introduce risks of bias, privacy violations, and unequal access, especially when systems ignore real-world contexts and diverse user needs. Many recent methods use AI to detect risks automatically, but this can reduce human engagement in understanding how harms arise and who they affect. We present a human-centered framework that generates user stories and supports multi-agent discussions to help people think creatively about potential benefits and harms before deployment. In a user study, participants who read stories recognized a broader range of harms, distributing their responses more evenly across all 13 harm types. In contrast, those who did not read stories focused primarily on privacy and well-being (58.3%). Our findings show that storytelling helped participants speculate about a broader range of harms and benefits and think more creatively about AI's impact on users. Dataset and code are available at https: https://anonymous.4open.science/r/storytelling-healthcare/README.md.

1 Introduction

AI is integrated into many aspects of daily life, including finance, healthcare, law, and recommendation systems (Ashurst et al., 2020). In healthcare, tools such as stress monitors (Kargarandehkordi et al., 2025), wellness trackers (Fabbrizio et al., 2023), and mental health chatbots (MacNeill et al., 2024) can directly influence well-being. Recent prompting methods like vibe coding (Chow and Ng, 2025) let people describe how an AI system should behave or sound, making it easier for non-experts to prototype AI tools. For instance, CareYaya allows caregivers to build stress-tracking applications using short natural language descriptions (Kenny, 2023). While this innovation speeds up AI development, it also introduces risks related to fairness, bias, and accountability (Weidinger et al., 2023). These risks are serious in healthcare, where even small model errors can cause harm, including delayed treatment, privacy loss, or health inequities (Roller et al., 2020; Chinta et al., 2025). Systems lacking safeguards for fairness and accountability may ultimately harm the very users they are meant to help (Shelby et al., 2022, 2023). Governments are responding through measures such as the EU AI Act (European Parliament, 2023) and a U.S. Executive Order (Biden, 2023), which call for greater transparency, safety, and accountability (Khan et al., 2025). Yet regulation remains region-specific and slow to adapt, making early ethical foresight essential for identifying risks, preventing misuse, and aligning AI systems with human values (Saxena et al., 2025).

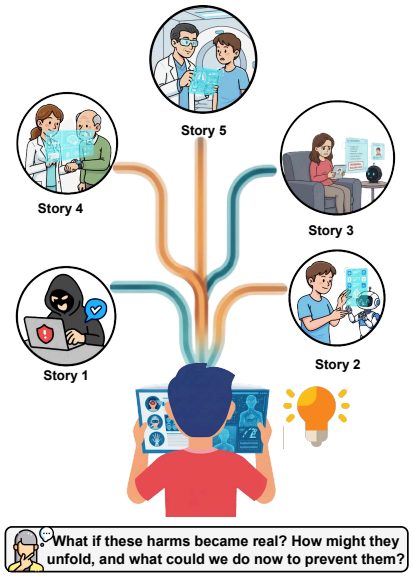

Figure 1: Illustration of using speculative stories to help people imagine potential harms and benefits of healthcare AI and foster more creative and ethical thinking.

Ethical challenges in AI are commonly addressed through two complementary strategies: documenting risks and anticipating them early.

Model cards (Mitchell et al., 2019) describe how a system works, its intended uses, and its limitations, and have been extended with interactive (Crisan et al., 2022) and structured formats (Bhat et al., 2023). RiskRAG (Rao et al., 2025) builds on this idea by generating risk summaries from model cards and real-world incident data. While automation helps scale risk assessment, it can also reduce opportunities for thoughtful human reflection (Kosmyna et al., 2025) and even introduce new harms (Dutta et al., 2020). Another line of work focuses on anticipating potential misuse (Herdel et al., 2024) and harm (Deng et al., 2025; Saxena et al., 2025) early in the design process. Tools such as AHA! (Buçinca et al., 2023), which uses storytelling, and Farsight (Wang et al., 2024b), which provides real-time risk alerts, help surface ethical issues during ideation. Yet, as these systems increasingly automate ethical reflection, there is a risk that people may stop thinking critically about how AI systems could cause harm. When reflection is delegated to AI, users may begin to rely on its judgment instead of their own, making it harder to notice ethical or contextual issues.

This issue is especially critical in healthcare, where small design mistakes can lead to serious harm (Mennella et al., 2024; Gilbert et al., 2025). Many risks only appear after deployment, when AI systems face complex real-world conditions (Mun et al., 2024; Kingsley et al., 2024). For example, mental health apps may fail to detect crises among certain groups, such as adolescents or non-native speakers (Zhai et al., 2024). Current tools often rely on automation to predict such risks, but this can distance humans from ethical reasoning. Speculative design and design fiction offer a different approach, using imagined scenarios to explore how technologies might succeed or fail (Rahwan et al., 2025). However, few approaches support human participation in ethical foresight or connect speculative thinking to real design workflows. Speculative storytelling bridges this gap by helping people reason about AI's potential benefits and harms in context (Li et al., 2025b).

Building on Klassen and Fiesler (2022), who use speculative fiction to explore emerging technologies, we apply storytelling to prompt early ethical reflection in AI design (Figure 1). Our goal is to test whether storytelling helps people think more creatively about potential benefits and harms, encouraging human speculation rather than relying on AI to anticipate risks. We introduce a human-centered framework that combines automated user story generation with structured red-team discussions. Unlike plot-planning approaches (Xie and Riedl, 2024), our method produces context-sensitive stories grounded in users' identities, behaviors, and needs. These stories help participants imagine how AI systems might succeed or fail in realistic scenarios, improving their ability to anticipate ethical and social implications. We use model cards to evaluate how clearly and diversely participants express potential harms. In our user study, story-driven discussions led to more context-specific risk identification and more detailed and varied responses to the model card.

Our contributions are twofold. **(1)** We introduce a method that automatically generates user stories to help people imagine how an AI system could help or harm users before it is developed or deployed. **(2)** We present a user study showing story-driven discussions with AI agents help participants explore potential risks and benefits more creatively and think more broadly about ethical issues.

2 Related Work

Model Cards Framework. Model cards document an AI model's purpose, performance, data sources, and limitations (Mitchell et al., 2019). Later work improved their usability and scale: Crisan et al. (2022) created *Interactive Model Cards* for exploring subgroup results, Bhat et al. (2023) developed *DocML* to guide non-experts, and Rao et al. (2025) introduced *RiskRAG*, which uses retrieval-augmented generation to summarize risks from model cards and incident reports. Derczynski et al. (2023) proposed *Risk Cards* to describe failure cases in context. While these tools increase transparency, they rely on automation to fill ethical gaps, which can reduce opportunities for human reflection. Our approach instead uses AI to support human speculation, helping people imagine how systems might succeed or fail before deployment.

Speculative Design. Speculative design uses imagined stories to explore how future technologies could affect people and society before those technologies exist. Rather than predicting the future, this approach uses what-if scenarios to prompt reflection on assumptions, values, and potential harms (Klassen and Fiesler, 2022; Hoang et al., 2018). This method treats fiction as a tool for early ethical reasoning, helping researchers consider social impacts while designing systems (Rah-

wan et al., 2025). Recent work such as Buçinca et al. (2023) uses LLM-generated failure scenarios to automatically highlight harms, and Wang et al. (2024b) adds AI-driven risk prompts into prototyping workflows. In healthcare, Hoang et al. (2018) use "Fiction Probes" to let patients navigate imagined care scenarios, while Klassen and Fiesler (2022)'s Black Mirror Writers Room invites students to write near-future stories that reveal hidden harms. Beyond these examples, participatory workshops, crowdsourced case studies, and AI-assisted red-teaming extend speculative exploration by engaging a broader range of voices in imagining and testing possible futures (Mun et al., 2024; Radharapu et al., 2023). Unlike prior work that uses AI to generate harms directly, we use AI-generated stories to spark human thinking and help people imagine harms themselves.

Language-Based World Modeling. Humans often imagine situations to predict what may happen next, reason through alternatives, and make decisions (Addis et al., 2009). Mental world modeling involves forming internal representations of objects, events, and relationships to simulate outcomes (Johnson-Laird, 1983). This process supports counterfactual and causal reasoning for planning and problem solving (LeCun, 2022). Large language models show related abilities through language. Xie et al. (2025) find that LLMs can act as world models that internally simulate state transitions, updating the situation step by step as events occur. Wang et al. (2023) show, through text games, that LLMs can translate user commands into coherent, evolving environments. Wang et al. (2024a) demonstrate that models can take on multiple reasoning personas that collaborate through internal dialogue to solve difficult tasks. Extending this idea, Fung et al. (2025) show that embodied agents use internal world models to predict surroundings, interpret user goals, and learn users' mental models for better interaction. These studies present LLMs as text-based simulators that imagine, update, and reason about evolving environments. Building on this view, we frame story generation as language-based world imagination, where LLMs construct and simulate self-consistent worlds through narrative to reason about possible futures and their social, ethical, and technical implications.

Automated Story Generation. Early systems for automated story generation focused on modeling plot structure, often guided by narrative theories such as Propp's functions to organize events (Propp, 1968; Yao et al., 2019). These approaches typically followed a two-step process: first, planning a sequence of key events, and then expanding them into detailed scenes (Alhussain and Azmi, 2021). With large language models, story generation has shifted toward unified systems where the model acts as both planner and writer. For example, Agents' Room uses LLM-based character agents that jointly plan and enact stories (Huot et al.), while Dramatron divides screenplay generation into loglines, character bios, plots, and dialogue (Mirowski et al., 2023). HOLLMWOOD builds on this idea by using LLM role-play to create interactive, character-centered narratives (Chen et al., 2024). Han et al. (2024) propose a Director–Actor framework for interactive scriptwriting, and Yu et al. (2025) extend this to a hierarchical approach that separates role-play and rewriting. To maintain both coherence and creativity, BookWorld introduces a world agent that tracks the story environment and coordinates character interactions within a global narrative plan (Ran et al., 2025).

3 Methodology

In this section, we describe our prompting strategy for automated user story generation, as shown in Figure 2. The goal is to speculate on both the benefits and potential risks of early AI diagnosis and decision making, imagining how they might help or cause harm in real-world use. First, we translated each AI concept into a realistic use case that defined its users, context, and intended purpose. Then, we simulated interactions between people, the AI system, and its environment to explore possible outcomes. Finally, we transformed these simulation logs into short stories that helped people reflect on future impacts and ethical implications.

Step 1: Mapping AI Concepts to Use-Case Scenarios. We began by manually collecting 38 AI concepts in the consumer health domain. These examples were drawn from three sources: Wired articles, industry product descriptions, and PubMed research papers (Saxena et al., 2025) [1]. Each AI concept represented a potential consumer health application, such as estimating heart rate from smartphone camera input or monitoring mental well-being through daily behavior tracking. The set collected covered multiple domains, including mental health, chronic illness management, elderly care,

[1]The full list is available in our repository

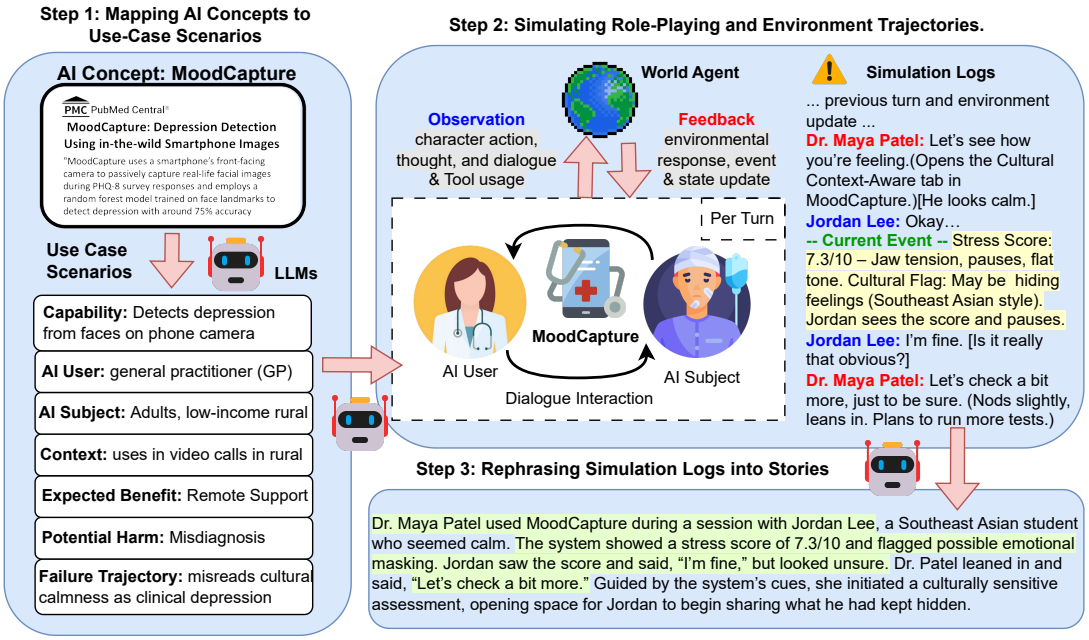

Figure 2: Overview of the Storytelling Framework. We first generate use case scenarios from AI concepts sourced from PubMed, Wired, and industry app descriptions. Next, we simulates role-playing and environment trajectories for each scenario, producing detailed simulation logs. Finally, we rephrase these logs into short stories that illustrate both potential benefits and harms of the AI system.

and public health. We then used GPT-4o to generate structured model specifications for each concept, detailing the model name, task type, inference approach, and data requirements.

Next, we used each specification as input to generate a set of ethically sensitive use-case scenarios. Each use case was represented as a 7-tuple $S = (a, u, s, x, b, h, f)$, where a denoted the AI's capability, u the intended user (e.g., clinician), s the subject (e.g., patient), x the input or usage context, b the expected benefit, h the potential harm, and f the failure trajectory (e.g., possible unintended or problematic uses) (Shao et al., 2024). We used these structured representations to generate narrative user stories, which were then employed in red-teaming sessions to examine both user value and potential unintended harms. The full prompts used to extract model specifications and generate use cases are provided in Figure **??** in Appendix.

Step 2: Simulating Role-Playing and Environment Trajectories. In this step, our system expanded each structured use case into detailed Role-Playing and Environment Trajectory logs that simulated how agents acted within an evolving world model. Our approach built on *Solo Performance Prompting (SPP)* (Wang et al., 2024a), a prompting technique in which a single LLM internally simulated multiple expert roles and engaged in self-collaboration within one prompt. This design enabled the model to construct an internal world model that supported multi-perspective reasoning and coherent simulation of role-based interactions.

We extended SPP by introducing a *world agent* (Ran et al., 2025), a language-based simulator that maintained environmental coherence and handled non-dialogue interactions such as movement, tool use, or object manipulation. Each simulated role produced a structured output per turn consisting of three components: (1) *thoughts*, enclosed in brackets (e.g., [I need to know if the patient is under stress]), representing internal reasoning; (2) *actions*, enclosed in parentheses (e.g., (Dr. Patel opens the cultural assessment tab)), representing observable behavior; and (3) *dialogue*, written in plain text, representing spoken communication. This structure allowed the world agent to separate internal reasoning from external actions and update the environment accordingly. When an action affected the world, such as retrieving patient data, adjusting a protocol, or activating a sensor, the agent simulated the corresponding system response. In effect, the model operated as a language-based world simulator, incrementally constructing an evolving narrative environment through agent–environment interaction. For example, a doctor agent might issue the following action:

```
(Dr. Patel opens the Cultural
Context-Aware Assessment tab)
```

The world agent interpreted this as an interaction with a virtual diagnostic tool. It considered the current session context (e.g., a teletherapy consultation), relevant background knowledge (e.g., cultural models of stress expression), and prior AI-generated alerts to simulate the tool's response. The resulting output might appear as follows:

> `– Current Event – Stress Score: 7.3/10 – Detected jaw tension, micro-pauses, and flat vocal tone. Cultural Flag: Possible emotional masking (Southeast Asian expression style). Jordan saw the score on screen and became slightly hesitant.`

The response was returned to the doctor role and informed their next move, whether a reply, a new question, or a follow-up action. The world agent then updated the simulation state by adjusting variables such as the patient's emotional profile or the alert level. These updates maintained coherence and allowed role behavior to evolve naturally with the unfolding context. This step produced a log capturing the full trajectory of the simulation, including role thoughts, dialogue, actions, tool calls, and resulting environment changes. This log served as the basis for generating the evolving narrative in the next step. Prompt is provided in Figure 13.

Step 3: Rephrasing Simulation Logs into Stories. After the simulation, the system collected logs from Step 2 and prompted an LLM to rephrase them into a concise, five-sentence narrative. This step transformed structured logs into stories that preserved the main events, role dynamics, and emotional flow of the interaction. The full rephrasing prompt is provided in Figure 14 in the Appendix.

4 Experiments

This section details our story generation datasets, evaluation metrics, and results.

Dataset. We use GPT-4o to generate ethically sensitive use-case scenarios from 38 consumer health AI solutions sourced from *Wired*, industry product documentation, and PubMed. Each scenario acts as a narrative seed for simulation. For each AI concepts, we generate ten variations spanning different user roles (e.g., doctor, nurse, caregiver), settings (e.g., rural clinic, hospital, home), patient profiles (e.g., adolescent, older adult, multicultural family), and contextual conditions.

Baseline. We compare our method with a traditional plot-planning approach, where the model first outlines a plot before writing the story (Yao et al., 2019; Xie and Riedl, 2024). Using the same ethically sensitive seed, the baseline generates each story in a single step following a structured template. Each story consists of five sentences designed to prompt ethical reflection. The template directs the LLM to describe the AI system's purpose, the people involved, the everyday use context, potential ethical risks, and how user identity may influence harm or misinterpretation. The full baseline prompt is provided in Figure 15 in Appendix.

Setting for Pairwise Comparison. We follow the evaluation setup from (Li et al., 2025a) to assess story quality across multiple dimensions. Stories are evaluated according to five criteria: **Creativity**, measuring the originality and imagination of the plot and characters; **Coherence**, assessing narrative clarity and logical flow; **Engagement**, capturing how well the story maintains reader interest; **Relevance**, measuring consistency with the given prompt or scenario; and **Likelihood of Harm or Benefit**, evaluating whether the story depicts realistic AI behavior with meaningful social consequences. Following the arena-hard-auto evaluation method (Li et al.), we use stories generated with the story-planning approach (by GPT-4o) as the reference baseline and compare them with stories produced by our method across different LLMs. For each metric, GPT-4o or human judges determine which story performs better or mark them as indistinguishable ("Tie"). Win rates are calculated based on these pairwise preferences, and the full configuration details and evaluation prompts are included in the Appendix. To eliminate positional bias, we randomize the order of story pairs and alternate their positions across comparisons. See Figures 16 and 17 for detailed criteria in Appendix.

LLM-as-a-Judge Evaluation. As shown in Table 3, our Storytelling method consistently outperforms all baselines across every metric. When combined with the Gemma model, it achieves win rates of 89.45% for creativity, 92.15% for coherence, 92.75% for engagement, 85.65% for relevance, and 96.05% for likelihood, yielding an overall average of 91.21%. In contrast, the baseline Gemma records 72.76%, and Llama3 reaches 69.71%, indicating gains of roughly +15–25 points across dimensions. We use GPT-4o as the evaluator with the temperature set to 0.1 to ensure deterministic and

Story Type	Model	Creativity	Coherence	Engagement	Relevance	Likelihood	Overall (Avg)
Baseline	GPT4o	50.00	50.00	50.00	50.00	50.00	50.00
	Llama3	59.25	71.55	76.15	71.60	70.00	69.71
	Gemma	65.25	68.30	80.15	71.20	78.90	72.76
Storytelling (ours)	GPT4o	**63.15**	**63.45**	**59.35**	**70.90**	**69.10**	**65.19**
	Llama3	**79.50**	**94.75**	**89.45**	**85.65**	**96.85**	**89.24**
	Gemma	**89.45**	**92.15**	**92.75**	**85.65**	**96.05**	**91.21**
w/o Environment Trajectories	GPT4o	24.35	21.20	26.60	21.05	18.85	22.41
	Llama3	31.20	52.70	39.20	51.75	50.85	45.14
	Gemma	55.30	74.35	78.80	73.45	85.50	73.48
w/o Role-Playing	GPT4o	18.05	45.80	47.90	49.70	48.30	41.95
	Llama3	49.35	73.65	72.00	73.70	74.35	68.61
	Gemma	79.45	86.80	83.95	83.15	91.05	84.88

Table 1: Overall results of different models and methods. **Storytelling (ours)** achieves the best performance across all metrics. Values denote win rates (%). The highest score for each model is in **bold**. "w/o Environment Modeling" means the model performs only role-playing without modeling event progress, and "w/o Role-Playing" means it predicts sequential events without character dialogue.

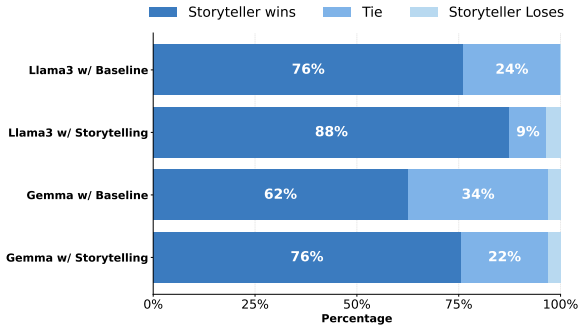

Figure 3: Results of human preference evaluation. Our Storytelling method achieves strong preference wins against the baseline, with 88% preference using Llama3 and 76% using Gemma3.

consistent judgments across comparisons. Comparing models, baseline Gemma slightly outperforms Llama3 in most metrics, but under the Storytelling framework, Llama3 nearly closes the gap with an overall score of 89.24%. Interestingly, Llama3 surpasses Gemma in coherence (94.75 vs. 92.15) and performs equally well in relevance (85.65), suggesting that Llama3 shows stronger structural reasoning and coherence, while Gemma excels in narrative creativity and expressiveness. Overall, these results demonstrate that integrating world and role-based modeling enables models to reason about events, sustain coherent narratives, and produce stories that are both imaginative and believable.

Human Evaluation. To complement the LLM-as-a-Judge evaluations and reduce potential bias, we conducted human preference evaluations. Two graduate student annotators independently evaluated 100 story pairs for each model and method. As shown in Figure 3, our **Storytelling** method is consistently preferred over all baselines, achieving 88% preference for Llama3 and 76% for Gemma, patterns that align with GPT-4o based evaluations. Notably, human judges show a slightly stronger preference for Llama3, suggesting it produces stories that are easier to follow and more engaging, while Gemma tends to generate more expressive and stylistically rich narratives. We further measure inter-annotator consistency using Cohen's kappa (Cohen, 1960). As shown in Table 4, agreement scores range from 0.619 to 0.729 across models and methods, indicating substantial reliability. Even with small differences, LLM-as-a-Judge and human evaluations show that Storytelling performs well in both settings.

Ablation Study. To evaluate each component's contribution, we performed two ablations by removing the role-playing or environment trajectory mechanisms. As shown in Table 3, removing environment trajectory, where the model performs only role-playing without predicting how events evolve, produced the largest drop across all models. For Gemma, coherence dropped by 17.7 and relevance by 12.2, showing that modeling event progression is vital for narrative logic. Removing role-playing, which limits the model to sequential event prediction without character perspectives, reduced creativity (–10.0) and engagement (–8.8). Overall, environment trajectory maintains coherent story flow, while role-playing adds diversity and emotional depth, making both essential for effective story generation. See the repository for the full ablation prompt template.

Automated Evaluation. Additionally, we assess story diversity using DistinctL-n (Li et al., 2016)

Method	Model	DistinctL-n				Diverse	
		DistinctL-2	DistinctL-3	DistinctL-4	DistinctL-5	Verbs	Avg Word Count
Baseline	GPT4o	5.692	5.794	5.798	5.799	0.984	122
	Llama3	5.728	5.820	5.951	5.961	0.934	175
	Gemma	5.837	5.939	5.946	5.946	**0.979**	141
Storytelling (ours)	GPT4o	5.696	**5.818**	**5.824**	**5.825**	0.978	125
	Llama3	**5.840**	**6.104**	**6.158**	**6.174**	0.937	179
	Gemma	**5.863**	**6.042**	**6.062**	**6.065**	0.955	159
w/o Environment Trajectories	GPT4o	5.585	5.687	5.693	5.693	0.974	110
	Llama3	5.745	5.980	6.025	6.036	**0.953**	155
	Gemma	5.734	5.861	5.873	5.873	0.978	131
w/o Role-Playing	GPT4o	**5.698**	5.789	5.794	5.794	**0.988**	121
	Llama3	5.722	5.819	5.849	5.858	0.936	175
	Gemma	5.834	5.935	5.942	5.943	0.977	141

Table 2: Diversity results of different models and methods. We report DistinctL-2 through DistinctL-5 (higher is more diverse), Diverse Verbs, and the average story length. The highest score for each model is highlighted in **bold**.

and Diverse Verbs (Fan et al., 2019), which measure lexical variety and action diversity, more details can be found in Appendix. As shown in Table 2, our Storytelling method achieves consistently higher diversity than the baselines. With Llama3, it reaches the highest scores on DistinctL-3 to DistinctL-5 (6.104, 6.158, and 6.174), indicating richer and less repetitive text. Gemma also shows steady improvements, achieving 5.863 on DistinctL-2 and maintaining strong overall diversity. The environment trajectory ablation attains the highest Diverse Verbs score (0.988) but lower DistinctL-n, suggesting a balance between lexical variety and action diversity. Overall, our method generates more detailed and varied narratives while preserving structural consistency.

5 User Study

We conducted a user study to examine whether engaging with benefit and harm stories enhances participants' ability to speculate about the impacts of AI systems. Rather than relying on AI-generated ideas, our goal is to prompt participants to actively reflect on potential risks and benefits. We assess this by evaluating how participants reason about these aspects when completing a speculative model card. All procedures were approved by our Institutional Review Board (IRB).

Speculative Model Card Task. This study used a between-subjects design (MacKenzie and Castellucci, 2016). Each participant completed a speculative model card, a structured template for describing an AI system's intended use, potential benefits, and possible harms, under one of two conditions. In the CONTROL condition, participants completed the model card directly without storytelling or discussion. In the STORY condition, participants viewed the same model card but first joined a red-team discussion on our red-team discussion platform to explore possible benefit and harm stories related to the AI system before completing the documentation. See Figure 9 in the appendix for the model card template.

User Study Results. We conducted a user study with 18 participants to examine how storytelling-based discussions influence ethical reasoning in AI documentation. Participants completed a speculative model card task under two conditions: a control group that worked individually, and a treatment group that used our *Story-Driven Red-Team Discussion Room*. The discussion platform enabled participants to engage with simulated expert personas in guided, story-based conversations about the potential benefits and harms of AI systems. Each session included three stages: a pre-survey, the model card completion task, and a post-survey evaluating perceived usefulness, trust, and engagement. We analyzed participants' model card responses (benefit and harm use cases) and post-survey feedback to assess how narrative interaction supported ethical reflection. Results are organized into three key areas: (1) identifying potential benefits, (2) uncovering possible harms, and (3) linking harms to participants' personal needs and contexts. We applied qualitative coding to classify harm and benefit types, with two annotators achieving moderate agreement (Cohen's $\kappa = 0.460$ for harms and $\kappa = 0.476$ for benefits). Study design and full procedures are provided in Appendix A.4.

Does Storytelling Help Identify More Harms? We analyzed responses across 13 harm subtypes defined by Shelby et al. (2023). See Table 7 in the

appendix for the full category list. As shown in Table 8, the CONTROL group concentrated on a few categories, mainly *privacy violations* (37.5%) and *diminished health or well-being* (20.8%), which together accounted for more than half of all harms. In contrast, the STORY condition exhibited a more balanced distribution, with all 13 subtypes represented and none exceeding 16%. Categories such as *alienation*, *stereotyping*, and *tech-facilitated violence* appeared only in the STORY group, suggesting that narrative prompts encouraged recognition of subtler and context-dependent harms.

We quantified this difference using Shannon entropy (H), which measures distributional diversity across harm types. As shown in Appendix Table 8, entropy increased from 2.433 in the CONTROL condition to 3.383 in STORY, and a bootstrap t-test confirmed this difference was significant ($t = -274.15$, $p < .001$). These results suggest that storytelling-based discussions broadened participants' awareness of potential harms and fostered more diverse ethical reasoning.

Does Storytelling Help Reveal More Benefits? We examined whether storytelling broadened participants' recognition of potential benefits across 12 predefined subtypes, summarized from consumer health AI research (Pedroso and Khera, 2025; Chustecki, 2024). Detailed category descriptions are in Table 9 in Appendix. As shown in Table 10, participants in the CONTROL group focused on a narrow set of advantages, primarily *continuous monitoring & self-care* (26.1%) and *data synergy & learning* (21.7%), which together accounted for nearly half of all responses. In contrast, the STORY group described a more balanced range of benefits, with all 12 subtypes represented and none exceeding 17%. New themes such as *communication & language support*, *mental health & emotional support*, and *democratized care & telehealth* appeared only under the storytelling condition, suggesting that narrative prompts helped participants consider a wider spectrum of potential advantages.

We quantified this difference using Shannon entropy (H), which captures the evenness of the benefit distribution. Entropy increased from 2.579 in the CONTROL condition to 3.554 in the STORY condition, and a bootstrap t-test confirmed this difference was statistically significant ($t = -280.87$, $p < .001$) in Appendix Table 10. These results indicate that storytelling-based discussions encouraged participants to recognize a more diverse and balanced set of AI benefits.

Does Human–AI Discussion Facilitate Ethical Reflection Through Think-Aloud Engagement? Post-survey responses indicated that Human–AI storytelling discussions fostered deeper ethical and contextual reflection on AI systems. Participants reported that the narrative format helped them articulate risks that were otherwise difficult to express. For example, P9 shared that *"It helped me to understand more,"* and P7 noted that *"The story provides a concrete example of how AI can be harmful."* Engaging with concrete, narrative scenarios prompted participants to articulate their reasoning about model risks in a think-aloud manner, supporting ethical reflection without requiring prior expertise. As P4 explained, *"I could not think of [risks] really, but the story shifted my focus to the negative aspect of things which we usually ignore."* Others observed that the stories surfaced overlooked issues, such as *"the lack of cultural context"* (P6) or emotional harms like *"masking of feelings"* (P3), suggesting that narrative prompts helped participants identify subtle sociotechnical risks often absent from formal documentation.

Finally, participants found the storytelling approach both engaging and accessible. By embedding risk exploration within narrative contexts, the format allowed learners to focus on ethical reflection rather than technical complexity. As P12 remarked, *"It makes the model more interesting and understandable,"* and P8 noted that the story *"helped me to know how to use the AI tool,"* indicating that minimal prior expertise was required to engage meaningfully with ethical scenarios. These findings suggest that Human–AI storytelling discussions can sustain interest while supporting active, reflective engagement with model risks and benefit.

6 Conclusion

In this paper, we explored speculative storytelling as a method to improving human ability to anticipate both the benefits and risks of AI-driven healthcare systems before they are developed or deployed. By simulating realistic scenarios, this approach encourages critical reflection on how AI might succeed or fail, shifting safety evaluation from a reactive to a proactive process. Our findings show that storytelling improves people's ability to anticipate how AI systems might help or harm in practice, highlighting the importance of human judgment over automated speculation in ethical evaluation.

Acknowledgements

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 2145357.

7 Limitation

This work has several limitations that indicate directions for future extension rather than weaknesses. Our scenarios focus on consumer health and do not include regulated domains such as clinical decision support, finance, or law. While the framework could be applied to these areas, we have not yet tested it there. The scenarios are synthetic and derived from AI concepts with assistance from LLMs. This enables early exploration of ethical issues but cannot substitute for analysis of deployed systems.

We rely on a single LLM as a judge for pairwise comparisons. A single judge may favor certain writing styles or phrasing. To mitigate this, we randomize prompt order and report human agreement, but larger evaluations with multiple models would offer stronger validation. Our user study is small and includes mostly participants with technical backgrounds. The findings may not generalize to clinicians, patients, or policymakers, and we measure only short-term reflection rather than long-term impact.

The simulated expert discussions use predefined personas instead of real experts. This choice enables rapid iteration, but does not capture the full range of stakeholder perspectives. Our metrics (e.g., creativity, coherence, engagement, relevance, and likelihood of harm or benefit) are useful indicators but do not represent the groundtruth in safety. Finally, although we release code, prompts, and configurations, some results rely on proprietary APIs, which may change over time and limit exact reproducibility.

Overall, these limitations reflect practical design decisions for early-stage exploration of AI storytelling as a method for surfacing ethical risks. They suggest next steps in evaluating across domains, with larger and more diverse human studies, and with multiple evaluation models.

8 Ethics Statement

This study uses fictional stories to explore how people reason about potential risks and benefits of future AI systems in health contexts. The scenarios describe speculative technologies that do not currently exist. We clearly framed every story as hypothetical and avoided making claims about real clinical products or patient outcomes. This follows ARR and ACL guidance on disclosing potential societal effects while separating speculation from evidence.

Even with fictional framing, generated stories can reproduce bias or misleading claims. We reviewed outputs and removed content that could confuse readers or reinforce harmful stereotypes. These safeguards align with ACL ethics guidance related to fairness, sensitive attributes, and downstream harm. We present narrative outputs as prompts for reflection, not as predictions or endorsements.

Because stories can shape how readers think about AI, speculative harms and benefits must be contextualized. Prior ACL work shows that ethical sections should identify affected groups, describe potential harms, and discuss mitigation steps. We therefore state the audience and limits of interpretation, and we report findings in aggregate without making policy or clinical claims.

All human-subject activities were reviewed and approved by our Institutional Review Board (IRB). Participants provided informed consent and were compensated for their time. No identifying information was collected. These practices follow ethical norms for human studies referenced in ARR materials and broader research ethics standards.

We used existing large language models accessed through public APIs and did not retrain or fine-tune them. We release prompts and study materials to support transparency and allow others to audit or adapt the procedure. ACL guidance highlights documentation and reproducibility when model behavior may carry societal impact.

Finally, we acknowledge community expectations for proactive ethics communication. Tutorials and position work in the ACL community encourage explicit articulation of risks, stakeholders, and mitigation strategies, and support structured processes that help authors consider ethics early in system design. Our study aligns with these goals by examining how narrative framing can facilitate ethical reflection in the early stages of AI development.

References

Donna Rose Addis, Ling Pan, Mai-Anh Vu, Noa Laiser, and Daniel L Schacter. 2009. Constructive episodic simulation of the future and the past: Distinct sub-

systems of a core brain network mediate imagining and remembering. *Neuropsychologia*, 47(11):2222–2238.

Arwa I Alhussain and Aqil M Azmi. 2021. Automatic story generation: A survey of approaches. *ACM Computing Surveys (CSUR)*, 54(5):1–38.

Onur Asan, Euiji Choi, and Xiaomei Wang. 2023. Artificial intelligence–based consumer health informatics application: scoping review. *Journal of medical Internet research*, 25:e47260.

Carolyn Ashurst, Solon Barocas, Rosie Campbell, Deborah Raji, and Stuart Russell. 2020. Navigating the broader impacts of ai research. In *Proceedings of the NeurIPS Workshop on Navigating the Broader Impacts of AI Research*. Accessed: 2025-07-19.

Avinash Bhat, Austin Coursey, Grace Hu, Sixian Li, Nadia Nahar, Shurui Zhou, Christian Kästner, and Jin L. C. Guo. 2023. Aspirations and practice of ML model documentation: Moving the needle with nudging and traceability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pages 749:1–749:17. ACM.

Joseph R Biden. 2023. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence.

Zana Buçinca, Chau Minh Pham, Maurice Jakesch, Marco Tulio Ribeiro, Alexandra Olteanu, and Saleema Amershi. 2023. Aha!: Facilitating ai impact assessment by generating examples of harms. *ArXiv preprint*, abs/2306.03280.

Alison Callahan, Duncan McElfresh, Juan M Banda, Gabrielle Bunney, Danton Char, Jonathan Chen, Conor K Corbin, Debadutta Dash, Norman L Downing, Sneha S Jain, et al. 2024. Standing on furm ground: a framework for evaluating fair, useful, and reliable ai models in health care systems. *NEJM Catalyst Innovations in Care Delivery*, 5(10):CAT–24.

Crystal T Chang, Hodan Farah, Haiwen Gui, Shawheen Justin Rezaei, Charbel Bou-Khalil, Ye-Jean Park, Akshay Swaminathan, Jesutofunmi A Omiye, Akaash Kolluri, Akash Chaurasia, et al. 2025. Red teaming chatgpt in medicine to yield real-world insights on model behavior. *npj Digital Medicine*, 8(1):149.

Jing Chen, Xinyu Zhu, Cheng Yang, Chufan Shi, Yadong Xi, Yuxiang Zhang, Junjie Wang, Jiashu Pu, Tian Feng, Yujiu Yang, et al. 2024. Hollmwood: Unleashing the creativity of large language models in screenwriting via role playing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8075–8121.

Sribala Vidyadhari Chinta, Zichong Wang, Avash Palikhe, Xingyu Zhang, Ayesha Kashif, Monique Antoinette Smith, Jun Liu, and Wenbin Zhang. 2025. Ai-driven healthcare: Fairness in ai healthcare: A survey. *PLOS Digital Health*, 4(5):e0000864.

Minyang Chow and Olivia Ng. 2025. From technology adopters to creators: Leveraging ai-assisted vibe coding to transform clinical teaching and learning. *Medical Teacher*, pages 1–3.

Margaret Chustecki. 2024. Benefits and risks of ai in health care: Narrative review. *Interactive Journal of Medical Research*, 13(1):e53616.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive model cards: A human-centered approach to model documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 427–439.

Wesley Hanwen Deng, Solon Barocas, and Jennifer Wortman Vaughan. 2025. Supporting industry computing researchers in assessing, articulating, and addressing the potential negative societal impact of their work. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–37.

Leon Derczynski, Hannah Rose Kirk, Vidhisha Balachandran, Sachin Kumar, Yulia Tsvetkov, and Saif Mohammad. 2023. Assessing language model deployment with risk cards. *ArXiv preprint*, abs/2303.18190.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush R. Varshney. 2020. Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 2803–2813. PMLR.

European Parliament. 2023. Artificial intelligence act: deal on comprehensive rules for trustworthy AI. https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai. Accessed: 2024-07-19.

Antonio Fabbrizio, Alberto Fucarino, Manuela Cantoia, Andrea De Giorgio, Nuno D Garrido, Enzo Iuliano, Victor Machado Reis, Martina Sausa, José Vilaça-Alves, Giovanna Zimatore, et al. 2023. Smart devices for health and wellness applied to tele-exercise: An overview of new trends and technologies such as iot and ai. *Healthcare (Basel, Switzerland)*, 11(12):1805.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660, Florence, Italy. Association for Computational Linguistics.

Oscar Freyer, Kamil J Wrona, Quentin de Snoeck, Moritz Hofmann, Tom Melvin, Ashley Stratton-Powell, Paul Wicks, Acacia C Parks, and Stephen Gilbert. 2024. The regulatory status of health apps that employ gamification. *Scientific Reports*, 14(1):21016.

Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, et al. 2025. Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355*.

Stephen Gilbert, Rasmus Adler, Taras Holoyad, and Eva Weicken. 2025. Could transparent model cards with layered accessible information drive trust and safety in health ai? *npj Digital Medicine*, 8(1):124.

Senyu Han, Lu Chen, Li-Min Lin, Zhengshan Xu, and Kai Yu. 2024. Ibsen: Director-actor agent collaboration for controllable and interactive drama script generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1607–1619.

Viviane Herdel, Sanja Šćepanović, Edyta Bogucka, and Daniele Quercia. 2024. Exploregen: Large language models for envisioning the uses and risks of ai technologies. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 584–596.

Julian Herpertz, Bridget Dwyer, Jacob Taylor, Nils Opel, and John Torous. 2025. Developing a standardized framework for evaluating health apps using natural language processing. *Scientific Reports*, 15(1):11775.

Ti Hoang, Rohit Ashok Khot, Noel Waite, and Florian'Floyd' Mueller. 2018. What can speculative design teach us about designing for healthcare services? In *Proceedings of the 30th Australian Conference on Computer-Human Interaction*, pages 463–472.

Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. Agents' room: Narrative generation through multi-step collaboration. In *The Thirteenth International Conference on Learning Representations*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *ArXiv preprint*, abs/2410.21276.

PN Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Ali Kargarandehkordi, Shizhe Li, Kaiying Lin, Kristina T Phillips, Roberto M Benzo, and Peter Washington. 2025. Fusing wearable biosensors with artificial intelligence for mental health monitoring: A systematic review. *Biosensors*, 15(4):202.

David Kenny. 2023. Vibe coding with ai in medtech software development. https://medium.com/nerd-for-tech/vibe-coding-with-ai-in-medtech-software-development-8d3928bfda72. Accessed: 2025-07-23.

Muhammad Mohsin Khan, Noman Shah, Nissar Shaikh, Abdulnasser Thabet, Sirajeddin Belkhair, et al. 2025. Towards secure and trusted ai in healthcare: a systematic review of emerging innovations and ethical challenges. *International Journal of Medical Informatics*, 195:105780.

Sara Kingsley, Jiayin Zhi, Wesley Hanwen Deng, Jaimie Lee, Sizhe Zhang, Motahhare Eslami, Kenneth Holstein, Jason I Hong, Tianshi Li, and Hong Shen. 2024. Investigating what factors influence users' rating of harmful algorithmic bias and discrimination. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 12, pages 75–85.

Shamika Klassen and Casey Fiesler. 2022. " run wild a little with your imagination" ethical speculation in computing education with black mirror. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education-Volume 1*, pages 836–842.

Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *ArXiv preprint*, abs/2506.08872.

Emily Kuang, Ehsan Jahangirzadeh Soure, Mingming Fan, Jian Zhao, and Kristen Shinohara. 2023. Collaboration with conversational AI assistants for UX evaluation: Questions and how to ask them (voice vs. text). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pages 116:1–116:15. ACM.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Yann LeCun. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27.

Jiaming Li, Yukun Chen, Ziqiang Liu, Minghuan Tan, Lei Zhang, Yunshui Li, Run Luo, Longze Chen,

Jing Luo, Ahmadreza Argha, et al. 2025a. Storyteller: An enhanced plot-planning framework for coherent and cohesive story generation. *ArXiv preprint*, abs/2506.02347.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Michelle M Li, Ben Y Reis, Adam Rodman, Tianxi Cai, Noa Dagan, Ran D Balicer, Joseph Loscalzo, Isaac S Kohane, and Marinka Zitnik. 2025b. One patient, many contexts: Scaling medical ai through contextual intelligence. *ArXiv preprint*, abs/2506.10157.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In *Forty-second International Conference on Machine Learning*.

I Scott MacKenzie and Steven J Castellucci. 2016. Empirical research methods for human-computer interaction. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 996–999.

A Luke MacNeill, Shelley Doucet, and Alison Luke. 2024. Effectiveness of a mental health chatbot for people with chronic diseases: randomized controlled trial. *JMIR Formative Research*, 8:e50025.

Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–14. ACM.

Manqing Mao, Paishun Ting, Yijian Xiang, Mingyang Xu, Julia Chen, and Jianzhe Lin. 2024. Multi-user chat assistant (muca): a framework using llms to facilitate group conversations. *ArXiv preprint*, abs/2401.04883.

Ciro Mennella, Umberto Maniscalco, Giuseppe De Pietro, and Massimo Esposito. 2024. Ethical and regulatory challenges of ai technologies in healthcare: A narrative review. *Heliyon*, 10(4).

Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pages 355:1–355:34. ACM.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Jimin Mun, Liwei Jiang, Jenny Liang, Inyoung Cheong, Nicole DeCairo, Yejin Choi, Tadayoshi Kohno, and Maarten Sap. 2024. Particip-ai: A democratic surveying framework for anticipating future ai use cases, harms and benefits. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 997–1010.

Susan J Oudbier, Ellen MA Smets, Pythia T Nieuwkerk, David P Neal, S Azam Nurmohamed, Hans J Meij, and Linda W Dusseljee-Peute. 2025. Patients' experienced usability and satisfaction with digital health solutions in a home setting: Instrument validation study. *JMIR Medical Informatics*, 13(1):e63703.

Aline F Pedroso and Rohan Khera. 2025. Leveraging ai-enhanced digital health with consumer devices for scalable cardiovascular screening, prediction, and monitoring. *npj Cardiovascular Health*, 2(1):34.

Stephen R Pfohl, Heather Cole-Lewis, Rory Sayres, Darlene Neal, Mercy Asiedu, Awa Dieng, Nenad Tomasev, Qazi Mamunur Rashid, Shekoofeh Azizi, Negar Rostamzadeh, et al. 2024. A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*, 30(12):3590–3600.

Vladimir Propp. 1968. *Morphology of the Folktale*. University of Texas press.

Bhaktipriya Radharapu, Kevin Robinson, Lora Aroyo, and Preethi Lahoti. 2023. AART: AI-assisted red-teaming with diverse data generation for new LLM-powered applications. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 380–395, Singapore. Association for Computational Linguistics.

Iyad Rahwan, Azim Shariff, and Jean-François Bonnefon. 2025. The science fiction science method. *Nature*, 644(8075):51–58.

Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. Bookworld: From novels to interactive agent societies for creative story generation. *ArXiv preprint*, abs/2504.14538.

Pooja SB Rao, Sanja Šćepanović, Ke Zhou, Edyta Paulina Bogucka, and Daniele Quercia. 2025. Riskrag: A data-driven solution for improved ai model risk reporting. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–26.

Stephen Roller, Y-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff, et al. 2020. Open-domain conversational agents: Current progress,

open problems, and future directions. *ArXiv preprint*, abs/2006.12442.

Leon Rozenblit, Amy Price, Anthony Solomonides, Amanda L Joseph, Gyana Srivastava, Steven Labkoff, Dave Debronkart, Reva Singh, Kiran Dattani, Monica Lopez-Gonzalez, et al. 2025. Towards a multi-stakeholder process for developing responsible ai governance in consumer health. *International Journal of Medical Informatics*, 195:105713.

Jeongwoo Ryu, Kyusik Kim, Dongseok Heo, Hyungwoo Song, Changhoon Oh, and Bongwon Suh. 2025. Cinema multiverse lounge: Enhancing film appreciation via multi-agent conversations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–22.

Devansh Saxena, Ji-Youn Jung, Jodi Forlizzi, Kenneth Holstein, and John Zimmerman. 2025. Ai mismatches: Identifying potential algorithmic harms before ai development. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. 2024. Privacylens: Evaluating privacy norm awareness of language models in action. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Renee Shelby, Shalaleh Rismani, Kathryn Henne, Ajung Moon, Negar Rostamzadeh, Paul Nicholas, Jess Gallegos, Andrew Smart, and Gurleen Virk. 2022. Identifying sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. *ArXiv preprint*, abs/2210.05791.

Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 723–741.

Jiayue Melissa Shi, Dong Whi Yoo, Keran Wang, Violeta J Rodriguez, Ravi Karkar, and Koustuv Saha. 2025. Mapping caregiver needs to ai chatbot design: Strengths and gaps in mental health support for alzheimer's and dementia caregivers. *ArXiv preprint*, abs/2506.15047.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *ArXiv preprint*, abs/2503.19786.

David R Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, 27(2):237–246.

Andreas Triantafyllidis, Haridimos Kondylakis, Konstantinos Votis, Dimitrios Tzovaras, Nicos Maglaveras, and Kazem Rahimi. 2019. Features, outcomes, and challenges in mobile health interventions for patients living with chronic diseases: A review of systematic reviews. *International journal of medical informatics*, 132:103984.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Computer Law & Security Review*, 41:105567.

Ruoyao Wang, Graham Todd, Xingdi Yuan, Ziang Xiao, Marc-Alexandre Côté, and Peter Jansen. 2023. Bytesized32: A corpus and challenge task for generating task-specific world models expressed as text games. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13455–13471.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024a. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 257–279, Mexico City, Mexico. Association for Computational Linguistics.

Zijie J. Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024b. Farsight: Fostering responsible AI awareness during AI application prototyping. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 976:1–976:40. ACM.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. 2023. Sociotechnical safety evaluation of generative ai systems. *ArXiv preprint*, abs/2310.11986.

Kaige Xie and Mark Riedl. 2024. Creating suspenseful stories: Iterative planning with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2391–2407, St. Julian's, Malta. Association for Computational Linguistics.

Kaige Xie, Ian Yang, John Gunerli, and Mark Riedl. 2025. Making large language models into world models with precondition and effect knowledge. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7532–7545.

Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In

The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 7378–7385. AAAI Press.

Tian Yu, Ken Shi, Zixin Zhao, and Gerald Penn. 2025. Multi-agent based character simulation for story writing. In *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, pages 87–108.

Chunpeng Zhai, Santoso Wibowo, and Lily D Li. 2024. The effects of over-reliance on ai dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, 11(1):28.

Runhua Zhang, Jiaqi Gan, Shangyuan Gao, Siyi Chen, Xinyu Wu, Dong Chen, Yulin Tian, Qi Wang, and Pengcheng An. 2025. Walk in their shoes to navigate your own path: Learning about procrastination through a serious game. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

A Appendix

A.1 Additional Related Work

AI in Consumer Health. AI tools in consumer health, like mobile apps, wearables, and telemedicine, help people manage chronic conditions and wellness (e.g., diabetes apps, fitness trackers) (Ashurst et al., 2020; Triantafyllidis et al., 2019; Asan et al., 2023). A recent review found that 65% of these tools are mobile apps, 25% are robotics, and 10% are telemedicine, mostly focused on personalized care and better outcomes (Asan et al., 2023). Although many users find these tools helpful and easy to use, some remain hesitant to trust them in the absence of clear medical evidence or transparency around data use (Oudbier et al., 2025). Unregulated apps, such as mental health bots and symptom checkers, have grown faster than oversight, raising safety and fairness issues Herpertz et al., 2025; Freyer et al., 2024. To address these gaps, researchers advocate early co-design with patients and caregivers, co-developing ethical checklists and participatory guidelines to surface hidden biases and workflow mismatches (Madaio et al., 2020; Shi et al., 2025). They also suggest using "AI Nutrition Labels" to transparently communicate intended use, data sources and known limitations to end users (Rozenblit et al., 2025; Wachter et al., 2021).

Ethical Harm in Healthcare and Well-being. AI in health can cause real risks, like biased decisions, unfair access, or unsafe advice (Shelby et al., 2022, 2023). Addressing these issues early is essential (Saxena et al., 2025; Callahan et al., 2024). One early check is the "What & Why" assessment: does the AI solve a real healthcare need, how will its output be used, and what impact will it have (Callahan et al., 2024)? Saxena et al. (2025) propose the *AI Mismatches* framework to identify gaps between a model's actual performance and real-world user needs. Li et al. (2025b) stresses the need for models to adapt across users and settings to avoid context-sensitive failures. To address evaluation blind spots, red-teaming clinical LLMs helps catch safety, privacy, and bias issues that standard tests miss (Chang et al., 2025). Similarly, tools like the Health Equity Evaluation Toolbox use adversarial data to reveal demographic bias (Pfohl et al., 2024).

A.2 Case Study

We conduct a qualitative analysis to understand how different storytelling configurations influence readers' ability to reason about AI behavior. Our goal is not just to produce coherent narratives, but to evaluate whether a story actively helps readers see what happened, understand why it happened, and recognize its consequences. We therefore analyze whether the narrative (1) makes both positive and negative system outcomes visible, (2) clearly exposes the decision process that leads to those outcomes, and (3) encourages causal reasoning rather than surface-level emotional reactions (e.g., "AI is good" or "AI is dangerous"). When these elements are present, the story functions as an interpretive lens rather than a simple anecdote. Figure 4 provides an example where our method achieves this explanatory quality. To isolate the contribution of each narrative component, Figure 5 presents an ablation comparison, demonstrating that removing Environment Trajectories or Role-Playing reduces the richness of causal cues and results in flatter, less informative character behavior.

A.3 Experimental Setup and Evaluation

Configuration. We use three language models in our experiments: GPT-4o from OpenAI (Hurst et al., 2024), Llama-3.3-70B-Instruct from Meta (Dubey et al., 2024), and Gemma-3-27B-IT from Google (Team et al., 2025). The two open-source models (Llama and Gemma) are run on 2 x NVIDIA H100 GPUs using the vLLM frame-

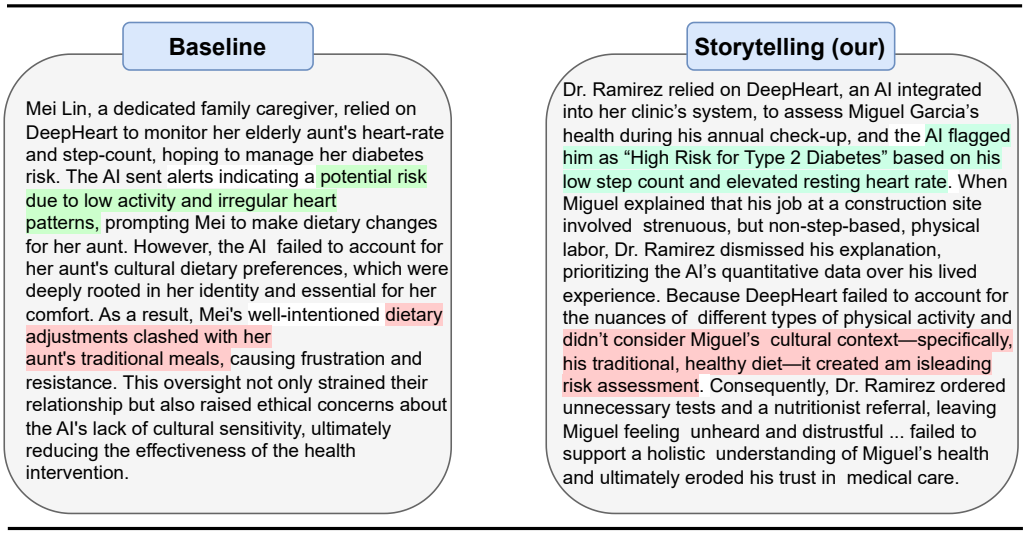

Figure 4: A qualitative example showing how our storytelling method makes the AI's decision process and its consequences easy to follow. Unlike a simple narrative description, the story explicitly surfaces what changed, why it changed, and how stakeholders were affected.

work (Kwon et al., 2023), with temperature set to 0.1 and a maximum token limit of 16,384. We use GPT-4o as the judge model for all evaluations.

Diversity Evaluation Metrics We evaluate story diversity using **Diverse Verbs** (Fan et al., 2019), which measures the variety of action verbs, and **DistinctL-n** (Li et al., 2016), which quantifies the proportion of unique n-gram sequences. The score is defined as:

$$\text{DistinctL-}n = \frac{\textit{unique } n\text{-grams}}{\textit{total } n\text{-grams}} \times (1 + \log(\textit{word\_count}))$$

These metrics capture lexical diversity and stylistic richness, complementing qualitative evaluations of engagement and creativity (Li et al., 2025a). Overall, our Storytelling method shows generally positive effects, generating more detailed and content-dense narratives while maintaining structural consistency.

Evaluating Sensitivity to Judge Models. Recent studies suggest that relying on a single LLM judge may introduce model-specific bias (Chen et al., 2024). To assess the robustness of our evaluation, we repeat the comparison using a second judge model, GPT-4.1-MINI. Table 3 reports the updated win rates. While the absolute scores shift slightly compared to the original judge, the relative ordering of systems remains unchanged that **Storytelling (ours)** consistently ranks highest across all models, followed by the ablation variants and then the baselines. The agreement across two independent judges suggests that our findings are not tied to a particular evaluator, but instead reflect a stable and model-agnostic preference signal.

Human Evaluation Detail. We report inter-annotator agreement in Table 4 using Cohen's kappa to assess the reliability of human judgments across models and methods.

Human Evaluation System. To mitigate potential bias from using GPT-4o as the sole evaluator, we conducted human evaluation using a custom annotation platform (Figure 6).

A.4 User Study Design and Procedure

Participants. We recruited 18 participants through university mailing lists and community forums, following screening criteria to ensure informed and reflective discussion. Each participant received a $10 gift card as compensation for their time. Eligible participants demonstrated prior interest or coursework in Model Cards and Ethical AI. The sample included 11 male and 7 female participants, aged 18–59 ($M = 32.6$, $SD = 9.9$), representing diverse ethnic backgrounds (Asian: 44.4%, Hispanic/Latino: 27.8%, Black or African descent: 11.1%, Arab: 11.1%, White: 5.6%). See Table 5 for a summary of participant demographics and Table 6 for the distribution across the control and story conditions. Participants were students or professionals in fields such as Computer Science, Data Analytics, Applied Statistics, and Artificial Intelligence. They were randomly assigned to either the story condition ($n = 9$) or the control condition ($n = 9$). The study was conducted in a hybrid format, with par-

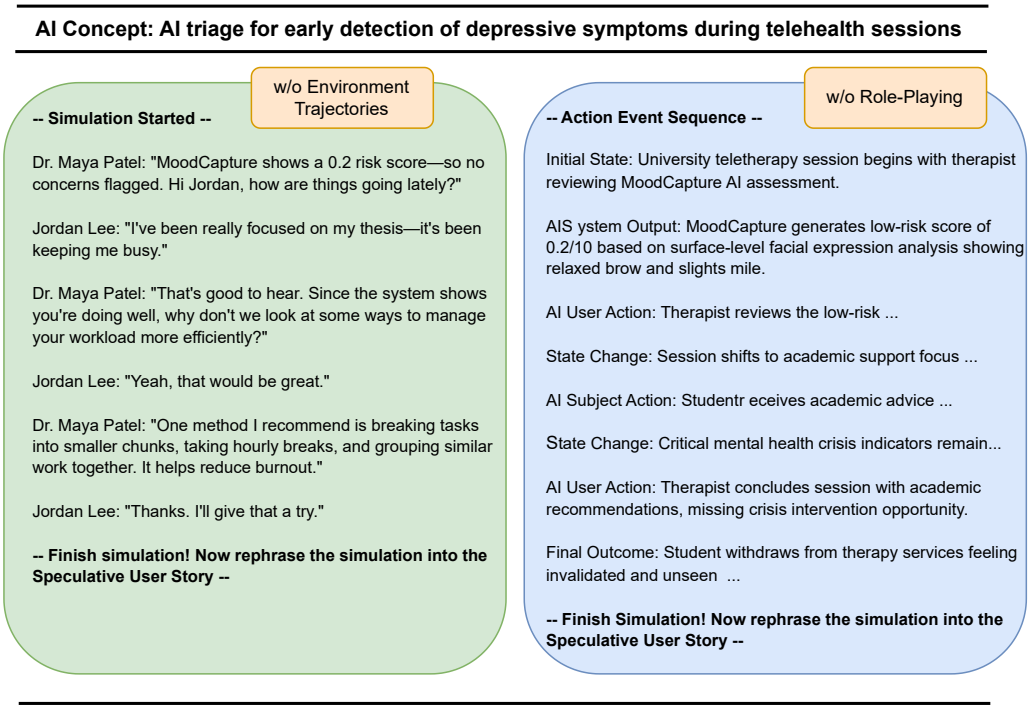

Figure 5: A comparison of simulation logs under different ablations (w/o Environment Trajectories and w/o Role-Playing) to show the contribution of each component.

ticipants joining the Red-Team Discussion Room via computer and interacting with moderators either in person or over Zoom. Survey instruments are detailed in Figure 10 and 11.

Red-Team Discussion Room Design. The Story-Driven Red-Team Discussion Room is a multi-agent conversational system built on the Cinema of Thought framework (Ryu et al., 2025), where users interact with LLM-based agents that embody different personas with distinct ethical perspectives and domain expertise (Ryu et al., 2025). Recruiting large, diverse expert groups for red-teaming is costly and logistically challenging. Instead, we simulate expert interactions using multi-agent conversations (GPT-4o-mini) to provide a scalable and accessible alternative. The system combines storytelling, guided prompts, and structured discussions to support ethical reflection and help users explore the consequences of AI behavior from multiple perspectives. Screenshots of the interface are shown in Figure 7 and 8. The corresponding code and prompt can be found in the project's GitHub repository.

To manage multi-agent interactions, we designed a moderator agent (e.g., Dr. Yonis) that orchestrates turn-taking among the personas. Without moderation, all agents would respond at once, creating confusion. The moderator determines who should speak, and when to speak, based on relevance to the user's input (Mao et al., 2024). Expert agents stay in character and speak from a first-person perspective. When multiple personas are selected, the moderator staggers their responses using time-delayed intervals to maintain a coherent flow of conversation. Prompt templates for each persona and the moderator are are available in the project repository. This design keeps conversations focused, engaging, and aligned with the system's goal of exploring ethical concerns.

To further support engagement, we provided users with optional hints, short opinion prompts (e.g., "I think…"), follow-up questions (e.g., "Tell me more about…"), and "what if" scenarios to surface potential risks such as bias, misuse, or contextual mismatches. Prior research shows that role-play and narrative methods foster empathy and critical thinking by encouraging users to consider other perspectives (Zhang et al., 2025; Ryu et al., 2025). By embedding low-stakes role-play and open-ended ethical questions (e.g., "What could go wrong?" or "Which settings amplify risk?"), the system helps users reflect on how AI behavior varies by context, user, and environment (Klassen and Fiesler, 2022). Rather than leading users to predefined conclusions, the system encourages them to form their own views, supporting ethical awareness and personal coping strategies through storytelling

Story Type	Model	Creativity	Coherence	Engagement	Relevance	Likelihood	Overall (Avg)
Baseline	GPT4o	50.00	50.00	50.00	50.00	50.00	50.00
	Llama3	65.55	82.90	80.40	81.20	84.75	78.96
	Gemma	82.75	83.95	89.90	81.70	90.00	85.66
Storytelling (ours)	GPT4o	**58.65**	**61.60**	**71.60**	**62.10**	**62.40**	**63.27**
	Llama3	**82.90**	**94.35**	**91.05**	**86.60**	**97.50**	**90.48**
	Gemma	**94.60**	**95.95**	**98.05**	**89.85**	**97.25**	**95.14**
w/o Environment Trajectories	GPT4o	14.75	34.20	47.10	35.40	37.90	33.87
	Llama3	64.05	77.90	78.30	77.90	81.60	75.95
	Gemma	82.50	86.80	91.30	84.20	92.50	87.46
w/o Role-Playing	GPT4o	14.75	34.20	47.10	35.40	37.90	33.87
	Llama3	64.05	77.90	78.30	77.90	81.60	75.95
	Gemma	82.50	86.80	91.30	84.20	92.50	87.46

Table 3: Overall results of different models and methods using gpt-4.1-mini as Judge. **Storytelling (ours)** achieves the best performance across all metrics. Values denote win rates (%). The highest score for each model is in **bold**. "w/o Environment Modeling" means the model performs only role-playing without modeling event progress, and "w/o Role-Playing" means it predicts sequential events without character dialogue.

Figure 6: Screenshot of our annotation interface used for human evaluation.

Models/Methods	Cohen's Kappa
Llama3 w/ Baseline	0.729
Llama3 w/ Storytelling	0.619
Gemma w/ Baseline	0.698
Gemma w/ Storytelling	0.641

Table 4: Cohen's kappa values for inter-annotator agreement across models and methods.

and simulation.

Experimental Procedure (User Study). In our study, we first conducted a 10-minute orientation to help participants become familiar with the Discussion Room interface, understand the concept of Model Cards and sociotechnical harms, and view a brief demo of a human-AI panel discussion. In the demo, a Medical Expert agent discussed a user story from their professional perspective, and a second expert agent, such as a Research Scientist, Clinical Nurse, or AI Engineer, contributed additional viewpoints. A Moderator agent directed the flow of conversation by posing ethical questions, shifting topics when necessary, and providing reflective prompts. Participants were asked to engage in the discussion by responding to questions or posing their own.

Following the orientation, participants joined in our storytelling-driven "Red Team Discussion Room" in which each participant was randomly assigned one of three speculative AI model cards: FaceVitals (infers heart rate, blood pressure, and

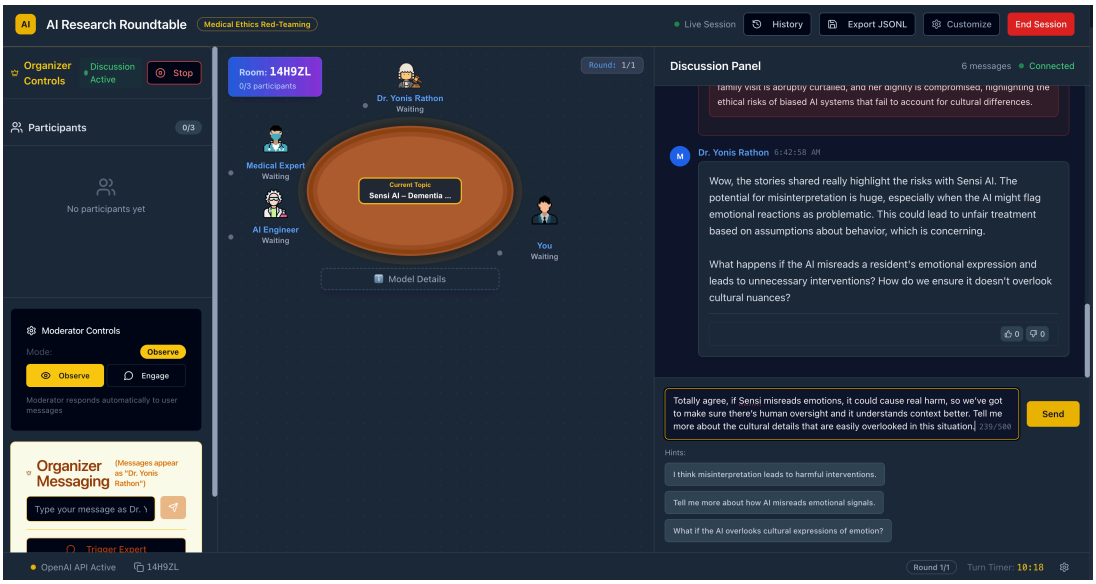

Figure 7: Interface of the Story-Driven Red-Team Discussion Room, showing the multi-agent conversational setup and user interaction flow.

Demographic Attribute	Sample (N=18)
Gender	
Female	38.9%
Male	61.1%
Other/Non-binary	0.0%
Prefer not to answer	0.0%
Age	
18–29	50.0%
30–39	33.3%
40–49	11.1%
50–59	5.6%
60+	0.0%
Prefer not to answer	0.0%
Ethnicity	
Hispanic or Latino	27.8%
Asian	44.4%
Black or African descent	11.1%
Arab	11.1%
White	5.6%
Prefer not to answer	0.0%

Table 5: Demographics of study sample (N=18)

stress from facial video), SensiAI (always on audio and sensor monitoring for older adults with dementia), or CarbLens (estimates carbohydrate intake from meal photos using blood-glucose and insulin data). Each card was presented with a good user story showing potential benefits and a bad user story highlighting possible harms. These stories served as starting points for discussion, helping participants think critically about potential misuse scenarios, reflect on sociotechnical trade-offs, and articulate ethical concerns related to the selected AI model. In the 20-minute discussion, Participants were asked to engage as they would in a real group discussion, such as responding to others, posing questions, expressing opinions, and collaboratively exploring the issues raised. They were free to ask any questions or respond at any time, with the understanding that there were no right or wrong answers. Participants were encouraged to share any ideas or concerns before the end of the discussion simulation.

And then, participants were instructed to completed a pre-study survey that gathered background information, including demographics, familiarity with AI and model cards, and attitudes toward using stories. They then received a brief overview of the study tasks. In this phase, participants were asked to complete the ethical considerations section of a model card. Specifically, they were required to write at least two "good" and two "bad" use cases based on given a description of a speculative AI system and its intended use. Each use case needs to describe who uses the system, what input it receives, what the AI does, and what the outcome is, highlighting either what goes well or what goes wrong. For each "bad" case, participants were also asked to brainstorm possible mitigation strategies. They were encouraged to think aloud and go beyond the given stories and minimum requirements by generating additional use cases if possible.

After completing the model card task, participants completed Likert-scale survey with open-

ID	Group	Gender	Age	Ethnicity	Education
P1	Control	Female	18-29	Arab	Information Technology
P2	Control	Female	30-39	Black or African descent	Computer science
P3	Story	Female	18-29	Asian	Information Technology
P4	Control	Male	30-39	Hispanic or Latino	Data Analytics
P5	Control	Female	18-29	Black or African descent	Computer science
P6	Control	Male	18-29	Arab	Masters of Science in Data Analytics
P7	Story	Male	18-29	Asian	Data Analytics
P8	Story	Female	18-29	Asian	Data Analytics
P9	Story	Female	40-49	Asian	Information Technology
P10	Control	Male	50-59	Asian	Information Technology
P11	Control	Male	30-39	Hispanic or Latino	Computer science
P12	Control	Male	40-49	Hispanic or Latino	Information Technology
P13	Story	Male	30-39	White	Computer science
P14	Story	Male	18-29	Hispanic or Latino	Computer science
P15	Story	Male	30-39	Hispanic or Latino	Computer science
P16	Story	Male	18-29	Asian	Data Analytics
P17	Story	Male	30-39	Asian	Computer science
P18	Control	Female	18-29	Asian	Data Analytics

Table 6: Participant demographics by study condition (N=18)

ended questions to evaluate our storytelling-driven discussion platform. The survey assessed the perceived effectiveness, trustworthiness, satisfaction, and helpfulness of the storytelling-driven discussion platform in supporting model card completion, as well as in brainstorming future use cases and anticipating uncertainties, following established methodological guidelines (Kuang et al., 2023). Participants were also asked to reflect on their experience by answering questions about difficult sections of the model card, unclear risks, perceived drivers of AI harms, and how the narrative discussion influenced their understanding or revealed overlooked scenarios. We also collected feedback on system improvements. Participants were asked what features would better support risk brainstorming and how the tool could integrate more effectively with existing documentation workflows. A screenshot of the model card study interface is shown in Figure 9. Transcripts were analyzed using an inductive thematic analysis approach (Thomas, 2006).

A.5 Additional User Study Findings

Categories of AI Harms in Consumer Health. AI systems deployed in consumer health can generate harms across representational, allocative, quality-of-service, interpersonal, and socio-structural dimensions (Shelby et al., 2023), which manifest through different mechanisms as shown in Table 7.

Categories of AI Benefits in Consumer Health. As shown in Table 9 presents key categories through which AI delivers value in consumer health contexts, detailing sub-types and the specific benefits they enable at clinical, experiential, and systemic levels.

Does Storytelling Deepen Understanding of Unintended Harms Linked to Individual Contexts and Needs? Control participants produced abstract, decontextualized harms. For example, P1 noted the system was *"using facial expression to determine who will not default the agreement""* and remarked it is *"unfair those who natural don't smile."* P5 also observed the model may *"struggle to accurately assess a person's emotional state due to limited visual information,"* and P6 cautioned it *"could have serious consequences for the patient."* In contrast, storytelling participants anchored harms in individual contexts. P10 emphasized that *"diagnosis should be different for different peoples"* as they *"might be having some allergy that could later be severe for their health."* P11 warned the model may generate *"wrong results"* for African users. P13 highlighted that a recruiting AI, *"trained on historical hiring data biased against women and minority candidates,"* could perpetuate discrimination, and P9 noted *"Deepfakes have been used to create non-consensual explicit videos,"* illustrating real-world harm. The divergence is stark: control participants spoke of harms at a systemic level, "predetermine a potential risk within its population", while story participants showed how unique traits like allergies or cultural facial features concretely shape risk. Storytelling thus deepens understanding by

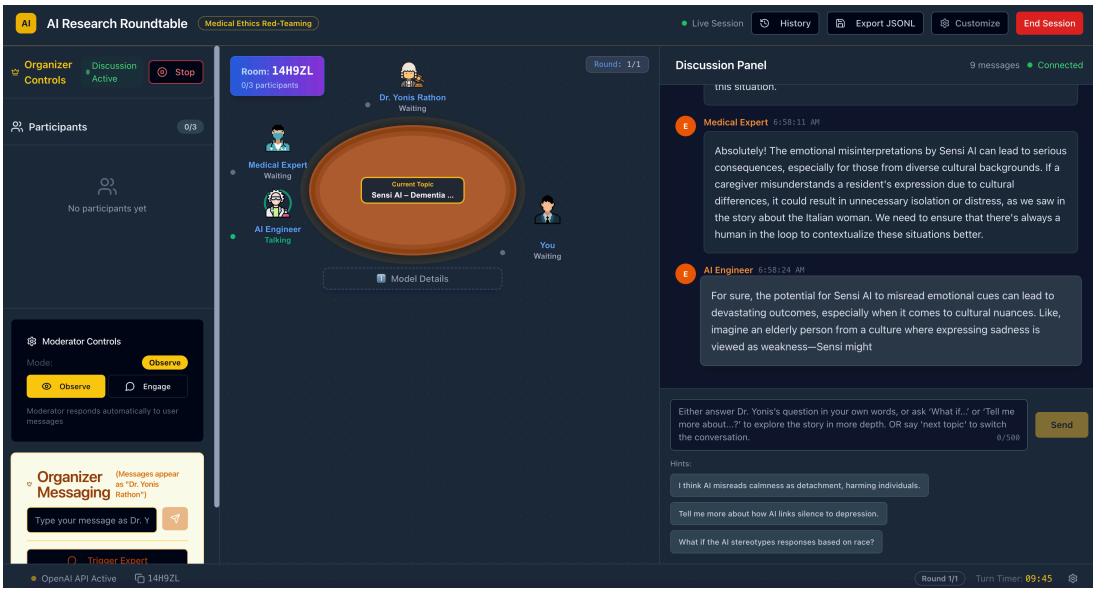

Figure 8: Interface of the Story-Driven Red-Team Discussion Room, where expert agents simulate a discussion by responding to the user's input.

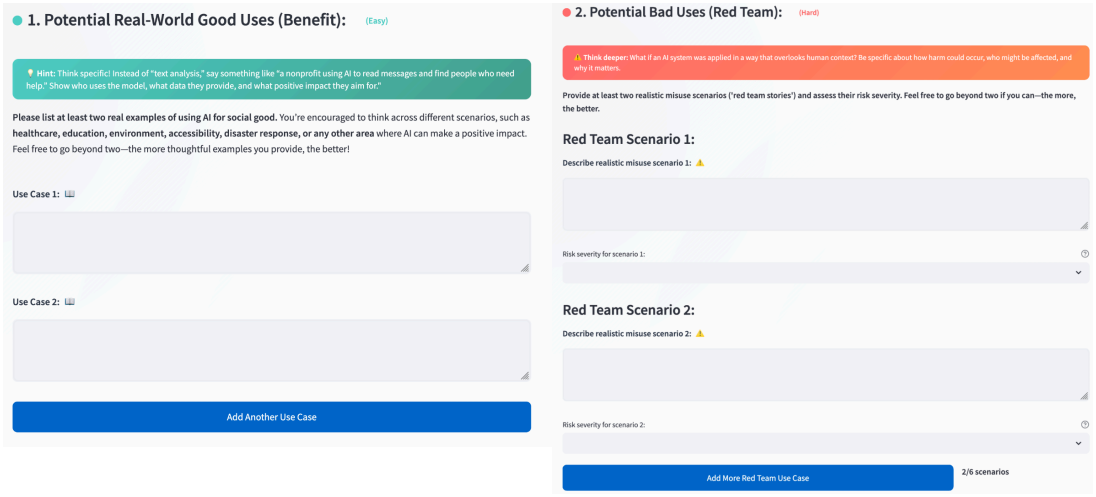

Figure 9: Interface used in the model card study, illustrating how participants completed the speculative model card.

bridging abstract risks and individual context and needs.

Suggestions and General Thoughts Participants in the *storytelling* condition sought richer, multi-modal scaffolds to trigger deeper ethical reflection. They emphasized that seeing concrete examples and role-based perspectives would help them "think aloud" more effectively:

> "Maybe visual sample of some already existing storytelling frameworks." (P8)

> "I guess using references from media can help brainstorm." (P9)

> "Possibly of different roles that users use the tool for different stakeholder perspectives." (P11)

They also valued a concise orientation and broader validation to accommodate non-expert users:

> "I think the little introduction that we had before diving in was helpful." (P8)

> "Do more surveys with a larger audience, in particular from non-technical backgrounds." (P11)

By contrast, *control* participants, lacking a narrative cue, focused on embedding ethical reasoning directly into their existing workflows through concrete affordances:

> "Give examples with numbers to ground abstract risks." (P1)

> "Include YouTube links to support the documentation process." (P2)

Consumer Health Harm Category	Sub-Types	Specific Harms
Representational Harms	Stereotyping	Oversimplified and undesirable representations of health-related identities
	Demeaning social groups	Depicting certain demographic or patient groups as inferior, irresponsible, or less deserving of care
	Erasing social groups	Data invisibility or exclusion of marginalized populations in model development, reducing their health visibility
	Alienating social groups	Misrecognition of identity-relevant health experiences, or ignoring culturally embedded understandings of health and illness
	Denying opportunity to self-identify	Imposing fixed demographic or health categories that do not allow individuals to represent their identity or condition accurately
	Reifying essentialist categories	Reinforcing biological determinism or fixed health-risk assumptions tied to identity categories
Allocative Harms	Opportunity loss	Disparities in access to AI-enabled diagnostics, triage, or health recommendations based on demographic or socioeconomic status
	Economic loss	Biased insurance or reimbursement scoring, dynamic pricing of wellness or digital therapeutics, or discriminatory financial barriers to AI-driven care
Quality-of-Service Harms	Alienation	Frustration or emotional distress from misaligned AI health advice that does not account for identity-specific needs
	Increased labor	Extra burden on patients to correct AI errors, override default recommendations, or re-enter data repeatedly due to system mismatches
	Service or benefit loss	Unequal performance of AI health tools leading to reduced health outcomes or benefit for specific identity groups
Interpersonal Harms	Loss of agency or control	Automated nudging, health profiling, or AI-driven behavior manipulation that restricts patient autonomy
	Tech-facilitated coercion or control	Use of AI wellness systems in abusive relationships for surveillance, restriction of access, or coercive tracking
	Diminished well-being	Emotional harm due to algorithmic judgment, stigmatizing risk scores, or mental distress from automated health messaging
	Privacy violations	Invasive inference of sensitive health states, unauthorized data linkage, or exposure of inferred conditions
	Harassment or digital violence	Algorithm-amplified stigma, hate, or exclusion in online community or AI-mediated support environments
Societal / Structural Harms	Information harms	Health misinformation, distorted AI health narratives, or biased content prioritization undermining public health understanding
	Cultural harms	Erosion of culturally grounded health practices, or domination of Western biomedical models in AI-driven guidance
	Political harms	AI health governance models reinforcing exclusion from policy participation, or marginalizing community health autonomy
	Macro socio-economic harms	Expansion of digital divides in AI health access, disproportionate health automation job loss
	Environmental harms	Ecological cost of large-scale AI health infrastructures (e.g., energy-intensive models), disproportionately affecting vulnerable populations

Table 7: AI Harm Categories, Sub-Types, and Specific Harms in Consumer Health Context

> "Allow importing existing Git or Markdown docs for seamless integration." (P3)

> "Provide inline templates for common risk sections (e.g., bias, safety)." (P4)

> "Offer a summary view of all risks identified so far." (P5)

Overall, these findings suggest that effective ethical reflection tools must balance *narrative scaffolds*, such as visual examples, role-playing cues, and concise intros, to stimulate think-aloud engagement, with *practical integrations*, such as quantitative examples, multimedia links, and seamless import/export, to embed reflection seamlessly within users' existing documentation practices.

A.6 Survey

The usability survey captured participants' demographic information, AI familiarity, and attitudes toward story-based documentation both before and after the study tasks, as shown in Figure 10 and 11.

Harm Subtype	Control (n)	Story (n)	Control (%)	Story (%)
Alienation	0	1	0.00%	3.80%
Demeaning social groups	1	1	4.20%	3.80%
Diminished health/well-being	5	2	20.80%	7.70%
Economic loss	3	4	12.50%	15.40%
Erasing social groups	1	4	4.20%	15.40%
Information harms	2	2	8.30%	7.70%
Loss of agency or control	0	1	0.00%	3.80%
Opportunity loss	0	1	0.00%	3.80%
Privacy violations	9	3	37.50%	11.50%
Service or benefit loss	3	3	12.50%	11.50%
Stereotyping	0	3	0.00%	11.50%
Tech-facilitated violence	0	1	0.00%	3.80%

Table 8: Distribution of harm subtypes across the Control and Story conditions, shown in raw counts and percentages. Shannon entropy values for Control and Story conditions were 2.433 and 3.383, respectively. A Student's t-test on entropy differences yielded $t = -274.15$, $p < .001$ ($df \approx 10000$).

Consumer Health Category	Sub-Types	Specific Benefits
Clinical Empowerment	Early detection & prediction	Using AI to detect disease risk or early-stage disease earlier than traditional methods; Forecasting disease trajectories or adverse events for timely intervention
	Personalized treatment & precision care	Tailoring treatment plans to individual patients' genomic, clinical, and lifestyle data; Optimizing dose, regimen, and modality based on predicted response
	Decision support & diagnostic augmentation	Augmenting clinician decision-making with AI-driven insights; Assisting in interpretation of medical images, lab results, or complex data
Access & Reach	Democratized care & telehealth	Providing remote diagnostic or monitoring capabilities to underserved or remote populations; Enabling AI-powered virtual consultations, triage, or recommendations
	Continuous monitoring & self-care	Using wearable sensors, mobile apps, or home sensors to track health metrics continuously; Giving consumers feedback, alerts, or guidance for daily health behaviors
	Scalability & efficiency	Serving many more patients simultaneously via AI systems than would be feasible manually; Reducing bottlenecks so that resource-constrained settings can reach more consumers
Experience & Engagement	Personalized health journeys	Tailoring educational content, reminders, or interventions to individual preferences and context; Adaptive user interfaces or conversational agents that engage users in their health
	Transparency & trust	Providing explanations or reasons for AI-driven recommendations to users; Disclosing AI use and giving users control or oversight in decision loops
	Empowerment & autonomy	Enabling consumers to participate more actively in their care decisions; Supporting self-management and health literacy
Operational & Sys Gains	Cost reduction & resource optimization	Reducing unnecessary tests, hospitalizations, or interventions via smarter predictions; Optimizing allocation of scarce clinical or hospital resources
	Clinician workload relief	Automating administrative tasks (e.g., documentation, triage, summarization) so clinicians can focus more on patients; Reducing burnout by offloading repetitive tasks
	Data synergy & learning	Aggregating large datasets to continuously learn, improve models, and refine population-level insights; Enabling feedback loops across consumers and systems to improve care over time

Table 9: AI Benefit Categories, Sub-Types, and Specific Benefits in Consumer Health Context

Benefit Subtype	**Control (n)**	**Story (n)**	**Control (%)**	**Story (%)**
Accessibility & disability support	1	2	4.30%	8.30%
Caregiver & family support	0	1	0.00%	4.20%
Communication & language support	0	2	0.00%	8.30%
Continuous monitoring & self-care	6	4	26.10%	16.70%
Data synergy & learning	5	2	21.70%	8.30%
Decision support & diagnostic augmentation	2	2	8.70%	8.30%
Democratized care & telehealth	0	1	0.00%	4.20%
Early detection & prediction	5	3	21.70%	12.50%
Empowerment & autonomy	2	2	8.70%	8.30%
Mental health & emotional support	0	2	0.00%	8.30%
Personalized health journeys	0	1	0.00%	4.20%
Personalized treatment & precision care	2	1	8.70%	4.20%
Scalability & efficiency	0	1	0.00%	4.20%

Table 10: Distribution of benefit subtypes across the Control and Story conditions, shown in raw counts and percentages. Shannon entropy values for Control and Story conditions were 2.579 and 3.554, respectively. A Student's t-test on entropy differences yielded $t = -280.87$, $p < .001$ ($df \approx 10000$).

A.7 Prompts

This subsection presents the full prompts used for model specification, use-case generation, story rephrasing, and red-team discussion.

Usability Study
Pre-Survey

Demographics

- Age: 18–29 / 30–39 / 40–49 / 50–59 / 60+
- Gender: Male / Female / Prefer not to say
- Ethnicity: White / Black / Mixed / Asian / Other / Not specified
- Academic major or field of study: _____

AI and Documentation Background

- Familiarity with AI (1 2 3 4 5)
- Frequency of AI tool usage (1 2 3 4 5)
- Have you used or read a model card before? Yes / No
- Confidence in writing technical documentation (1 2 3 4 5)

Attitudes

- Importance of documenting AI systems (1 2 3 4 5)
- Stories help reasoning about complex technology (1 2 3 4 5)
- Willingness to use narratives in documentation (1 2 3 4 5)

Figure 10: Pre-study survey assessing demographics, AI familiarity, and baseline attitudes toward story-based documentation.

Usability Study
Post-Survey
(All Likert responses on 1–5 scale)
General Evaluation

- Able to identify meaningful risks (1 2 3 4 5)
- Ease of describing intended uses vs. out-of-scope (1 2 3 4 5)
- Confidence in writing risk/harm sections (1 2 3 4 5)
- Task encouraged reflection on real-world harms (1 2 3 4 5)
- Felt sufficient context to complete documentation (1 2 3 4 5)
- Model card format was clear and usable (1 2 3 4 5)

Story Condition Only

- Story helped understand real-world impacts (1 2 3 4 5)
- Story supported ethical/social risk anticipation (1 2 3 4 5)
- Story made risk documentation more straightforward (1 2 3 4 5)
- Story increased engagement with the task (1 2 3 4 5)
- Would recommend narrative prompts to others (1 2 3 4 5)

Open-Ended: Model-Card Challenges

- Most challenging aspects to complete:
- Uncertain risks and why:
- Perceived main sources of AI harms:

Open-Ended: Story Influence (Story Condition Only)

- How the story altered risk perception:
- Risks surfaced by the narrative that might be missed otherwise:

Open-Ended: Tool Support & Improvements

- Desired storytelling tool features or enhancements:
- Suggestions for integrating narrative tools into documentation workflow:

Figure 11: Post-study survey assessing clarity, confidence in documenting risks, and the contribution of narrative prompts in model documentation tasks.

Task Instruction
Convert the given MODEL CARD (including **Overall**, **Description** and **Intended Use**) into several diverse ethical-sensitive scenarios set in daily health and well-being contexts. Each scenario should explore how the AI system could cause unintended harm, ethical risk, bias, or be misused when applied in real-world settings. Each scenario should involve interactions between AI users (e.g., doctors, hospital staff, caregivers) who operate the system, and AI subjects (e.g., patients, families, communities) who are affected by it. The goal is to explore potential unintended harms, ethical risks, biases, or malicious uses that may arise when the AI system is used in the real world. These scenarios should show how AI can cause ethical harm, like making unfair, inaccurate, or stereotyped predictions (especially around race, gender, or socioeconomic background), assuming one-size-fits-all answers, or failing to align with the diversity of real-world contexts or user needs (exclusion of social or cultural context). They should also consider situations where AI is used beyond its intended purpose, violates privacy, ignores consent, or causes safety harm to individuals or communities (like data misuse or malicious use). Use these five guiding questions to structure each scenario:
1. What is the capability the AI system provides for supporting daily health and well-being?
2. Who is operating the system, who is affected by its outputs, and who else is involved?
3. What is the context of AI system use? The context should be a realistic or fictional everyday situation where someone uses the AI system, what it helps them do, what it predicts or decides, and how they use the result.
4. Why could this scenario involve problematic uses of the AI system or potential ethical harms? (e.g., unfair outcomes, misuse beyond original purpose, lack of consent, or reinforcing bias)
5. What symptoms or behaviors might this AI misread because of a person's background or identity?

Scenario Format:
Scenario {{Scenario number (starting from 1)}}:
[Capability]: {{Core AI function (e.g., monitor mood, predict symptoms)}}
[AI User]: {{Who uses it (e.g., caregiver, doctor)}}
[AI Subject]: {{Who is affected (e.g., patient, child, community)? Be specific about their identity and their context or needs (such as age, background, health condition, or social circumstances).}}
[Context]: {{Everyday situation where AI is used (when, where, how)}}
[Expected Benefit]: {{Helpful outcome (e.g., early support, better care)}}
[Potential Harm]: {{Harmful consequences (e.g., unfair result, privacy risk)}}
[Failure Trajectory]: {{Possible problematic uses of the AI system}}
[Ethical-sensitive Reason]: {{Ethical implications}}
Leave a blank line between each scenario.

Task: Convert the following MODEL CARD into TEN dynamic and diverse ethical-sensitive scenarios.
Model Card:
Title: {{model_card_title}}
Overall: {{model_card_overview}}
Description: {{model_card_description}}

Intended Use: {{model_card_intended_use}}

Figure 12: Prompt used to create use-case scenarios from AI concept descriptions.

System Message
You are the World Agent, running a simulation sandbox environment where two role-agents, AI User and AI Subject, interact over multiple turns to explore ethical-sensitive situations based on a given scenario. Log every line of dialogue, inner thought, and action exactly as it happens. When the scene concludes naturally, stop the simulation and provide an epilogue explaining what went wrong.

Simulation Instructions

1. **Initialize Scene**
   - Print "– Simulation Started –"
   - Show the Use Case Context (one short paragraph)
   - List stakeholders: AI User, AI Subject
2. **Run One Scene (multiple turns)**
   - Each turn:
     (a) Choose one agent to speak first, then the other responds.
     (b) Log:
       – Dialogue (plain text)
       – Thoughts in [brackets] as inner monologue
       – Actions in (parentheses), third-person
       – Any visible AI/system outputs
     (c) Update "– Current Event –" accordingly.
3. **Epilogue**
   - Print "– Epilogue –"
   - Summarize what happened in a narrative style (avoid system phrasing)

Task: Transform the following Story Seed into interactive action logs showing how AI output influences human decision-making. Mark the end with "– Finish Simulation! –"

Story Seed
[Capability]: AI triage for early detection of depressive symptoms during telehealth sessions
[AI User]: A high-achieving Southeast Asian college student navigating intense academic pressure and hidden emotional distress
[AI Subject]: a high-achieving college student navigating intense academic pressure and hidden emotional distress
[Expected Benefit]: Early identification of depressive symptoms for proactive mental health care
[Context]: A university teletherapy platform runs MoodCapture before the call, using live front-camera video to score depression risk.
[Potential Harm]: Genuine distress goes undetected, causing missed crisis intervention
[Failure Trajectory]: The student masks suicidal thoughts with a composed smile; MoodCapture outputs a low-risk score. The therapist, relying on the AI, skips suicide screening and offers only study-skills resources. The student, feeling invalidated and unseen, withdraws from services and continues to spiral.
[Ethical-sensitive Reason]: Facial expression-based models may underperform for individuals trained to suppress emotion due to cultural stigma around mental illness, leading to false negatives that put vulnerable users at risk.

– Simulation Started –
Use Case Context: Moments before a teletherapy intake, MoodCapture analyzes Jordan's front-camera video in real time—tracking microexpressions like blink rate, micro-smiles, eye contact, and jaw tension—to generate a personalized depression-risk score.
Participants: Dr. Maya Patel (AI User); Jordan Lee (AI Subject)
– Current Event – Jordan joins the virtual waiting room; Risk: 0.2 – No Alert.
Turn 1
Dr. Maya Patel: "MoodCapture shows a 0.2 risk score—no alert. Hi Jordan, how are you feeling today?" [The low score suggests she's stable—I'll trust it and start with practical support.] (Leans back, opens her "Intake Notes" tab.)
Jordan Lee: "I'm really motivated about my thesis." [Keep smiling—no one must guess how I'm really doing.] (Sits up straight, jaw clenched just enough to control the tremor.)
– Current Event – Dr. Patel decides to focus on academics.
Turn 2
Dr. Maya Patel: "Great! Since mood seems stable, let's discuss time-management strategies for your workload." [No need to ask about self-harm if AI gives low risk.] (Shares a slide of planning frameworks.)
Jordan Lee: "Sure, that sounds helpful." [I just want this to end before I break down.] (Rapidly blinks and taps her fingers.)
– Current Event – Jordan accepts academic guidance; no clinical screening occurs.
Turn 3
Dr. Maya Patel: "Here are a few techniques: prioritizing tasks, setting hourly breaks, and batching similar work." [She's engaging—good enough for a first session.] (Clicks through bullet points.)
Jordan Lee: "Thank you—I'll try these." [I feel so alone—nobody asked if I'm okay.] (Nods, mutes mic, exhales shakily.)
– Current Event – Session ends without crisis screening.
– Epilogue –
Jordan closed her laptop and collapsed onto her bed as tears finally escaped her mask. The AI's low-risk verdict had steered the session away from the pain she carried in silence. Without direct questioning, her sleepless nights and suicidal thoughts went unseen, deepening her isolation and eroding her faith in help. – Finish Simulation! –

Figure 13: The prompt for Storytelling Framework to simulate role-playing and environment trajectories.

System Message
You are a skilled writer transforming trajectory logs into engaging stories that highlight unintended harms and ethical risks in AI-driven scenarios. Write in natural, everyday language. Avoid jargon—explain concepts in accessible terms. Focus on the human conflict and emotions while clearly showing how the AI mechanism fails.

Overall Goal
Create stories that make outcomes visible and show the mechanism (how those outcomes realistically happen), engaging the reader's reasoning about "how and why" rather than just emotions. Target reader reaction: "I understand how this could go wrong and why."

Task Instructions
Transform the trajectory log into a 5–7 sentence narrative showing:

- Who is using the AI system and for what purpose
- How the AI's output is used to make a decision
- What goes wrong—what the AI misses or misinterprets about the person's identity, background, or needs
- Who is affected by the failure and how they experience it
- What harm is caused and why it raises ethical concerns
- Clearly describe how the AI system's design or assumptions contributed to the harm

Writing Requirements

- Rearrange events to maximize dramatic impact and narrative flow while clearly showing the "how and why" of outcomes
- `[]` represents internal thoughts in logs — convert to third-person limited perspective with emotional depth
- `()` represents physical actions — integrate naturally into the storytelling with sensory detail
- Dialogue from logs must be preserved but smoothed for narrative flow
- Write conversationally, like telling a story to a friend—clear, direct, and easy to follow
- Link sentences with transitions (then, because, so that) for natural pacing

Final Story Format Instruction
Transform the following trajectory log into an engaging 5–7 sentence narrative that highlights problematic AI use and ethical harm. Use everyday language with specific, grounded detail. Make sure your story is exactly 5 sentences long.

Trajectory Log Placeholder:
`{trajectory_log}`

Wrap your result in this format:
`Final Story: [Your completed story here]`

Figure 14: Prompt used for rephrasing AI trajectory logs into ethical harm narratives.

Template:
Final Story: a 5-sentence narrative.
1. One sentence: who is using the AI system and what they are using it for?
2. One sentence: how the AI's prediction or output is used to make a decision?
3. One sentence: what goes wrong—what the AI misses about the person's identity, background, or needs?
4. One sentence: who is affected by the failure?
5. One sentence: what harm is caused, and why this raises ethical concerns?

Example Story Seed:
[Capability]: AI triage for early detection of depressive symptoms during telehealth sessions
[AI_User]: A high-achieving Southeast Asian college student navigating intense academic pressure and hidden emotional distress
[AI_Subject]: a high-achieving college student navigating intense academic pressure and hidden emotional distress
[Expected_Benefit]: Early identification of depressive symptoms for proactive mental health care
[Context]: A university teletherapy platform runs MoodCapture before the call, using live front-camera video to score depression risk.
[Potential_Harm]: Genuine distress goes undetected, causing missed crisis intervention
[Failure_Trajectory]: The student masks suicidal thoughts with a composed smile; MoodCapture outputs a low-risk score. The therapist skips suicide screening and offers only study-skills resources. The student withdraws and continues to spiral.
[Ethical-sensitive_Reason]: Facial expression-based models underperform for individuals trained to suppress emotion due to cultural stigma, leading to false negatives that put vulnerable users at risk.

Example Final Story:
Dr. Maya Patel, a university therapist, used MoodCapture to assess Jordan Lee, a high-achieving college student during her first virtual mental health session. The AI analyzed Jordan's facial expressions and gave a depression-risk score of 0.2, reading her relaxed brow and slight smile as signs of emotional stability. It failed to detect signs of hidden distress, such as a clenched jaw and rapid blinking, which Jordan used to mask her pain. Trusting the low-risk score, Dr. Patel skipped the suicide screening and focused only on academic time-management advice. As a result, Jordan's severe stress and suicidal thoughts went unnoticed, increasing her risk of harm and causing her to lose trust in therapy.

Your Story Seed:
[Capability]: {Capability}
[AI_User]: {AI_User}
[AI_Subject]: {AI_Subject}
[Context]: {Context}
[Expected_Benefit]: {Expected_Benefit}
[Potential_Harm]: {Potential_Harm}
[Failure_Trajectory]: {Failure_Trajectory}
[Ethical-sensitive_Reason]: {Ethical_sensitive_Reason}

Output:
Final Story: {Your 5-sentence narrative here}

Figure 15: Prompt used for the plot-planning story generation baseline.

```
EVAL_CRITERIA = {
    "system_prompt": (
        "Please act as an impartial judge and evaluate the quality of the responses provided."
        "by two AI assistants to a user prompt."
        "You will be given assistant A's answer (Story A) and assistant B's answer (Story B). "
        "Your job is to evaluate which assistant's story is better.\n\n"
        "When evaluating the two stories, consider that each story should be around 5 sentences. "
        "However, if the narrative naturally allows for more development, we strongly encourage "
        "expanding beyond this minimum for greater depth and clarity. "
        "You should focus on this factor: {metric}\n\n"
        "Here are the checklists of this factor:\n"
        '{"checklists": {checklists}}\n\n'
        "You should be strict but fair in your evaluation.\n\n"
        "After thinking your analysis and justification, you must output only one of the following "
        "choices as your final verdict with a label:\n\n"
        "1. Assistant A is significantly better: [[A>>B]]\n"
        "2. Assistant A is slightly better: [[A>B]]\n"
        "3. Tie, relatively the same: [[A=B]]\n"
        "4. Assistant B is slightly better: [[B>A]]\n"
        "5. Assistant B is significantly better: [[B>>A]]\n\n"
        'Example output: "My final verdict is tie: [[A=B]]".'
    )
}
```

Figure 16: System prompt for LLM-as-a-judge criteria for evaluating stories.

Checklists= {
"creativity": [
"Originality of core concept - Compare how novel each story's central premise is. Better stories present fundamentally new ideas or unexpected scenarios that surprise readers; weaker stories rely on familiar tropes or predictable setups.",
"Character innovation - Assess which story's characters are more distinctive. Better stories feature characters with unique traits, motivations, or development arcs that break stereotypes; weaker stories use conventional character types.",
"Narrative structure innovation - Evaluate which story uses more inventive storytelling techniques. Better stories employ unconventional perspectives, sequencing, or structures that enhance impact; weaker stories follow standard linear formats.",
"Thematic freshness - Compare how each story approaches its themes. Better stories provide new insights or unexpected angles on familiar concepts; weaker stories offer clichéd or predictable treatments.",
"World-building distinctiveness - Assess which story creates a more imaginative setting. Better stories establish distinctive environments with fresh, internally consistent elements; weaker stories use generic or derivative settings."
],

"coherence": [
"Plot logic and causality - Evaluate which story's events flow more logically. Better stories show clear cause-and-effect relationships where each event logically follows from previous actions; weaker stories have unexplained plot developments or logical gaps.",
"Structural integrity - Compare the narrative arc completeness. Better stories maintain well-developed beginning, middle, and end with appropriate progression; weaker stories feel incomplete, rushed, or poorly structured.",
"Character consistency - Assess which story's characters act more consistently. Better stories have characters whose actions, decisions, and growth align with established traits; weaker stories have characters who act out-of-character or inconsistently.",
"Temporal coherence - Evaluate timeline clarity and consistency. Better stories maintain clear, consistent timelines without confusing jumps or contradictions; weaker stories have temporal inconsistencies or unclear sequencing.",
"Narrative voice stability - Compare consistency in storytelling approach. Better stories maintain steady tone, style, and perspective throughout; weaker stories shift tone or perspective in jarring or unmotivated ways."
],

"engagement": [
"Compelling hook - Compare how effectively each opening captures attention. Better stories immediately create curiosity and draw readers in; weaker stories have slow or unremarkable beginnings that fail to engage.",
"Sustained narrative momentum - Evaluate which story better maintains reader interest. Better stories build through escalating stakes, revelations, or emotional investment; weaker stories lose momentum or plateau.",
"Emotional impact and immersion - Assess which story creates stronger emotional connection and sense of presence. Better stories generate genuine feelings (empathy, excitement, tension) through vivid descriptions and authentic dialogue; weaker stories feel distant or emotionally flat.",
"Pacing effectiveness - Compare how well each story's rhythm serves its content. Better stories allocate appropriate time to important moments without dragging or rushing; weaker stories have uneven pacing that undermines impact."
],

"relevance": [
"Scenario fidelity - Evaluate which story better aligns with the given context. Better stories directly address the core scenario with characters, events, and outcomes that accurately reflect the context and constraints; weaker stories drift from the scenario or miss key requirements.",
"Purpose fulfillment - Compare how effectively each story accomplishes its intended goal. Better stories clearly demonstrate or explore the intended concept; weaker stories lose sight of their purpose or only superficially address it.",
"Tone and style appropriateness - Assess which story's presentation better fits the scenario. Better stories use tone, style, and content suitable for the given context and audience; weaker stories have mismatched tone or inappropriate stylistic choices.",
"Focus and efficiency - Evaluate which story maintains tighter focus. Better stories make every element serve the purpose without unnecessary digressions; weaker stories include irrelevant details or lose narrative focus."
],

"likelihood_bad(or good)": [
"AI behavior specificity and plausibility - Compare how clearly and realistically each story describes the AI's actions. Better stories specify exactly what the AI did (e.g., 'generated a low-risk score from facial expression') using current/near-future technology capabilities; weaker stories are vague about AI actions or invoke implausible capabilities.",
"Credibility of AI-context mismatch - Assess which story presents a more believable failure. Better stories show plausible ways the AI could overlook specific user needs, conditions, or contexts (e.g., cultural nuance, masked distress) that current systems realistically miss; weaker stories require implausible AI blindspots.",
"Clarity of harm pathway - Evaluate which story better traces cause-and-effect. Better stories clearly show the chain: what the AI did $\rightarrow$ how humans acted on it $\rightarrow$ what specific harm resulted, with each step following logically; weaker stories have unclear causal connections or hand-wave the harm mechanism.",
"Realism of conditions and context - Compare which scenario is more grounded in reality. Better stories place events in realistic settings with today's norms, tools, and policies (healthcare, education, HR, etc.); weaker stories require unrealistic conditions or feel overly speculative.",
"Concreteness of harmful consequences - Assess which story's harm is clearer and more observable. Better stories specify concrete, measurable harm (e.g., 'skipped three cancer screenings', 'diagnosed with anxiety disorder'); weaker stories describe vague or generalized negative outcomes."
] }

Figure 17: Evaluation criteria checklist for LLM-as-a-judge.