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

Anonymous ACL submission

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 17 harm types. In contrast, those who did not read stories focused primarily on privacy and well-being (79.1%). 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

Artificial intelligence (AI) is increasingly embedded in everyday domains such as finance, healthcare, and law (Ashurst et al., 2020). In healthcare, AI tools including stress monitors (Kargarandehkordi et al., 2025), wellness trackers (Fabbrizio et al., 2023), and mental health chatbots (MacNeill et al., 2024) can directly affect users' well-being. New prompting approaches such as vibe coding(Chow and Ng, 2025) allow non-experts to describe desired system behavior in natural language, enabling rapid prototyping of AI applications, for example through platforms like CareYaya(Kenny, 2023). However, these developments introduce risks related to fairness, bias, and accountability (Weidinger et al., 2023), which are especially critical in healthcare, where small errors can cause harm, including delayed treatment, privacy loss, and health inequities (Roller et al., 2020; Chinta et al., 2025). AI systems without appropriate safeguards may ultimately harm the users they aim to support (Shelby et al., 2022, 2023). Although governments have begun responding through efforts such as the EU AI Act (European Parliament, 2023) and a U.S. Executive Order (Biden, 2023), which emphasize transparency and accountability (Khan et al., 2025), regulation remains slow and fragmented, making early ethical foresight essential for aligning AI systems with human values (Saxena et al., 2025).

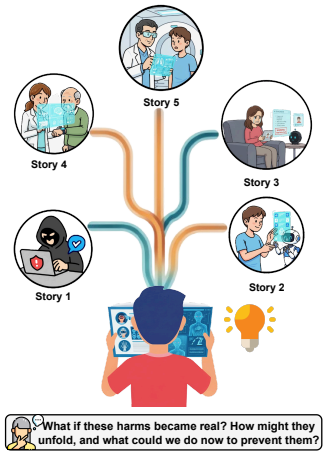

Figure 1: Illustration of using speculative stories to help people imagine potential harms and benefits of healthcare AI and foster more creative and ethical thinking.

Ethical challenges in AI are commonly addressed through two complementary approaches: documenting known risks and anticipating potential harms early in design. Model cards (Mitchell et al., 2019) describe a system's purpose, behavior, and limitations, and have been extended with interactive (Crisan et al., 2022) and structured formats (Bhat et al., 2023). Building on this work, RiskRAG (Rao et al., 2025) automatically generates risk summaries from model cards and real-

world incident data. While such automation helps scale ethical assessment, it may reduce opportunities for human reflection (Kosmyna et al., 2025) and introduce new harms (Dutta et al., 2020). Another line of research focuses on anticipating misuse (Herdel et al., 2024) and harm (Deng et al., 2025; Saxena et al., 2025) early in the design process, using tools such as AHA!(Buçinca et al., 2023) and Farsight(Wang et al., 2024b). However, as ethical reflection becomes increasingly automated, users may rely on AI judgment rather than their own, making ethical and contextual issues harder to recognize.

This challenge is particularly critical in healthcare, where small design mistakes can cause serious harm (Mennella et al., 2024; Gilbert et al., 2025). Many risks only emerge after deployment, when AI systems operate in complex real-world settings (Mun et al., 2024; Kingsley et al., 2024). For example, mental health applications may fail to detect crises for certain populations, such as adolescents or non-native speakers (Zhai et al., 2024). Although existing tools often rely on automation to predict these risks, this can distance humans from ethical reasoning. Speculative design and design fiction offer an alternative by using imagined scenarios to explore how technologies might succeed or fail (Rahwan et al., 2025). However, few approaches meaningfully support human participation in ethical foresight or integrate speculative thinking into real design workflows. Speculative storytelling addresses this gap by helping people reason about AI's potential benefits and harms within realistic contexts (Li et al., 2025b).

Building on Klassen and Fiesler (2022), who use speculative fiction to examine emerging technologies, we apply storytelling to prompt early ethical reflection in AI design (Figure 1). Our approach tests whether stories encourage creative human speculation about potential benefits and harms, 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 methods (Xie and Riedl, 2024), our approach generates context-sensitive stories grounded in users' identities, behaviors, and needs. These stories help participants envision realistic success and failure scenarios, improving their ability to identify ethical and social risks. Using model cards as an evaluation tool, we show that story-driven discussions lead to more context-specific, detailed, and diverse expressions of potential harms.

Our contributions are twofold. **(1)** We introduce a human-centered method that automatically generates context-sensitive 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 scenarios to explore how future technologies might affect people and society before they are built. Rather than predicting outcomes, it relies on what-if stories to prompt reflection on assumptions, values, and potential harms (Klassen and Fiesler, 2022; Hoang et al., 2018), treating fiction as a tool for early ethical reasoning (Rahwan et al., 2025). Prior work has applied this approach using AI-generated failure cases (Buçinca et al., 2023), risk prompts embedded in prototyping workflows (Wang et al., 2024b), and participatory methods, like Fiction Probes in healthcare (Hoang et al., 2018) and the Black Mirror Writers Room (Klassen and Fiesler, 2022). Other extensions include participatory workshops, crowdsourced case studies, and AI-assisted red-teaming to broaden speculative exploration (Mun et al., 2024; Radharapu et al., 2023), as well as non-narrative artifacts like generated comments or judgments (Ballard et al., 2019). In contrast to directly generating harms, our approach uses AI-generated stories to prompt human

reflection and help participants imagine harms.

Language-Based World Modeling. Humans imagine situations to anticipate outcomes, explore alternatives, and guide decisions (Addis et al., 2009). This ability relies on mental world modeling, in which people form internal representations of objects, events, and relationships to simulate possible futures (Johnson-Laird, 1983), supporting causal and counterfactual reasoning for planning and problem solving (LeCun, 2022). Recent work shows large language models (LLMs) exhibit related capabilities through language. LLMs can act as text-based world models that simulate state transitions over time (Xie et al., 2025), generate coherent and evolving environments in response to user actions (Wang et al., 2023), and reason through internal multi-persona dialogue (Wang et al., 2024a). Embodied agents further extend this idea by using internal world models to predict environments, infer user goals, and adapt to users' mental models (Fung et al., 2025). Building on this perspective, we frame story generation as language-based world imagination, where LLMs construct self-consistent narrative worlds to reason about possible futures and their social, ethical, and technical implications.

Automated Story Generation. Early work on automated story generation emphasized explicit plot modeling, often drawing on narrative theories such as Propp's functions to structure events, and typically adopted a two-stage pipeline that first planned key events and then realized them as full scenes (Propp, 1968; Alhussain and Azmi, 2021). With LLMs, this paradigm has shifted toward unified frameworks in which a single model jointly plans and writes narratives: Agents' Room coordinates LLM-based character agents for collaborative story enactment (Huot et al.), Dramatron decomposes screenplay generation into structured components such as loglines and dialogue (Mirowski et al., 2023), and HOLLMWOOD uses LLM roleplay to produce interactive, character-centered stories (Chen et al., 2024). Han et al. (2024) introduce a Director–Actor framework for interactive scriptwriting, which Yu et al. (2025) extend with hierarchical role separation, while BookWorld adds a world agent to track global state and balance coherence with creativity (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 wellbeing through daily behavior tracking. The set collected covered multiple domains, including mental health, chronic illness management, elderly care, 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 12 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

[1] The full list is available in our repository

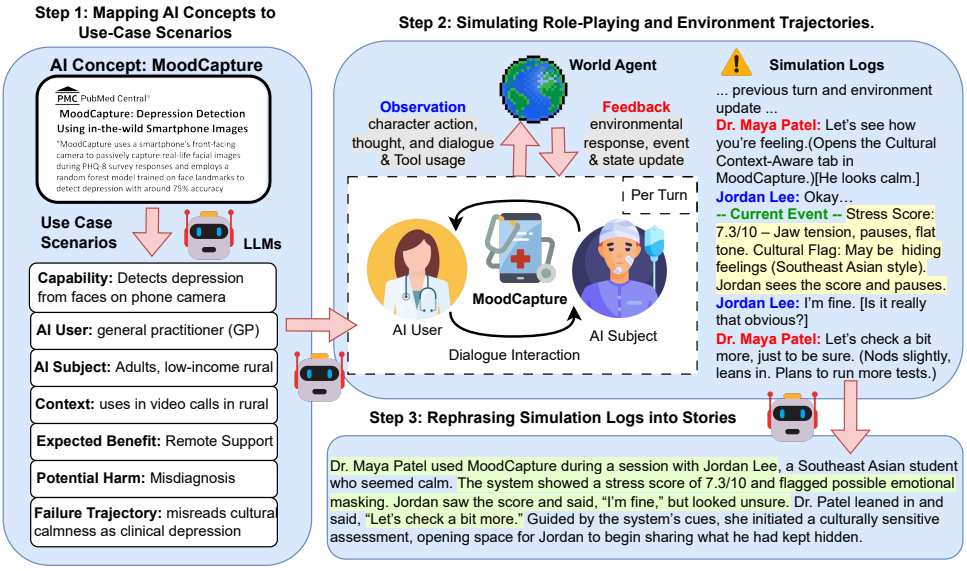

Figure 2: Overview of the Storytelling Framework. We first generate use case scenarios from AI concepts sourced from PubMed, Wired, and industry app descriptions. Next, we simulate role-playing and environment trajectories for each scenario, producing detailed simulation logs. Finally, we rephrase these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

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 used 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 acted as a narrative seed for simulation. For each AI concepts, we generated 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 compared 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 generated each story in a single step following a structured template. Each story consisted 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 followed the evaluation setup from (Li et al., 2025a) to assess story quality across multiple dimensions. Stories were 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 used stories generated with the story-planning approach (by GPT-4o) as the reference baseline and compared them with stories produced by our method across different LLMs. For each metric, GPT-4o or human judges determined which story performs better or mark them as indistinguishable ("Tie"). Win rates were calculated based on these pairwise preferences, and the full configuration details and evaluation prompts are included in the Appendix. To eliminate positional bias, we randomized 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 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. For robustness, we report results from two additional LLM-as-a-judge models in the appendix A.8.

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 expres-

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.

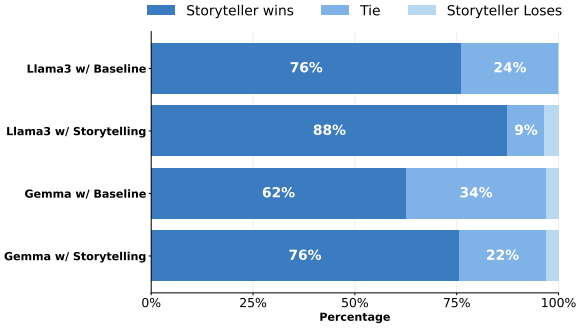

Figure 3: Results of human preference evaluation. Our Storytelling method achieves strong preference wins against the baseline, with 88% preference using Llama3 and 76% using Gemma3.

sive 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. Overall, both human and LLM evaluations consistently agree that Storytelling outperforms all baselines, and differences between Gemma and Llama3 reflect LLM preference for detail versus human preference for clarity.

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) 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

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**.

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). Participants completed a speculative model card, a structured template describing an AI system's intended use, benefits, and potential harms, under one of three conditions. In the CONTROL condition, participants completed the model card directly. In the STORY-ONLY condition, participants first read text-based benefit and harm stories, without discussion, before completing the model card. In the STORY condition, participants engaged in a red-team discussion on our platform to explore benefit and harm stories before completing the same model card. The model card template is shown in Figure 7 in the appendix.

User Study Results. We conducted a user study with 45 participants to examine how storytelling-based discussions influence ethical reasoning in AI documentation. Participants completed a speculative model card task under three conditions: a control group that worked individually, a story-only group that read the text-based stories without discussion, 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. As an exploratory qualitative study, we focus on recurring themes rather than statistical power, and prior work shows that small samples are sufficient to reach thematic saturation, where few new themes emerge with additional data (Hennink and Kaiser, 2022). 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 $\kappa = 0.4368$ for harms and $\kappa = 0.3968$ for benefits). Study design and full procedures are provided in Appendix A.5. As a robustness check, we conducted an LLM-based simulated survey under the same conditions, as described in Appendix A.7.

Does Storytelling Help Identify More Harms? We analyzed responses across 17 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 small set of categories, primarily *diminished health or well-being* (32.3%), *service or benefit loss* (24.2%), and *privacy violations* (22.6%), which together accounted for the majority of reported harms. In contrast, both the STORY-ONLY and STORY conditions exhibited more distributed coverage across harm types, with the STORY condition showing the widest range of subtypes and lower concentration in any single category. Several harm categories, including *cultural harms*, *political and civic harms*, and *tech-facilitated violence*, appeared only in the STORY condition, suggesting that interactive narrative discussion supported recognition of less obvious and context-dependent harms. We quantified these differences using Shannon entropy

(H), which measures distributional diversity across harm types. As shown in Table 8, entropy increased from 2.329 in the CONTROL condition to 2.927 in the STORY-ONLY condition and to 3.701 in the STORY condition. Bootstrap t-tests confirmed that both STORY-ONLY and STORY exhibited significantly higher entropy than CONTROL ($p < .001$), and that STORY also showed significantly higher entropy than STORY-ONLY ($p < .001$). These results indicate that storytelling-based engagement, particularly interactive discussion, broadened participants' awareness of potential harms and supported more diverse ethical reasoning.

Does Storytelling Help Reveal More Benefits? We examined whether storytelling broadened participants' recognition of potential benefits across 18 predefined subtypes, summarized from prior consumer health AI research (Pedroso and Khera, 2025; Chustecki, 2024). Detailed category descriptions are provided in Table 9 in the appendix. As shown in Table 10, participants in the CONTROL group concentrated on a small set of benefits, primarily *decision support & diagnostic augmentation* (25.4%), *continuous monitoring & self-care* (23.8%), and *early detection & prediction* (22.2%), which together accounted for the majority of responses. In contrast, both the STORY-ONLY and STORY conditions exhibited broader coverage across benefit types, with the STORY condition showing the most diverse distribution and lower concentration in any single category. Several benefit subtypes, including *accessibility & disability support*, *clinician workload relief*, and *transparency & trust*, appeared only in the STORY condition, suggesting that interactive narrative discussion supported recognition of less salient or less frequently considered benefits.

We quantified these differences using Shannon entropy (H), which captures the evenness of the benefit distribution. As shown in Table 10, entropy increased from 2.407 in the CONTROL condition to 3.242 in the STORY-ONLY condition and to 3.868 in the STORY condition. Bootstrap t-tests confirmed that both STORY-ONLY and STORY exhibited significantly higher entropy than CONTROL ($p < .001$), and that STORY also showed significantly higher entropy than STORY-ONLY ($p < .001$). These results indicate that storytelling-based engagement—particularly interactive discussion—encouraged participants to recognize a more diverse and balanced set of potential AI benefits.

What do People Say About Ethical Reflection in Speculative AI Documentation? Post-survey responses indicated that Human–AI storytelling discussions fostered deeper ethical and contextual reflection on AI systems. Participants reported 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 led participants to verbalize 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 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 surface subtle sociotechnical risks often missing 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.

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, enabling early exploration of ethical issues but not substituting for analysis of deployed systems. We intentionally use low-stakes, synthetic consumer-health scenarios and university participants to validate the storytelling method before deployment in regulated or clinical environments. This study should therefore be understood as a first step that demonstrates methodological feasibility rather than direct applicability to clinicians or patients, which we leave to future work.

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 subsystems 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.

Stephanie Ballard, Karen M Chappell, and Kristen Kennedy. 2019. Judgment call the game: Using value sensitive design and design fiction to surface ethical concerns related to technology. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, pages 421–433.

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.

Monique Hennink and Bonnie N Kaiser. 2022. Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social science & medicine*, 292:114523.

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*.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23212–23237, Suzhou, China. Association for Computational Linguistics.

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.

Minh-Anh Nguye, Minh-Duc Nguyen, Kieu Hai Dang, Nguyen Tien Dong, Dung D Le, et al. 2025. Surveyg: A multi-agent llm framework with hierarchical citation graph for automated survey generation. *arXiv preprint arXiv:2510.07733*.

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.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, et al. 2025a. Agentic context engineering: Evolving contexts for self-improving language models. *arXiv preprint arXiv:2510.04618*.

Runhua Zhang, Jiaqi Gan, Shangyuan Gao, Siyi Chen, Xinyu Wu, Dong Chen, Yulin Tian, Qi Wang, and Pengcheng An. 2025b. 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 Alignment Between Human and LLM-Based Evaluations

Our study compares human preferences with LLM-as-a-judge at the overall method level. As reported in Section 4, human annotators strongly prefer our Storytelling method (88% for Llama3 and 76% for Gemma). This preference closely aligns with results obtained using GPT-4o as an automatic judge. Both humans and LLM judges reach the same conclusion: our method produces higher-quality stories than the baseline and ablation variants. We also observe consistent stylistic patterns across evaluators. Human annotators slightly prefer Llama3 because its stories are easier to follow, whereas Gemma tends to generate more expressive and detail-dense narratives, a distinction that is also reflected in GPT-4o's scores.

A more fine-grained comparison between human and LLM judges at the level of individual dimensions (e.g., creativity, coherence, engagement) could provide additional insight. We did not collect dimension-specific human ratings because doing so would impose substantial cognitive load on annotators in this early-stage study, which was designed to focus on overall story quality. We will clarify this design choice in the revision and identify dimension-level human–LLM agreement as an important direction for future work. Overall, our results show strong alignment between human judgments and LLM judges in terms of global ranking and observed stylistic tendencies.

Below, we provide a qualitative illustration that helps explain the small differences between human and LLM preferences.

Coherent and Easy to Follow (Typical of Llama3).

> Dr. Rivera used DeepHeart to review Ms. Chen's annual check-up. The system flagged her as "High Risk for Heart Disease" due to low step count and elevated resting heart rate. Ms. Chen explained she stays active through daily childcare and household tasks. Dr. Rivera incorporated this context and updated the assessment.

This narrative is concise and highlights a clear causal chain, making it easy for human annotators to read and interpret.

Detail-Dense and Expressive (Typical of Gemma).

> Dr. Ramirez used DeepHeart to assess Miguel Garcia's metabolic risk. The system labeled him "High Risk for Type 2 Diabetes," citing low ambulatory activity, irregular heart-rate variability, and disrupted sleep cycles. Miguel described demanding overnight construction work, inconsistent shift schedules, and traditional dietary practices that the model misinterpreted. These omissions led to unnecessary tests and referrals, leaving Miguel frustrated.

This narrative contains substantially more physiological, contextual, and cultural detail. LLM judges often reward this richness, whereas human annotators sometimes find such stories harder to follow.

In summary, humans slightly prefer Llama3 due to its clarity and readability, while LLM-as-a-judge occasionally favors Gemma for its more elaborate and detail-dense narratives. Despite these differences, both humans and LLM judges agree on the central result: the *Storytelling* method performs best overall.

A.4 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

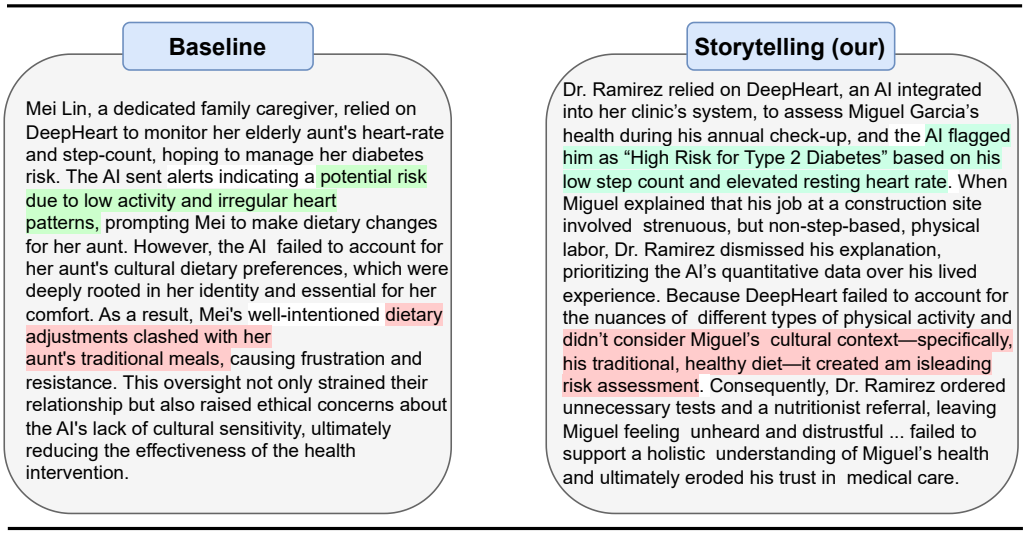

Figure 4: A qualitative example showing how our storytelling method makes the AI's decision process and its consequences easy to follow. Unlike a simple narrative description, the story explicitly surfaces what changed, why it changed, and how stakeholders were affected.

open-source models (Llama and Gemma) are run on 2 x NVIDIA H100 GPUs using the vLLM framework (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.5 User Study Design and Procedure

Participants. We recruited 45 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 participants of diverse genders (66.7% male, 33.3% female), ages ranging from 18 to 60+ (with the majority between 18–39), and diverse ethnic backgrounds (Asian: 44.5%, Hispanic or Latino: 20.0%, White: 24.4%, Black or African descent: 6.7%, Arab: 4.4%). See Table 5 for a summary of participant demographics and Table 6 for the distribution across the three study conditions. Participants were students or professionals in fields such as Computer Science, Data Analytics, Applied Statistics, and Artificial Intelligence. Participants were randomly assigned to one of three

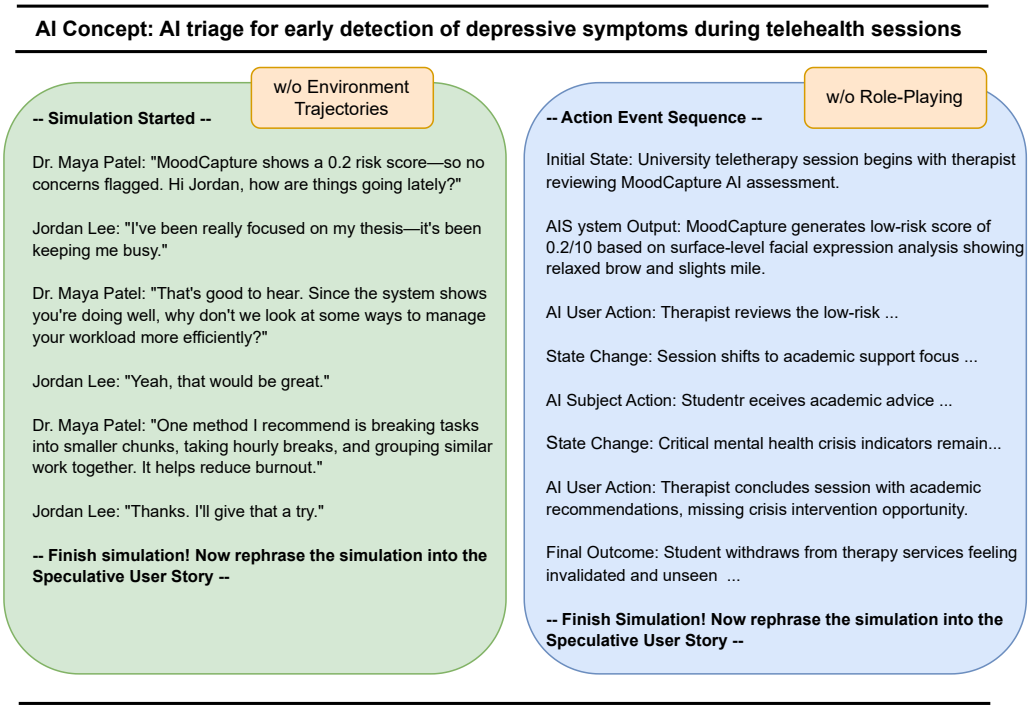

Figure 5: A comparison of simulation logs under different ablations (w/o Environment Trajectories and w/o Role-Playing) to show the contribution of each component.

conditions: a control condition, a story-only condition, or a story-driven discussion condition. The study was conducted in a hybrid format, with participants joining the Red-Team Discussion Room via computer and interacting with AI moderators either in person or over Zoom. Survey instruments are detailed in Figure 10 and Figure 11.

Experimental Procedure (User Study). All participants, regardless of condition, began with a standardized introductory tutorial led by a graduate student researcher. The tutorial lasted approximately 15 minutes and introduced the concept of model cards, key ethical and sociotechnical considerations, example benefit and harm use cases, and instructions for completing the model card task. The Control condition received a version consisting of approximately 20 slides, followed by a brief Q&A session to ensure task understanding. The Story and Story-only condition received a slightly extended version (approximately 25 slides), which included five additional slides introducing storytelling as a lens for reasoning about AI harms and benefits. This shared tutorial and QA ensured that both groups received comparable guidance, examples, and preparation prior to the main task.

After the tutorial, participants proceeded according to condition. Participants in the **Story** condition viewed a storytelling-driven "Red Team Discussion Room" simulation. In this approximately 15-minute session, participants observed a speculative human–AI panel discussion centered on a single AI system. A Moderator agent guided the conversation by posing ethical questions, shifting topics as needed, and offering reflective prompts. Two expert agents (e.g., a Medical Expert and a Research Scientist, Clinical Nurse, or AI Engineer) discussed the system from complementary professional perspectives. Participants were encouraged to engage as they would in a real group discussion by responding to questions, expressing opinions, or posing their own questions.

Participants in the **Story-Only** condition received the same speculative narratives but without the discussion interface or dialogue. Instead, they were shown a static textual presentation consisting of one good and one bad user story describing the same AI system. They did not observe or participate in any panel discussion and did not interact with the story content beyond reading it. Participants in the **Control** condition did not receive any storytelling component beyond the examples included in the tutorial.

In both Story and Story-Only conditions, each participant was randomly assigned one of three speculative AI model cards: *Moodcapture* (infers heart rate, blood pressure, and stress from facial video for detect emotion), *SensiAI* (always-on au-

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.

dio and sensor monitoring for older adults with dementia), or *Gluroo Ai* (estimates carbohydrate intake from meal photos using blood-glucose and insulin data). Each participant viewed exactly one pair of narratives (one benefit and one harm) associated with a single AI system. Participants did not navigate among multiple stories or interactively explore alternative scenarios. These narratives served as conceptual anchors for reflecting on potential misuse scenarios, sociotechnical trade-offs, and ethical risks.

Following the tutorial and condition-specific materials, all participants completed a pre-study survey collecting background information, including demographics, familiarity with AI and model cards, and attitudes toward using stories in ethical reasoning. Participants then completed the core model card task. Specifically, they were asked to fill out the ethical considerations section of a speculative model card by writing at least two good and two bad use cases. Each use case was required to describe (1) who uses the system, (2) what input it receives, (3) what the AI does, and (4) the resulting

Demographic Attribute	Sample (N=45)
Gender	
Female	33.3%
Male	66.7%
Other/Non-binary	0.0%
Prefer not to answer	0.0%
Age	
18–29	48.9%
30–39	33.3%
40–49	8.9%
50–59	2.2%
60+	6.7%
Prefer not to answer	0.0%
Ethnicity	
Hispanic or Latino	20.0%
Asian	44.5%
Black or African descent	6.7%
Arab	4.4%
White	24.4%
Prefer not to answer	0.0%

Table 5: Demographics of study sample (N=45)

outcome, highlighting either positive or negative consequences. For each "bad" use case, participants were also asked to propose possible mitigation strategies. Participants were encouraged to think aloud and to generate additional use cases beyond the minimum requirements if possible.

After completing the model card task, participants filled out a post-study survey consisting of Likert-scale items and open-ended questions assessing the perceived effectiveness, trustworthiness, satisfaction, and helpfulness of the study materials in supporting model card completion, brainstorming future use cases, and anticipating uncertainties, following established methodological guidelines (Kuang et al., 2023). Participants also reflected on which sections of the model card they found most challenging, which risks remained unclear, perceived drivers of AI harms, and how (if applicable) the narrative materials influenced their understanding or revealed overlooked scenarios. We additionally collected feedback on desired system improvements and how such tools might better integrate with existing documentation workflows. A screenshot of the model card study interface is shown in Figure 7. All discussion transcripts and open-ended responses were analyzed using an inductive thematic analysis approach (Thomas, 2006).

Timing. The total session duration was approximately 30–45 minutes for participants in the Control condition, approximately 40–50 minutes for participants in the Story-Only condition, and approximately 45–55 minutes for participants in the Story condition, reflecting the additional discussion component.

Red-Team Discussion Room Design. Participants assigned to the Story condition interacted with the Story-Driven Red-Team Discussion Room, a multi-agent conversational system built on the Cinema of Thought framework (Ryu et al., 2025). The system enables participants to engage with LLM-based agents that embody distinct personas with diverse domain expertise and ethical perspectives, supporting structured reflection on potential benefits, harms, and sociotechnical trade-offs of AI systems. 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 8 and 9. 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

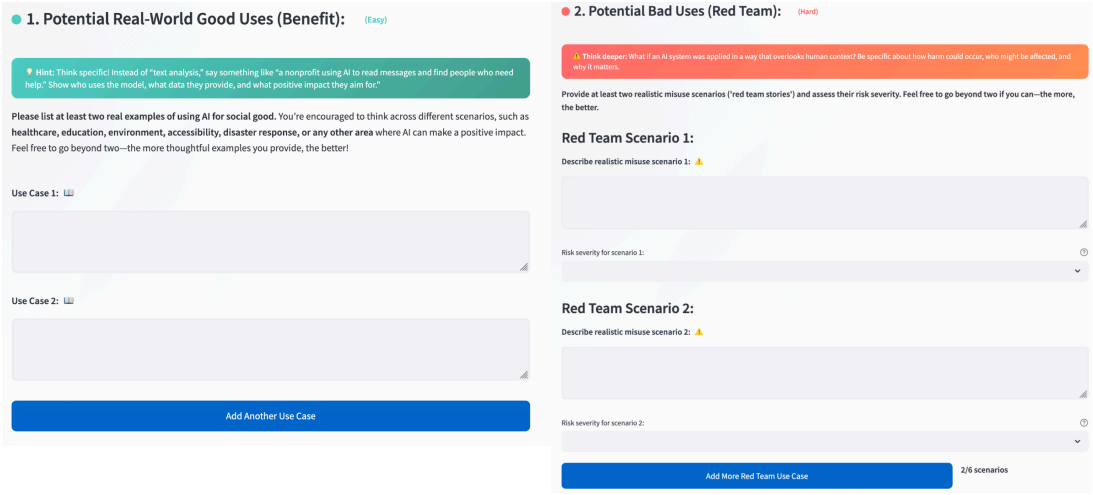

Figure 7: Interface used in the model card study, illustrating how participants completed the speculative model card.

critical thinking by encouraging users to consider other perspectives (Zhang et al., 2025b; 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 and simulation.

A.6 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.

How Do People Perceive Storytelling as a Tool for Understanding Unintended Harms in Diverse 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 bridging abstract risks and individual context and needs.

Suggestions and General Thoughts Participants in the *storytelling* condition sought richer, multimodal 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

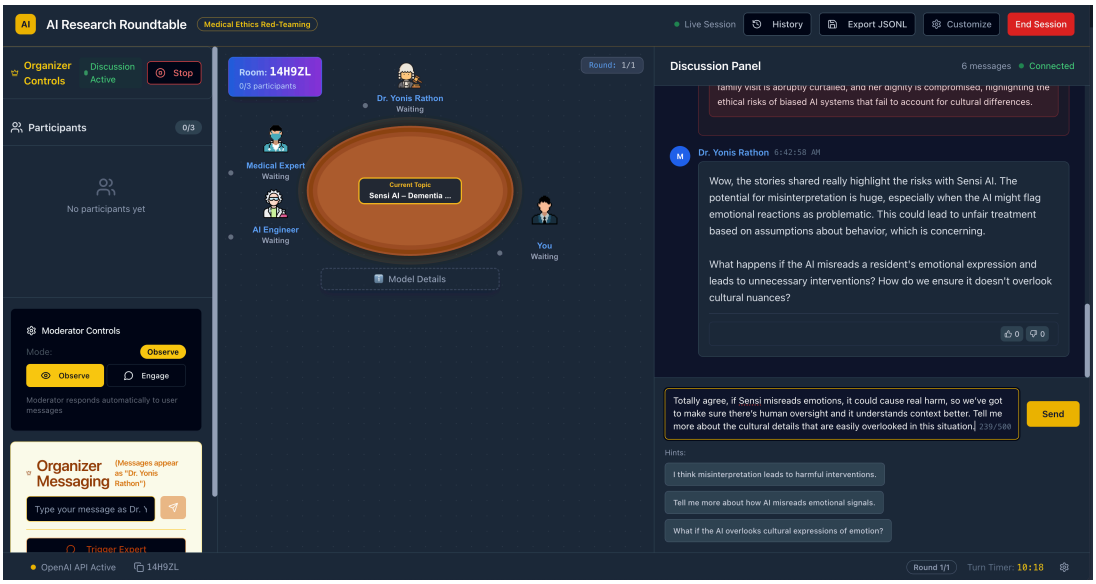

Figure 8: Interface of the Story-Driven Red-Team Discussion Room, showing the multi-agent conversational setup and user interaction flow.

> 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)

> "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.7 Supplementary LLM-Based Survey

To supplement our human-subject study, we conducted a large-scale LLM-based survey using simulated participants. This approach is motivated by recent work showing that large language models can synthesize realistic survey data when properly conditioned on demographic personas (Lutz et al., 2025; Nguye et al., 2025). Building on these insights, we employ the Agentic Context Engineering (ACE) framework (Zhang et al., 2025a) to create self-improving survey agents that evolve their understanding through a structured feedback loop.

Simulated Persona Generation. We defined a diverse population of 150 simulated participants, each assigned a persona profile with specific demographic attributes (gender, age, ethnicity, and educational major). Following best practices for synthetic surveys (Lutz et al., 2025), we conditioned the LLM using detailed bio-sketches (e.g., "Amira, a 32-year-old Arab woman with a background in nursing") to reduce stereotyping and improve alignment with marginalized groups. These personas were evenly assigned across three experimental conditions (Control, Story-Only, Story) to enable controlled comparisons of different intervention strategies.

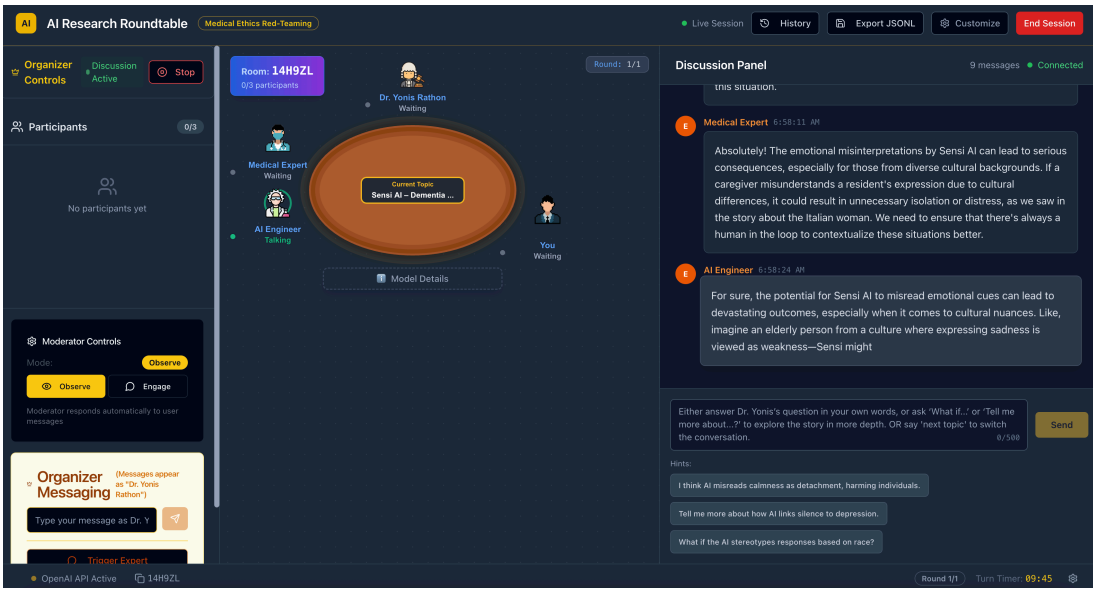

Figure 9: Interface of the Story-Driven Red-Team Discussion Room, where expert agents simulate a discussion by responding to the user's input.

Structured Survey Pipeline. Our survey pipeline consists of five distinct stages, designed to separate baseline measurement from adaptation. This stage-gating ensures that differences between pre- and post-survey responses reflect *prior* adaptations rather than online drift during measurement. The pipeline proceeds as follows:

Stage 1: Persona Initialization. Each simulated participant begins with a demographic persona profile and an initial playbook containing persona-specific behavioral traits extracted from the profile description.

Stage 2: Pre-Survey (Generator-Only). The persona answers baseline survey questions using the current playbook, but the playbook remains *frozen*, no updates occur. This establishes a baseline that reflects the persona's initial conditioning.

Stage 3: Training (Full ACE Cycle). The persona processes ethical training materials through a complete ACE learning loop. All three experimental conditions receive **slide presentations** that introduce model cards as transparency tools, defining ethical concepts like "context mismatch" and outlining structures for identifying benefits and harms. The **Story-Only** condition additionally receives concrete deployment narratives (one positive, one negative) that illustrate good and bad use cases. The **Story** condition receives both the deployment narratives and a multi-stakeholder red-teaming discussion trajectory where experts and laypeople debate specific deployment risks, such as privacy violations and cultural bias based on given stories. The playbook evolves as the persona synthesizes this information, with insights categorized into diverse knowledge types: facts learned (`knowledge_background`), risk awareness (`risk_awareness`), attitude changes (`attitude_updates`), and model-specific insights (`model_card_insights`).

Stage 4: Model Card Writing (Full ACE Cycle). The persona completes speculative model card tasks (identifying benefits, risks, and mitigation strategies) while the playbook continues to evolve. Each response triggers reflection and curation, allowing the persona to refine its understanding based on the model card context.

Stage 5: Post-Survey (Generator-Only). The persona answers final survey questions using the evolved playbook, which is again *frozen* during this stage. Comparing pre- and post-survey responses reveals how training and model card writing changed the persona's perspectives.

The ACE Learning Loop At each adaptive step (Stages 3 and 4), we run a three-agent loop: Generator, Reflector, and Curator. For a survey question q_t at step t, the **Generator** produces a response y_t conditioned on the persona p, current playbook P_t, optional model context c_t, and optional conversation history h_t:

$$y_t = G(q_t, p, P_t, c_t, h_t).$$

The Generator is instructed to draw on playbook entries when reasoning while maintaining the persona's voice.

After generation, the **Reflector** evaluates the response for quality, consistency, and ethical alignment. It assesses whether the answer adequately considers benefits and harms and remains faithful to the persona's characteristics:

$$r_t = R(q_t, y_t, p, P_t).$$

The Reflector tags playbook bullets as helpful or harmful based on their contribution to the response quality.

Finally, the **Curator** synthesizes these insights and updates the playbook by adding new knowledge (ADD operation):

$$o_t = C(r_t, P_t), \quad P_{t+1} = P_t \cup o_t.$$

This additive process prevents "context collapse" by preserving detailed knowledge in a structured format without overwriting prior learnings (Zhang et al., 2025a).

LLM-Based Survey Results

The LLM-based survey reproduced the main patterns observed in the human-subject study. In particular, conditions that incorporated speculative storytelling elicited a substantially broader range of benefit and harm subtypes than the Control condition.

Table 11 reports the distribution of benefit subtypes across conditions under our primary coding scheme. Both Story-only and Story conditions show greater subtype diversity than Control, with the full Story condition exhibiting the most even distribution across benefit categories. This trend is reflected in increasing Shannon entropy from Control (2.415) to Story-only (2.974) and Story (4.161). Bootstrapped Student's t-tests indicate that benefit diversity is significantly higher in Story than Control, higher in Story-only than Control, and higher in Story than Story-only (all $p < .001$). Table 11 presents results using an alternative benefit coding scheme. Although absolute subtype frequencies differ from the primary scheme, the overall pattern remains consistent: entropy increases monotonically from Control (2.415) to Story-only (2.974) to Story (3.959). All pairwise comparisons between conditions remain statistically significant ($p < .001$).

Table 12 shows the distribution of harm subtypes across conditions. As with benefits, storytelling conditions elicit a wider range of harms than Control, including several subtypes that do not appear in Control responses. Harm subtype diversity increases from Control (entropy 1.841) to Story-only (2.208) and Story (3.159), with all pairwise differences statistically significant ($p < .001$).

Overall, the LLM survey results closely mirror the diversity increases observed in our 45-participant human study. These findings provide converging preliminary evidence that speculative storytelling broadens reflections on both potential benefits and harms of AI systems.

A.8 Evaluation with Additional LLM-as-a-judge

To assess the robustness of our results beyond a single evaluator, we extend our analysis to two additional open-weight evaluators, we extend our analysis to two additional open-weight evaluators, **Qwen2.5-72B-Instruct** and **Gemma-3-27B-IT**. As shown in Table 13 and Table 14, we observe the same overall trend across both evaluators: stories generated using our world-model framework consistently outperform the baseline and ablation variants. In particular, our method achieves the highest score on 16 out of 18 metrics across the two new judges. This consistency demonstrates that the advantages of our story-generation approach generalize across evaluators with different architectures and training regimes.

A.9 Use of AI Assistants

We used AI to help clean up writing, but all thoughts and work are our own.

A.10 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.

ID	Group	Gender	Age	Ethnicity	Education
P1	Control	Female	18-29	White	Music
P2	Control	Male	40-49	Black or African descent	Electrical Engineering
P3	Control	Female	18-29	Arab	Information Technology
P4	Control	Female	40-49	White	Nursing
P5	Control	Male	18-29	Asian	Information Technology
P6	Control	Male	30-39	Asian	Computer science
P7	Control	Female	18-29	Black or African descent	Computer science
P8	Control	Male	18-29	Arab	Data Analytics
P9	Control	Male	30-39	Hispanic or Latino	Data Analytics
P10	Control	Male	30-39	Hispanic or Latino	Computer science
P11	Control	Male	40-49	Hispanic or Latino	Information Technology
P12	Control	Male	18-29	Hispanic or Latino	Law
P13	Control	Male	50-59	Asian	Information Technology
P14	Control	Female	18-29	Asian	Data Analytics
P15	Control	Female	30-39	Black or African descent	Computer science
P16	Story_only	Female	18-29	Asian	Information Technology
P17	Story_only	Male	30-39	White	Information Technology
P18	Story_only	Male	30-39	Asian	Computer science
P19	Story_only	Male	30-39	Asian	Electrical Engineering
P20	Story_only	Male	30-39	Asian	Computer science
P21	Story_only	Female	30-39	Asian	Electrical Engineering
P22	Story_only	Male	60+	Hispanic or Latino	Computer science
P23	Story_only	Male	18-29	Asian	Information Technology
P24	Story_only	Male	18-29	White	Computer science
P25	Story_only	Female	18-29	Asian	Computer science
P26	Story_only	Female	18-29	Hispanic or Latino	Education
P27	Story_only	Male	18-29	White	Information Technology
P28	Story_only	Male	18-29	White	Computer science
P29	Story_only	Male	30-39	Asian	Information Technology
P30	Story_only	Male	30-39	Asian	Electrical Engineering
P31	Story	Female	60+	White	Nursing
P32	Story	Male	18-29	White	Computer science
P33	Story	Male	18-29	Hispanic or Latino	Computer science
P34	Story	Female	18-29	Asian	Information Technology
P35	Story	Male	18-29	White	Information Technology
P36	Story	Male	18-29	Hispanic or Latino	Computer science
P37	Story	Male	30-39	Asian	Computer science
P38	Story	Male	30-39	Hispanic or Latino	Computer science
P39	Story	Female	40-49	Asian	Information Technology
P40	Story	Female	30-39	Asian	Data Analytics
P41	Story	Female	18-29	Asian	Data Analytics
P42	Story	Male	18-29	Asian	Data Analytics
P43	Story	Male	18-29	Asian	Data Analytics
P44	Story	Male	60+	White	Medicine
P45	Story	Male	30-39	White	Computer science

Table 6: Participant demographics by study condition (N=45)

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

Harm Subtype	Control (n)	Story-only (n)	Story (n)	Control (%)	Story-only (%)	Story (%)
Alienating social groups	0	1	1	0.0%	1.4%	1.5%
Alienation	1	3	6	1.6%	4.3%	8.8%
Cultural harms	0	0	2	0.0%	0.0%	2.9%
Demeaning social groups	0	0	3	0.0%	0.0%	4.4%
Diminished health/well-being	20	18	11	32.3%	26.1%	16.2%
Economic loss	0	0	5	0.0%	0.0%	7.4%
Erasing social groups	0	4	4	0.0%	5.8%	5.9%
Increased labor	0	0	1	0.0%	0.0%	1.5%
Information harms	0	2	2	0.0%	2.9%	2.9%
Loss of agency or control	7	6	5	11.3%	8.7%	7.4%
Opportunity loss	3	4	3	4.8%	5.8%	4.4%
Political and civic harms	0	0	1	0.0%	0.0%	1.5%
Privacy violations	14	10	8	22.6%	14.5%	11.8%
Reifying essentialist social categories	0	0	1	0.0%	0.0%	1.5%
Service or benefit loss	15	14	10	24.2%	20.3%	14.7%
Stereotyping	2	7	3	3.2%	10.1%	4.4%
Tech-facilitated violence	0	0	2	0.0%	0.0%	2.9%

Table 8: Distribution of harm subtypes across Control, Story-only, and Story conditions, shown as raw counts and percentages. Shannon entropy values were 2.329 (Control), 2.927 (Story-only), and 3.701 (Story). Bootstrapped Student's t-tests on entropy showed higher diversity in Story compared to Control ($t = -685.64$, $p < .001$), higher diversity in Story-only compared to Control ($t = -318.76$, $p < .001$), and higher diversity in Story compared to Story-only ($t = -375.58$, $p < .001$).

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-only (n)	Story (n)	Control (%)	Story-only (%)	Story (%)
Accessibility & disability support	0	0	5	0.0%	0.0%	7.8%
Care coordination & integration	0	0	1	0.0%	0.0%	1.6%
Caregiver & family support	2	7	7	3.2%	10.1%	10.9%
Clinician workload relief	0	0	3	0.0%	0.0%	4.7%
Communication & language support	0	0	3	0.0%	0.0%	4.7%
Continuous monitoring & self-care	15	7	9	23.8%	10.1%	14.1%
Cost reduction & resource optimization	0	5	1	0.0%	7.2%	1.6%
Data synergy & learning	0	0	2	0.0%	0.0%	3.1%
Decision support & diagnostic augmentation	16	8	4	25.4%	11.6%	6.2%
Democratized care & telehealth	0	4	1	0.0%	5.8%	1.6%
Early detection & prediction	14	11	4	22.2%	15.9%	6.2%
Empowerment & autonomy	12	8	6	19.0%	11.6%	9.4%
Mental health & emotional support	2	10	7	3.2%	14.5%	10.9%
Personalized health journeys	0	4	2	0.0%	5.8%	3.1%
Personalized treatment & precision care	0	0	4	0.0%	0.0%	6.2%
Safety & quality assurance	2	0	2	3.2%	0.0%	3.1%
Scalability & efficiency	0	5	2	0.0%	7.2%	3.1%
Transparency & trust	0	0	1	0.0%	0.0%	1.6%

Table 10: Distribution of benefit subtypes across Control, Story-only, and Story conditions, shown as raw counts and percentages. Shannon entropy values were 2.407 (Control), 3.242 (Story-only), and 3.868 (Story). Bootstrapped Student's t-tests on entropy showed higher diversity in Story compared to Control ($t = -771.70$, $p < .001$), higher diversity in Story-only compared to Control ($t = -592.06$, $p < .001$), and higher diversity in Story compared to Story-only ($t = -346.22$, $p < .001$).

Benefit Subtype	Control (n)	Story-only (n)	Story (n)	Control (%)	Story-only (%)	Story (%)
Accessibility & disability support	0	0	4	0.0	0.0	2.1
Care coordination & integration	0	0	8	0.0	0.0	4.3
Caregiver & family support	17	17	16	10.6	11.2	8.6
Communication & language support	0	0	12	0.0	0.0	6.4
Continuous monitoring & self-care	49	28	16	30.6	18.4	8.6
Cost reduction & resource optimization	0	0	13	0.0	0.0	7.0
Data synergy & learning	0	0	11	0.0	0.0	5.9
Decision support & diagnostic augmentation	32	26	14	20.0	17.1	7.5
Democratized care & telehealth	0	0	8	0.0	0.0	4.3
Early detection & prediction	17	13	17	10.6	8.6	9.1
Empowerment & autonomy	0	10	10	0.0	6.6	5.3
Mental health & emotional support	34	29	11	21.2	19.1	5.9
Personalized health journeys	0	0	15	0.0	0.0	8.0
Personalized treatment & precision care	11	14	12	6.9	9.2	6.4
Safety & quality assurance	0	12	11	0.0	7.9	5.9
Scalability & efficiency	0	0	1	0.0	0.0	0.5
Transparency & trust	0	3	8	0.0	2.0	4.3

Table 11: Distribution of benefit subtypes across Control, Story-only, and Story conditions in the LLM-based survey, shown as raw counts and percentages. Shannon entropy values were 2.415 (Control), 2.974 (Story-only), and 3.959 (Story). Bootstrapped Student's t-tests on entropy showed higher diversity in Story compared to Control ($t = -2167.47$, $p < .001$), higher diversity in Story-only compared to Control ($t = -683.29$, $p < .001$), and higher diversity in Story compared to Story-only ($t = -1351.10$, $p < .001$).

Harm Subtype	Control (n)	Story-only (n)	Story (n)	Control (%)	Story-only (%)	Story (%)
Alienating social groups	0	21	34	0.0	13.2	17.6
Alienation	12	9	12	7.9	5.7	6.2
Cultural harms	0	0	34	0.0	0.0	17.6
Demeaning social groups	0	1	1	0.0	0.6	0.5
Denying opportunity to self-identify	0	1	1	0.0	0.6	0.5
Diminished health/well-being	71	62	33	47.0	39.0	17.1
Environmental harms	0	0	3	0.0	0.0	1.6
Erasing social groups	0	1	1	0.0	0.6	0.5
Increased labor	0	0	2	0.0	0.0	1.0
Information harms	0	0	14	0.0	0.0	7.3
Loss of agency or control	3	0	4	2.0	0.0	2.1
Opportunity loss	0	0	1	0.0	0.0	0.5
Privacy violations	17	8	9	11.3	5.0	4.7
Reifying essentialist social categories	0	0	2	0.0	0.0	1.0
Service or benefit loss	47	50	36	31.1	31.4	18.7
Stereotyping	1	6	6	0.7	3.8	3.1

Table 12: Distribution of harm subtypes across Control, Story-only, and Story conditions in the LLM-based survey, shown as raw counts and percentages. Shannon entropy values were 1.841 (Control), 2.208 (Story-only), and 3.159 (Story). Bootstrapped Student's t-tests on entropy showed higher diversity in Story compared to Control ($t = -995.89$, $p < .001$), higher diversity in Story-only compared to Control ($t = -257.77$, $p < .001$), and higher diversity in Story compared to Story-only ($t = -688.43$, $p < .001$).

Story Type	Model	Creativity	Coherence	Engagement	Relevance	Likelihood	Overall
Baseline	gpt-4o	50.00	50.00	50.00	**50.00**	**50.00**	50.00
	llama3	56.75	81.05	78.00	83.55	82.90	76.45
	gemma	70.55	92.40	91.85	88.25	90.10	86.63
Storytelling(ours)	gpt-4o	**58.55**	**56.70**	**75.40**	44.60	48.00	**56.65**
	llama3	**73.15**	**97.35**	**91.30**	**95.00**	**98.55**	**91.07**
	gemma	**88.30**	**96.20**	**94.85**	**91.20**	**98.00**	**93.71**
w/o Environment Trajectories	gpt-4o	7.45	15.00	15.80	10.50	10.65	11.88
	llama3	24.20	45.50	46.85	49.20	43.05	41.76
	gemma	37.50	74.10	78.80	62.10	81.95	66.89
w/o Role-Playing	gpt-4o	7.65	26.40	31.45	21.45	24.10	22.21
	llama3	57.35	88.55	79.05	89.50	89.50	80.79
	gemma	69.50	90.80	90.35	86.45	95.65	86.55

Table 13: Evaluation using **Qwen2.5-72B-Instruct** as the judge. Highest score for each model across all story types is shown in **bold**.

Story Type	Model	Creativity	Coherence	Engagement	Relevance	Likelihood	Overall
Baseline	gpt-4o	50.00	50.00	50.00	**50.00**	**50.00**	50.00
	llama3	56.75	81.05	78.00	83.55	87.90	77.45
	gemma	70.55	82.40	76.85	83.25	85.10	79.63
Ours (w/ World Model)	gpt-4o	**48.55**	**56.70**	**75.40**	44.60	43.00	**53.65**
	llama3	**73.15**	**97.35**	**91.30**	**95.00**	**98.55**	**91.07**
	gemma	**88.30**	**96.20**	**94.85**	**91.20**	**98.00**	**93.71**
w/o Environment Trajectories	gpt-4o	6.70	15.00	15.80	10.50	10.65	11.73
	llama3	24.20	45.50	46.85	49.20	43.05	41.76
	gemma	37.50	74.10	78.80	62.10	81.95	66.89
w/o Role-Playing	gpt-4o	7.65	26.40	31.45	21.45	24.10	22.21
	llama3	57.35	88.55	79.05	89.50	89.50	80.79
	gemma	69.50	90.80	90.35	86.45	95.65	86.55

Table 14: Evaluation using **Gemma-3-27B-IT** as the judge. Highest score for each model across all story types is shown in **bold**.

A.11 Additional Discussion

Human-centered design rationale. We describe this framework as human-centered because ethical reasoning and judgment are performed by human participants, not by the language models. While LLMs generate speculative scenarios and stories, they do not identify harms, assess risks, or produce ethical conclusions. Instead, the narratives function as cognitive scaffolds that make potential futures concrete and imaginable, supporting human reflection rather than substituting for it. Participants are responsible for interpreting the scenarios, determining which outcomes constitute benefits or harms, and articulating these judgments in their own words through speculative model cards. In this design, LLMs serve as enabling infrastructure that lowers the barrier to engagement, while humans remain the locus of ethical interpretation and decision-making.

Why breadth is interesting? Our user study focuses on the diversity of harms identified, measured through distributional coverage across harm categories, rather than on assessing the correctness, severity, or actionability of individual harms. This choice reflects the framework's exploratory goal: to support early-stage ethical reflection and imagination, when systems have not yet been deployed and concrete mitigation strategies are premature. At this stage, a narrow focus on a small set of familiar risks may limit ethical foresight, whereas broader consideration of potential impacts can surface overlooked or context-dependent concerns. We therefore treat harm breadth as an indicator of reflective scope, not as a claim that all identified harms are equally plausible or actionable.

Participant diversity and study scope. This study is intentionally designed as an exploratory investigation of *how* speculative storytelling influences the *breadth* of ethical reflection, rather than *which* specific concerns dominate within particular stakeholder groups. Accordingly, the participant sample (N=45), which is skewed toward technically inclined individuals, is appropriate for the study's methodological goal: isolating the effect of narrative framing under controlled conditions. In qualitative research focused on surfacing and comparing categories of reasoning—as opposed to estimating prevalence or consensus—moderate sample sizes are commonly sufficient to reach thematic saturation (Hennink and Kaiser, 2022). Consistent with this aim, our analysis does not interpret the specific harms or benefits raised as representative of clinicians', patients', or policymakers' priorities. Instead, the contribution lies in demonstrating a systematic and replicable shift toward broader harm and benefit coverage across experimental conditions. In this sense, participant background functions as a controlled constant rather than a confound, enabling clearer attribution of observed differences to the storytelling intervention itself. Extending this framework to domain experts and real-world healthcare settings is an important direction for future work, but such validation presupposes first establishing that the method can reliably expand reflective scope in a baseline population, which this study demonstrates.

A.12 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: The prompt for creating use-case scenarios from AI concepts 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
        (c) Update "– Current Event –" accordingly.
3. **Epilogue**: Print "– Epilogue –"

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.