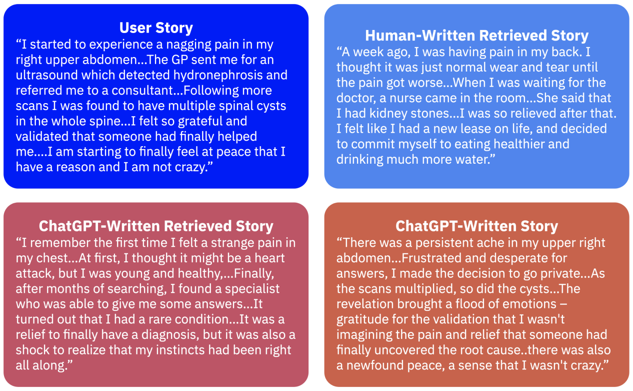
# Stories Without Heart? Empathy towards Human vs AI Stories is Modulated by Author Disclosure

**Anonymous**

## Abstract

People often empathize with one another via stories and sharing of experiences. With the rise of generative AI systems that interact with us in daily life, it is important to understand how empathy unfolds toward stories from human vs. AI narrators and the role transparency plays in this effect. In this work, we conduct four crowd-sourced studies with $N = 985$ participants to understand how empathy shifts across human-written vs AI-written stories. We find that participants consistently and significantly empathize with *human-written over machine-written stories* in almost all conditions, regardless of whether they are aware that an AI wrote the story. We also find that participants reported a greater *willingness* to empathize with AI-written stories if there is transparency about the story author. Our work sheds light on how empathy towards AI or human narrators is tied to the way the story is presented, thus informing ethical considerations of artificial agent interactions that are intended to evoke empathetic reactions.



**User Story**
"I started to experience a nagging pain in my right upper abdomen...The GP sent me for an ultrasound which detected hydronephrosis and referred me to a consultant...Following more scans I was found to have multiple spinal cysts in the whole spine...I felt so grateful and validated that someone had finally helped me....I am starting to finally feel at peace that I have a reason and I am not crazy."

**Human-Written Retrieved Story**
"A week ago, I was having pain in my back. I thought it was just normal wear and tear until the pain got worse...When I was waiting for the doctor, a nurse came in the room...She said that I had kidney stones...I was so relieved after that. I felt like I had a new lease on life, and decided to commit myself to eating healthier and drinking much more water."

**ChatGPT-Written Retrieved Story**
"I remember the first time I felt a strange pain in my chest...At first, I thought it might be a heart attack, but I was young and healthy,...Finally, after months of searching, I found a specialist who was able to give me some answers...It turned out that I had a rare condition...It was a relief to finally have a diagnosis, but it was also a shock to realize that my instincts had been right all along."

**ChatGPT-Written Story**
"There was a persistent ache in my upper right abdomen...Frustrated and desperate for answers, I made the decision to go private...As the scans multiplied, so did the cysts...The revelation brought a flood of emotions – gratitude for the validation that I wasn't imagining the pain and relief that someone had finally uncovered the root cause..there was also a newfound peace, a sense that I wasn't crazy."

Figure 1: Examples of a user story and corresponding retrieved human-written story, pre-generated ChatGPT story retrieved based on the user's story, and on-the-fly generated ChatGPT story. Participants read different stories depending on the study condition, and rate their empathy towards the narrator of the story as well as general willingness to empathize with AI.

## 1 Introduction

Empathy, the sharing of emotions with a social other, is foundational in developing strong interpersonal ties [Decety and Jackson, 2004; Cuff *et al.*, 2016]. People often empathize with others through storytelling and sharing lived experiences, which can often be a powerful force for bridging political and interpersonal divides [Kalla and Broockman, 2021]. As machines are increasingly capable of telling human-like stories in daily life, this raises important questions about how people might empathize with machine-written stories and the ethical implications of empathy towards AI "experiences" [Spitale *et al.*, 2022; Schaaff *et al.*, 2023a]. Humans can breathe life into inanimate or artificial systems [Pelau *et al.*, 2021; Lee *et al.*, 2019; Lv *et al.*, 2022; Chung and Kang, 2023], and are able to relate to fictional experiences when they are human-like or realistic in the scope of one's own life [Oatley, 2016; Djikic *et al.*, 2013]. As such, these questions, by nature, call for ethical and philosophical concerns about differences in empathy towards humans and AI – Machines have no lived experiences, yet can produce stories as their "own" [Abercrombie *et al.*, 2023; Alonso, 2023]. If the machine uses these fabricated experiences to elicit a particular behavior from the user, is this considered manipulation? How are behaviors shifted if the user is aware that the experiences are fabricated?

Since outputs from generative AI are not an artificial agent's actual experiences [Alonso, 2023], but rather a probabilistically sampled sequence of text from human experiences, it is important to be precise and nuanced when communicating results from generated text to ensure ethical deployment of such systems [Brandtzaeg *et al.*, 2022; Croes and Antheunis, 2020]. In the field of social psychology, researchers have explored how nudges, small subtle changes that can inspire big changes in actions, can modulate empathy [Zaki, 2019]. Yet, in the AI domain, few works have explored how subtle design changes in the presentation of AI-written stories can significantly shift attitudes and empathy towards AI systems [Giorgi *et al.*, 2023].

Prior works do, however, generally indicate that perceptions of AI can change depending on transparency. Most works find that knowledge of AI involvement reduces the perception of the agent or quality of interaction and that there are fundamental qualities of "humanness" in texts written by people [Ishowo-Oloko *et al.*, 2019; Giorgi *et al.*, 2023; Straten *et al.*, 2020], but that fostering trust and acceptance can lead to more empathy towards an AI agent [Pelau *et al.*, 2021]. Grounded by these works, we hypothesize that empathy towards AI-written stories, both generated and retrieved in response to a user's own personal story, will be significantly lower than empathy towards human-written stories whether or not the author is disclosed **[H1]**. We hypothesize that people are more willing to empathize with AI stories when the author of the story is made transparent, as the output could be perceived as more trustworthy **[H2]**. To test our hypotheses, in this work, we investigate the following:

1. How does empathy change when stories, human or AI-written, are retrieved vs. generated directly by a language model?
2. How does transparency about the author of a story play a role in empathy towards human vs AI narrators?

To this end, we conduct four crowd-sourced studies with $N = 985$ participants to study how situational empathy (empathy evoked by a specific event) changes when receiving a human-written or AI-written story in response to a user's own story. Through a mixed methods analysis across these measures, we explored how and why empathy unfolds across different story scenarios and use these results to inform ethical discussion of asking people to empathize with the "experiences" of AI.

# 2 Related Work

## 2.1 Empathy with AI

Current AI systems hold the ability to express social and emotional influences through the mechanisms of empathy, which can lead to downstream impacts in the real world. For example, in AI service applications, increased empathy improves service acceptance and user compliance [Adam *et al.*, 2021; Yoon and Lee, 2021]. Such empathetic relationships with synthetic agents can unfold in the following ways: (1) in the behaviour of the agent, where the agent behaves in an empathic way towards other agents and towards the user, and (2) in the relation the agent establishes with the user, where the agent looks like and acts in a way that leads the user to establish an empathic relation with it [Paiva, 2011; Paiva *et al.*, 2004]. Thus, "an agent that is able to, by its behaviour and features, allow the users to build an empathic relation with it."

Various factors for social agents influence their perception as targets of empathy, such as the situation or context that the agent is in, the features of the observer or empathizer, the empathy triggering mechanism used, and characteristics of the agent including degree of agency, physical appearance, expressivity and cognitive capabilities [Kim and Hur, 2023]. Crucially, prior works indicate that acceptance and trust towards AI devices is directly related to how much people empathize with an AI-system [Pelau *et al.*, 2021]. Further works validate that the more personal information AI agents disclose, the more empathetic human conversation partners are towards these agents [Tsumura and Yamada, 2023].

As AI-generated content becomes commonplace, it is important to understand people's empathic reactions towards generative stories and how factors such as transparency modulate these psychological responses. Little work has been done to explore the difference of perceived empathy between stories created by humans and AI. In this work, we study how much people empathize with stories created by AI compared to stories created by other humans as well as how author disclosure affects perceived empathy.

## 2.2 Deception

Some may propose that AI cannot explicitly convey empathy; and if it shares an empathetic motivation, that is considered deception. Deception broadly refers to the act of withholding information or giving false information, often to the benefit of the communicator of the information. According to the taxonomy of Masters et al., chatbots could be deceptive because they seek to *intimidate*, or *generate a simulation indistinguishable from the thing being simulated* [Masters *et al.*, 2021]. Chatbots seek to emulate human dialogue, and though there are characteristics that can reveal machine generation, empirical research has found that it is sometimes difficult to identify if a text has been written by a human or a machine [Sadasivan *et al.*, 2023].

Transparency has been discussed as a tool to mitigate deception of AI systems [Natale, 2021], and in our work, we define transparency as disclosing the author of a text. Research has looked at the impact of author disclosure on various user behaviors. For example, one study showed that people were approximately 79% less likely to purchase a product if they knew the seller was a chatbot [Luo *et al.*, 2019]. Another study found that while a bot was better at inducing cooperation in a game, when the bot was disclosed to the player, they cooperated significantly less with it [Ishowo-Oloko *et al.*, 2019]. In our work, we use this definition of transparency around author disclosure to analyze the effects of empathy told by AI vs human narrators as well as the willingness to empathize with AI storytellers.

# 3 Methods

## 3.1 Study Procedure

We conducted four crowd-sourced studies with a total of $N = 985$ participants to assess the effects of author origin on empathy. Our study was approved by our institution's ethics board as an exempt protocol. Within each session, participants wrote their own personal stories and rated empathy towards stories written by people or by ChatGPT.

The retrieved stories are matched based on similarity of the embeddings of stories, and generated stories are generated on-the-fly given the user's story as a prompt. We used ChatGPT to generate a set of $1,568$ stories using seed stories from the EMPATHICSTORIES dataset [Shen *et al.*, 2023]. Stories generated by ChatGPT were prompted with a context story and the following instruction: *Write a story from your own life that the narrator would empathize with. Do not refer to the narrator explicitly.* The study's four comparisons are as follows:

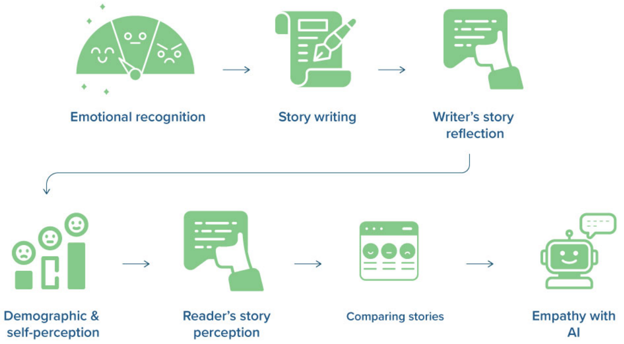

Emotional recognition
Story writing
Writer's story reflection
Demographic & self-perception
Reader's story perception
Comparing stories
Empathy with AI

Figure 2: Flow of user study procedure. Participants identify their own emotions, write their personal story and reflections, then fill out demographic information, their State Empathy, and ranking of stories. Finally, at the end of the study, they rate how willing they are to empathize with AI in general.

- **H-CR:** We compared empathy towards the narrator across *human-written retrieved* stories and ***ChatGPT retrieved*** stories
- **H-CR+T:** We compared empathy towards the narrator across *human-written retrieved* stories and ***ChatGPT retrieved*** stories, making transparent to the user whether the story they read was written by a human or an AI before they rated their empathy (repeat H-CR with transparency)
- **H-CG:** We compared empathy empathy towards the narrator towards *human-written retrieved* stories and ***ChatGPT generated*** stories (in response to the user's story as a context).
- **H-CG+T:** We compared empathy towards the narrator towards *human-written retrieved* stories to ***ChatGPT generated*** stories, making the author of the story transparent (repeat H-CG with transparency)

Finally, in all studies, participants reported how their empathy towards the stories would change if the stories were written by an AI. Examples of stories across conditions are shown in Figure 1.

### Retreival Dataset and Model

Since our study aims to assess differences in empathy towards human vs AI-written stories, both the user's experiences and the stories returned by our system are important. Returning a story at random could undermine the user's experiences and hinder their empathy towards the retrieved story. While many methods exist to retrieve semantically similar pieces of text [Reimers and Gurevych, 2019], few focus on retrieving stories that users would emotionally resonate with given their own story context. As such, we use a fine-tuned BART-base model from Shen et al., which is trained on the EMPATHICSTORIES dataset, a corpus containing pairs of stories each annotated with an "empathic similarity" score from 1-4, where empathic similarity refers to how likely the narrators of both stories would empathize with one another [Shen *et al.*, 2023]. Using this model, we improved retrieval of stories that are empathetically relevant to a user's own personal story.

### User Study Interface

We deployed a web interface similar to a guided journaling app where users write and read personal stories. The interface connects to a server run on a GPU machine (4x Nvidia A40s, 256GB of RAM, and 64 cores), which retrieves story responses in real time. In addition, the server connects the front-end to Firebase Realtime storage in order to track interaction data throughout the course of the study.

### User Story Prompts and Retrieved Stories

To prompt vulnerable and meaningful personal stories from users, we used questions from the Life Story Interview, an approach from social science that gathers key moments from a person's life [Atkinson, 1998]. Stories retrieved by our model were either pulled from the EMPATHICSTORIES dataset ($1,568$ stories) or generated by ChatGPT. In order to ensure topics were constrained to stories present in our retrieval database, we used topic modeling to identify key clusters in the personal narratives from EMPATHICSTORIES. To identify these topics, we used Latent Dirichlet Allocation (LDA) and KeyBERT on the clusters [Grootendorst, 2020]. Users were instructed to reflect on their life in relation to one of the chosen topics.

## 3.2 Participants and Recruitment

We recruited a pool of 985 participants from Prolific.[1] Participants across the studies were predominantly female and white. All participants on average had high trait empathy and neutral arousal and valence prior to starting the study. Full demographic distributions across the four studies are shown in the supplementary material (Table 1).

| | H-CG | H-CR | H-CR+T | H-CG+T |
|---|---|---|---|---|
| **Num Participants** | 300 | 299 | 197 | 189 |
| **Age** | $37.60 \pm 12.54$<br>min : 18<br>max : 75 | $40.18 \pm 14.31$<br>min : 18<br>max : 79 | $40.16 \pm 13.76$<br>min : 19<br>max : 77 | $38.82 \pm 13.52$<br>min : 18<br>max : 79 |
| **Gender** | 173 women, 120 man, 5 non binary, 2 na | 161 women, 132 man, 3 non binary, 3 na | 100 women, 93 men, 2 non binary, 2 na | 111 women, 76 men, 1 non binary, 1 na |
| **Ethnicity** | 228 white, 24 black, 14 asian, 13 other, 10 indian, 5 na, 4 hispanic, 1 middle eastern, 1 native | 242 white, 16 black, 15 asian, 8 na, 7 other, 7 indian, 2 hispanic, 1 middle eastern, 1 islander | 160 white, 20 black, 6 asian, 4 other, 4 indian, 3 na | 145 white, 13 black, 13 indian, 9 asian, 5 other, 2 middle eastern, 1 na, 1 hispanic |

Table 1: Demographic distribution of of all participants across the four studies.

## 3.3 Data Collection and Analysis

At the beginning of the study, we measured the user's valence and arousal, as current emotional state could influence empathy. For our empathy measurement, we used a shortened version of the State Empathy Scale [Shen, 2010], which contains 7 questions covering affective (sharing of others' feelings), cognitive (adopting another's point of view), and associative (identification with others) aspects of situational empathy. Users additionally provided free-text responses about their empathy towards the story as well as multiple choice questions listing reasons why they did or did not empathize

[1] https://www.prolific.com/

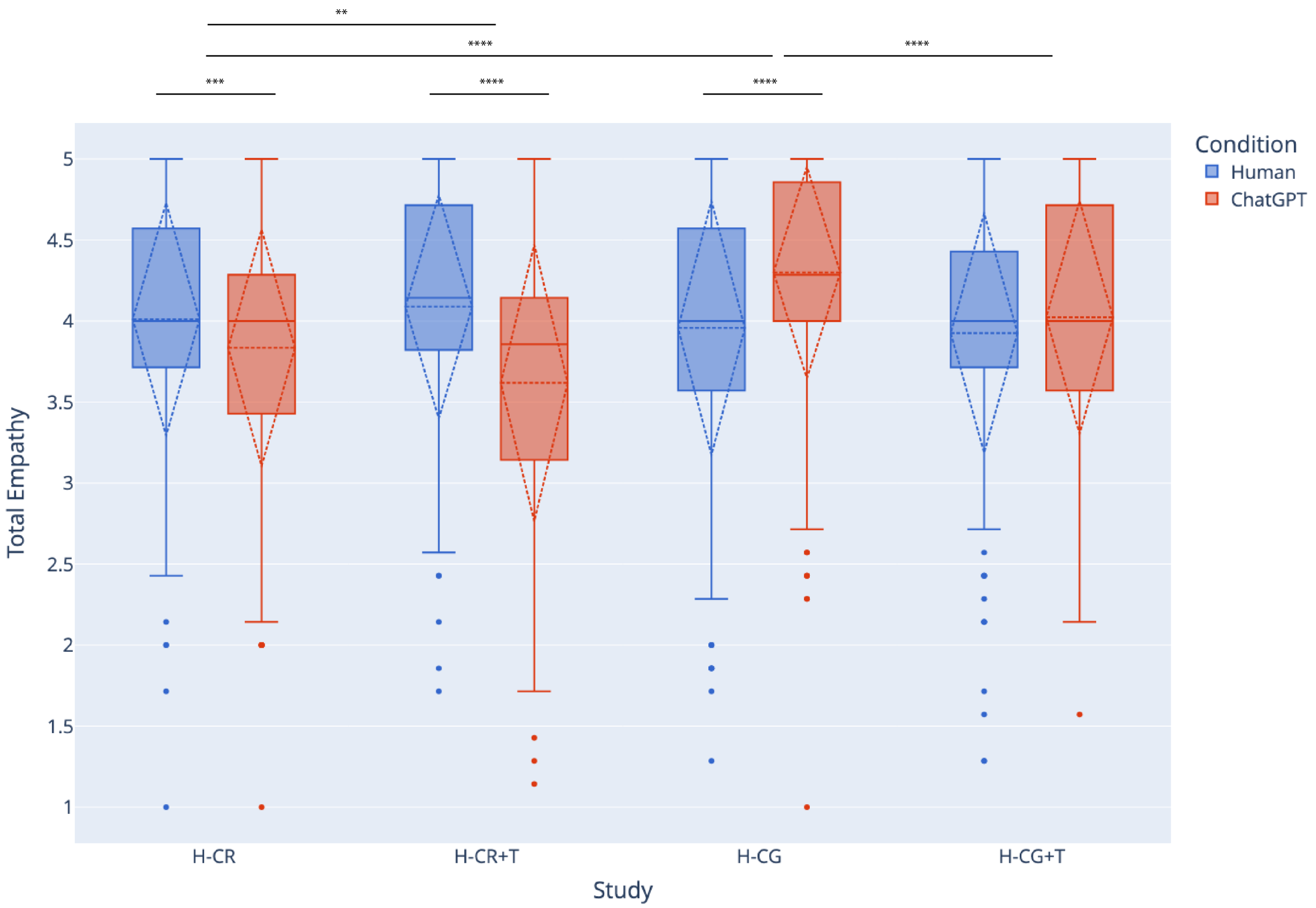


Figure 3: Changes in total empathy towards stories participants read across conditions (human-written vs AI-written story) and studies (author made transparent vs author not transparent, AI story was retrieved vs generated)

with the story (ie. how well-written the story was and how consistently it read). At the end of the study, users self-reported how their empathy would change if the stories they read in the session were written by AI (which we term as perceived empathy with AI).

We used both quantitative and qualitative approaches to understand the effects of empathy towards a story from human vs. AI narrators and offer insights around why empathy shifts under certain conditions. To analyze differences in empathy with the State Empathy Scale, we used a paired t-test, as we identified through a Shapiro-Wilke test that the data is normally distributed. Note that we computed total empathy towards a story using the mean of the State Empathy Scale survey questions. To compare perceived empathy across studies, we used an independent t-test. A full flow of the user study procedure is shown in Figure 2.

For qualitative analysis, open-ended explanations for the empathy rating were thematically coded using an inductive approach [Patton, 2005]. Two researchers independently coded a subset of the data and reached substantial agreement with a Cohen's Kappa value of .70.

# 4 Results

## 4.1 Effects on Empathy towards Stories

**Participants Generally Felt More Empathy for Human-Written Stories than AI-Written Stories.** When we instead retrieve stories from a corpus of narratives generated by ChatGPT (**H-CR**), total empathy decreases across AI-written vs. human-written stories ($t(298) = 3.46, p < 1e-5$,Cohen's $d = 0.24$). This indicates a noticeable difference between human vs. AI-written stories, which we explore further through qualitative analysis in Section 4.3. When we make transparent to the user the author of the retrieved story (**H-CR+T**), we see an even greater decrease in total empathy towards AI-written stories relative to human-written stories ($t(196) = 7.07, p < 1e-10$,Cohen's $d = 0.60$).

When comparing *human retrieved* stories and *ChatGPT generated* stories based on the user's original story (**H-CG**), we find that participants empathize significantly more with ChatGPT generated stories than retrieved human stories ($t(299) = 6.14, p < 1e-08$,Cohen's $d = 0.47$). Following this trend, we find that there is no statistically significant dif-

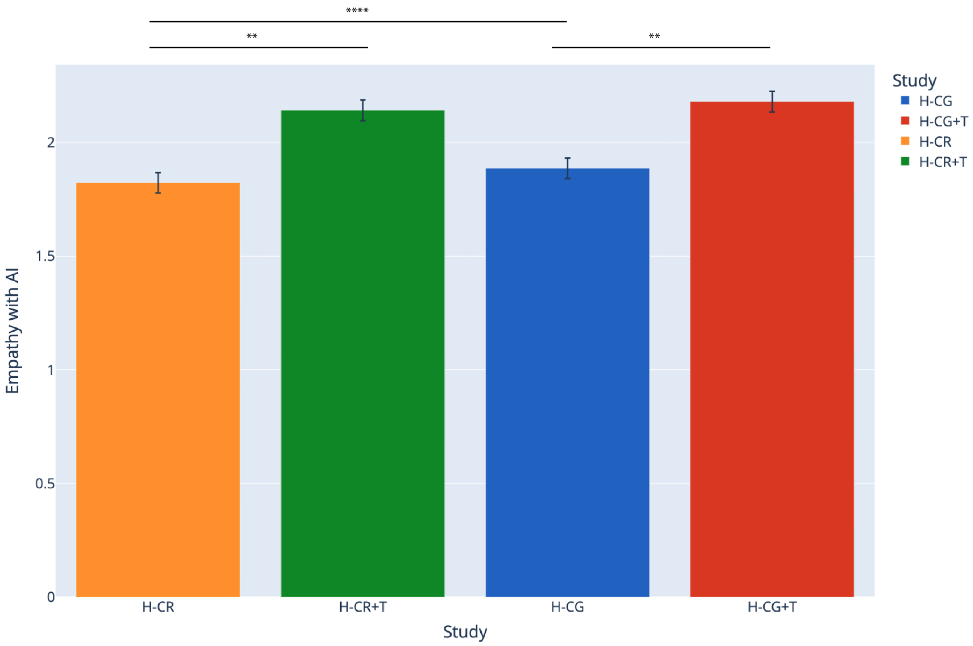


Figure 4: Self-reported willingness to empathize with AI written stories across all four studies

ference between empathy towards *human retrieved* and *ChatGPT generated* stories when the author is made transparent (**H-CG+T**).

**Generated Stories Elicit More Empathy Than Retrieved Stories.** Next, we cross-compared total empathy towards AI-written stories in **H-CG** (mean = 4.3, s.d. = 0.65) and **H-CR** (mean = 3.83, s.d. = 0.73), allowing us to explore differences in ChatGPT responding directly to a user's personal story context as compared to retrieving a relevant AI-generated story. From Figure 3, we see that empathy statistically significantly decreases in H-CR, when stories are retrieved instead of generated directly from the user's written story (t(597)=8.20, $p < 1e-14$, Cohen's d = 0.67).

**Disclosure of Story Author Reduces Empathy in ChatGPT Generated Stories.** We cross-compared total empathy towards AI-written stories in **H-CR** and **H-CR+T** (mean = 3.62, s.d. = 0.86), allowing us to assess how transparency about a story being written by ChatGPT shifts empathy. We find that empathy towards the AI-written stories statistically significantly decreases when users are told before reading that the story is written by ChatGPT (t(494) = 3.02, $p < 1e-4$, Cohen's d = 0.27), as shown in Figure 3.

Finally, we cross-compared total empathy towards AI-written stories in **H-CG** and **H-CG+T** (mean = 4.02, s.d. = 0.72), and see that empathy in **H-CG+T** statistically significant decreases (t(487) = 4.37, $p < 1e-6$, Cohen's d = 0.40). This confirms the aforementioned result that telling participants a story is written by an AI will decrease empathy (Figure 3).

## 4.2 Effects on Willingness to Empathize with AI

**People are More Willing to Empathize with AI-Written Stories if Author is Transparent.** In addition to raw, self-reported empathy towards the narrator of each story, we also ask participants to rate how much they believe their empathy would shift if the stories they read were all written by AI, where scores are from likert 1 (empathize a lot less) to 4, (empathize a lot more). As shown in Figure 4, we find that across all four studies, participants would, on average, empathize less (scores are generally at or below 2) with AI-written stories using our survey measurements (**H-CG**: mean = 1.88, s.d. = 0.91; **H-CR**: mean = 1.82, s.d. = 0.90; **H-CR+T**: mean = 2.14, s.d. = 0.89; **H-CG+T**: mean = 2.18, s.d. = 0.87,). However, interestingly, we see that willingness to empathize with AI-written stories statistically significantly increases when we are transparent about the story being written by ChatGPT (ie. participants read a story knowing it was generated by ChatGPT). These results are shown in cross comparing **H-CR** and **H-CR+T** for retrieved ChatGPT stories ( t(494)=-5.49, $p < 1e-7$, Cohen's d = 0.36 ) as well as cross-comparing **H-CG** and **H-CG+T** for directly generated ChatGPT stories (t(494)=-4.99, $p < 1e-6$, Cohen's d = 0.33 ).

## 4.3 Understanding Mechanisms Behind Empathy Towards Human vs AI Stories

Through qualitative coding of participant free responses, conducted by two independent coders, we reveal 9 unique themes around why participants did or did not empathize with the stories (Table 2). Participants explained their reasoning by commenting on the narrator's perspective, including empathizing

| Code | Definition | Total | H-CG | H-CR | H-CG+T | H-CR+T |
|---|---|---|---|---|---|---|
| Emotional | Empathize with the emotions that the narrator describes in the story | **30.38%** | **35.38%** | **31.99%** | **23.40%** | **27.29%** |
| Situational | Empathize with the situation or context that the narrator is in | 24.74% | 26.88% | 25.94% | 23.19% | 21.18% |
| Story Confusion | Mention of specific details in the story that aren't clear, including details or logic that doesn't add up | 11.41% | 8.91% | 12.97% | 11.49% | 12.88% |
| Not Relatable | Explicit mention of not empathizing because the story was not relatable or they did not agree with the narrator | 11.20% | 9.33% | 14.41% | 10.43% | 10.04% |
| Word Choice | Mention of the writing style, phrasing, or grammar, typically to reduce feelings of empathy | 7.39% | 6.55% | 5.19% | 9.57% | 9.83% |
| Authenticity | Explicit mention of the story being "real" or "fake", any mention of believability or originality | 6.67% | 5.15% | 4.03% | 10.85% | 8.73% |
| Mentions AI | Explicitly mentions AI or automation | 4.15% | 3.06% | 1.15% | 7.87% | 6.55% |
| Personality | Mention of personal ability to empathize | 0.47% | 0.56% | 0.29% | 0.43% | 0.66% |
| Other | Does not fit into any category, restates the question or generic | 3.59% | 4.18% | 4.03% | 2.77% | 2.84% |

Table 2: Themes resulting from a qualitative analysis across all four studies. Percentages are shown as the number of times a code was mentioned out of the total number of participants *within* each study. We **bold** the top code in each column and underline the top percent in each row.

with the **situation** in the story or the **emotions** the narrator describes. Some participants did express that the story was **not relatable** enough for them to empathize with. Two themes appeared around the way that the story was written: some expressed **story confusion** due to some of the logic of the story not being clear, and there was also a common theme around the **word choice** of the narrator, such as the writing style or phrasing. There were some participants who **mentioned AI** explicitly, and others who talked about the **authenticity** of the story, or whether it was real or fake. Some participants spoke about their **personality** being a factor in whether or not they were able to empathize. The **other** category was used if a response did not fit into an existing category.

We assigned a theme (or themes) to each response and a percentage was calculated in order to account for the number of participants in each study. As a whole, the **emotional** (30.38%) and **situational** (24.74%) codes showed up most frequently across all conditions. One notable difference is that participants in **H-CG** (35.38%) and **H-CR** (31.99%) had a higher percentage of **emotional** codes than **H-CR+T** (23.40%) and **H-CG+T** (27.29%). **H-CR+T** and **H-CG+T** had a higher percentage of **word choice**, **authenticity**, and **mention AI** codes than **H-CG** and **H-CR**.

We broke themes down into individual studies and conditions. Conditions **H-CR** and **H-CR+T** were compared, as they compared the same types of stories (human-retrieved vs. AI-retrieved), with **H-CR+T** explicitly telling participants when the stories were AI generated. Interestinglyg, codes for **emotional** were less common in the **H-CR+T** condition. **H-CG** and **H-CG+T** were compared (human-retrieved vs. AI-generated), and showed a similar decrease in **emotional** codes.

## 5 Discussion

From our work, we show that it is important to be intentional in how one presents outputs from generative AI systems.

**Generated vs Retrieved Stories.** Firstly, through cross-comparisons between ChatGPT-written retrieved stories (**H-CR**) and ChatGPT-generated stories (**H-CG**), we find that empathy is higher for ChatGPT-generated stories rather than ChatGPT-retrieved stories. Interestingly, we find that empathy is higher towards ChatGPT-generated stories than human-written retrieved stories. Thus, we did not validate that humans would empathize more with human-written stories in all conditions **[H1]**. These results on generated vs. retrieved stories highlight the importance of context awareness. Generated stories directly respond to the user's story, and previous literature shows that a direct response to one's story increases empathy [Rashkin *et al.*, 2019]. Output that is generated from conditioning on the stories can take much more from the input story, thus probably reaching a higher level of similarity, beyond what our retrieval algorithm is based on [Krishna *et al.*, 2023].

**Transparent vs. Opaque Story Author.** In studies **H-CR** and **H-CR+T** we find that people significantly empathize less with retrieved AI-written stories than human-written stories, which is in line with and supports previous research findings [Ishowo-Oloko *et al.*, 2019; Straten *et al.*, 2020]. We find that empathy decreases most between human-written and AI-retrieved stories in **H-CR+T** when we are transparent about the author of the story. This indicates that knowing when a story is written by an AI alters our empathy towards that story and ability to relate to the narrator, possibly because of the fact that an AI is conveying experiences that are not its "own."

Interestingly, participants' willingness to empathize with AI systems significantly increases across both retrieval and generation conditions when the author of the story is made transparent (validating **[H2]**). Prior works indicate that transparency about an AI's lack of human qualities can reduce perceived similarity [Straten *et al.*, 2020], but that transparency can increase trust towards AI systems [Liu, 2021]. Our results may indicate that disclosing a story's author could increase willingness to empathize through trust, or through demonstration that AI stories contain relatable qualities.

In the **H-CR+T** condition, participants' reasoning for not empathizing with AI-written stories was more centered around themes relating to how the story was written, including "story confusion" and "word choice", similar to research that showed "linguistic style" was a reported indicator for AI generated text [Jones and Bergen, 2023]. For example, one participant stated, "*The story and feelings described feel really fake and over the top. It does not feel genuine and has clearly been written by a robot.*"

Others mention not being able to empathize with the story because the story did not actually happen, but they are still capable of engaging with it as a made-up story. For example, one participant shared, "*Because I know it's written by AI then I can't think that it is genuine. However as a work of fiction I can immerse myself in it and connect with the characters portrayed.*" This sentiment opens up the potential for AI-written stories to be contextualized for the user in a way that doesn't feel like they are being deceived by a fake story.

We see no difference in empathy between retrieved human stories and ChatGPT stories generated in direct response to the user (**H-CG+T**), indicating that responding directly to a user's story might overshadow the underlying empathic benefits of human-written stories. In this condition, more participants mentioned the "authenticity" of the story or mentioned AI explicitly as a factor against empathizing with the story they read. Participants tended to focus more on the author of the story instead of the content of the story in their open-ended responses. One participant shared, "*The story felt similar to the content of my story, which made me feel like I could empathize with it. But knowing the story was written by an AI makes me feel less connected to the story because I know it's not real.*"

## 6 Ethical Considerations

Our experimental protocol was reviewed by our institution's ethics review boards, and we ensured that stories shown to participants were not toxic or harmful. From our studies, we show that retrieval of human-written stories can encourage human-human empathy rather than empathy towards AI systems. Large, pre-trained generative models do not truly experience the situations present in stories. Language models such as ChatGPT, represent a population sourced from large quantities of human data, but still fall short of human-written stories in their empathic quality [Giorgi *et al.*, 2023; Schaaff *et al.*, 2023b; Montemayor *et al.*, 2022; Pelau *et al.*, 2021]. This appropriation of human experiences could be subverted by using AI to instead, retrieve more empathically similar texts between human authors [Shen *et al.*, 2023] or to mediate human-human communications [Hohenstein and Jung, 2020].

It is clear that people bring their experiences and their identities to AI interactions [Alonso, 2023]. As such, it is also important to consider how our definition of empathy, which focuses on similarity in experience, might be limited in encompassing the rich diversity of experiences humans are able to empathize with [Lahnala *et al.*, 2022]. Future work can explore whether the goal of AI systems should be to make people more empathetic to any target (human vs. AI), more empathetic to human targets, or more empathetic to *diverse* human targets.

Finally, we show the importance of framing in interactions with stories, as a one-sentence disclosure of the author significantly shifted empathy. The field of AI advocates for transparency as an ethical design tenet [Yampolskiy, 2015]. The more transparent a system is, the more agency one has in the way they use it. In our study, we find that disclosure decreases empathy. This finding might be in tension with systems that rely on empathy for efficacy, such as in persuasive technologies that use bond with the robot to improve outcomes [Rahmanti *et al.*, 2022; Lv *et al.*, 2022]. However, the empathy and transparency trade-off might not be mutually exclusive, as transparency can breed trust, which also influences interaction.

## 7 Conclusion

In this work, we conducted four crowdsourced studies to assess how empathy differs across human-written vs AI-written stories, varying how stories are selected (generation vs retrieval) and author disclosure (transparency that story was written by an AI author vs. no transparency). While we utilize current state-of-the-art empathetic retrieval and generation in this work, our findings provide more generalized future insights around human behavior when interacting with AI. Crucially, we find that transparency of the author plays an important role in empathy towards an AI story as well as people's willingness to empathize towards machines. Our work motivates future directions regarding the social, psychological, and ethical implications of nuanced AI system design considerations that can drastically affect the ways in which humans extend empathy to artificial agents.

# References

[Abercrombie *et al.*, 2023] Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. Mirages: On Anthropomorphism in Dialogue Systems, October 2023. arXiv:2305.09800 [cs].

[Adam *et al.*, 2021] Martin Adam, Michael Wessel, and Alexander Benlian. Ai-based chatbots in customer service and their effects on user compliance. *Electronic Markets*, 31(2):427–445, 2021.

[Alonso, 2023] Marcos Alonso. Can Robots have Personal Identity? *International Journal of Social Robotics*, January 2023.

[Atkinson, 1998] Robert Atkinson. The Life Story Interview. page 21, 1998.

[Brandtzaeg *et al.*, 2022] Petter Brandtzaeg, Marita Skjuve, and Asbjørn Følstad. My AI Friend: How Users of a Social Chatbot Understand Their Human–AI Friendship. *Human Communication Research*, 48, April 2022.

[Chung and Kang, 2023] Liz L. Chung and Jeannie Kang. \I'm Hurt Too\: The Effect of a Chatbot\s Reciprocal Self-Disclosures on Users' Painful Experiences. *Archives of Design Research*, 36(4):67–84, November 2023.

[Croes and Antheunis, 2020] Emmelyn Croes and Marjolijn Antheunis. Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot. *Journal of Social and Personal Relationships*, 38:026540752095946, September 2020.

[Cuff *et al.*, 2016] Benjamin M.P. Cuff, Sarah J. Brown, Laura Taylor, and Douglas J. Howat. Empathy: A Review of the Concept. *Emotion Review*, 8(2):144–153, April 2016. Publisher: SAGE Publications.

[Decety and Jackson, 2004] Jean Decety and Philip L. Jackson. The Functional Architecture of Human Empathy. *Behavioral and Cognitive Neuroscience Reviews*, 3(2):71–100, June 2004.

[Djikic *et al.*, 2013] Maja Djikic, Keith Oatley, and Mihnea C. Moldoveanu. Reading other minds: Effects of literature on empathy. *Scientific Study of Literature*, 3(1):28–47, January 2013. Publisher: John Benjamins.

[Giorgi *et al.*, 2023] Salvatore Giorgi, David M Markowitz, Nikita Soni, Vasudha Varadarajan, Siddharth Mangalik, and H Andrew Schwartz. "I Slept Like a Baby": Using Human Traits To Characterize Deceptive ChatGPT and Human Text. 2023.

[Grootendorst, 2020] Maarten Grootendorst. Keybert: Minimal keyword extraction with bert., 2020.

[Hohenstein and Jung, 2020] Jess Hohenstein and Malte Jung. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior*, 106:106190, May 2020.

[Ishowo-Oloko *et al.*, 2019] Fatimah Ishowo-Oloko, Jean-François Bonnefon, Zakariyah Soroye, Jacob Crandall, Iyad Rahwan, and Talal Rahwan. Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence*, 1(11):517–521, November 2019. Number: 11 Publisher: Nature Publishing Group.

[Jones and Bergen, 2023] Cameron Jones and Benjamin Bergen. Does gpt-4 pass the turing test? *arXiv preprint arXiv:2310.20216*, 2023.

[Kalla and Broockman, 2021] Joshua L. Kalla and David E. Broockman. Which Narrative Strategies Durably Reduce Prejudice? Evidence from Field and Survey Experiments Supporting the Efficacy of Perspective-Getting. *American Journal of Political Science*, n/a(n/a), 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12657.

[Kim and Hur, 2023] Woo Bin Kim and Hee Jin Hur. What makes people feel empathy for ai chatbots? assessing the role of competence and warmth. *International Journal of Human–Computer Interaction*, pages 1–14, 2023.

[Krishna *et al.*, 2023] Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense, October 2023. arXiv:2303.13408 [cs].

[Lahnala *et al.*, 2022] Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. A Critical Reflection and Forward Perspective on Empathy and Natural Language Processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2139–2158, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[Lee *et al.*, 2019] Yeonjoo Lee, Miyeon Ha, Sujeong Kwon, Yealin Shim, and Jinwoo Kim. Egoistic and altruistic motivation: How to induce users' willingness to help for imperfect AI. *Computers in Human Behavior*, 101:180–196, December 2019.

[Liu, 2021] Bingjie Liu. In AI We Trust? Effects of Agency Locus and Transparency on Uncertainty Reduction in Human–AI Interaction. *Journal of Computer-Mediated Communication*, 26(6):384–402, November 2021.

[Luo *et al.*, 2019] Xueming Luo, Siliang Tong, Zheng Fang, and Zhe Qu. Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science*, 38(6):937–947, 2019.

[Lv *et al.*, 2022] Xingyang Lv, Yufan Yang, Dazhi Qin, Xingping Cao, and Hong Xu. Artificial intelligence service recovery: The role of empathic response in hospitality customers' continuous usage intention. *Computers in Human Behavior*, 126:106993, January 2022.

[Masters *et al.*, 2021] Peta Masters, Wally Smith, Liz Sonenberg, and Michael Kirley. Characterising deception in ai: A survey. In *Deceptive AI: First International Workshop, DeceptECAI 2020*, pages 3–16. Springer, 2021.

[Montemayor *et al.*, 2022] Carlos Montemayor, Jodi Halpern, and Abrol Fairweather. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. *AI & SOCIETY*, 37(4):1353–1359, December 2022.

[Natale, 2021] Simone Natale. *Deceitful media: Artificial intelligence and social life after the Turing test*. Oxford University Press, USA, 2021.

[Oatley, 2016] Keith Oatley. Fiction: Simulation of Social Worlds. *Trends in Cognitive Sciences*, 20(8):618–628, August 2016.

[Paiva *et al.*, 2004] Ana Paiva, Joao Dias, Daniel Sobral, Ruth Aylett, Polly Sobreperez, Sarah Woods, Carsten Zoll, and Lynne Hall. Caring for agents and agents that care: Building empathic relations with synthetic agents. In *Autonomous Agents and Multiagent Systems, International Joint Conference on*, volume 2, pages 194–201. IEEE Computer Society, 2004.

[Paiva, 2011] Ana Paiva. Empathy in social agents. *International Journal of Virtual Reality*, 10(1):1–4, 2011.

[Patton, 2005] Michael Quinn Patton. Qualitative research. *Encyclopedia of statistics in behavioral science*, 2005.

[Pelau *et al.*, 2021] Corina Pelau, Dan-Cristian Dabija, and Irina Ene. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior*, 122:106855, September 2021.

[Rahmanti *et al.*, 2022] Annisa Ristya Rahmanti, Hsuan-Chia Yang, Bagas Suryo Bintoro, Aldilas Achmad Nursetyo, Muhammad Solihuddin Muhtar, Shabbir Syed-Abdul, and Yu-Chuan Jack Li. SlimMe, a Chatbot With Artificial Empathy for Personal Weight Management: System Design and Finding. *Frontiers in Nutrition*, 9, 2022.

[Rashkin *et al.*, 2019] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y.-Lan Boureau. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset. Technical Report arXiv:1811.00207, arXiv, August 2019. arXiv:1811.00207 [cs] type: article.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, August 2019. Number: arXiv:1908.10084 arXiv:1908.10084 [cs].

[Sadasivan *et al.*, 2023] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected?, 2023.

[Schaaff *et al.*, 2023a] Kristina Schaaff, Caroline Reinig, and Tim Schlippe. Exploring ChatGPT's Empathic Abilities, September 2023. arXiv:2308.03527 [cs].

[Schaaff *et al.*, 2023b] Kristina Schaaff, Caroline Reinig, and Tim Schlippe. Exploring chatgpt's empathic abilities. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2023.

[Shen *et al.*, 2023] Jocelyn Shen, Maarten Sap, Pedro Colon-Hernandez, Hae Park, and Cynthia Breazeal. Modeling empathic similarity in personal narratives. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6237–6252, Singapore, December 2023. Association for Computational Linguistics.

[Shen, 2010] Lijiang Shen. On a Scale of State Empathy During Message Processing. *Western Journal of Communication*, 74(5):504–524, October 2010. Publisher: Routledge _eprint: https://doi.org/10.1080/10570314.2010.512278.

[Spitale *et al.*, 2022] Micol Spitale, Sarah Okamoto, Mahima Gupta, Hao Xi, and Maja J Matarić. Socially Assistive Robots as Storytellers That Elicit Empathy. *ACM Transactions on Human-Robot Interaction*, page 3538409, May 2022.

[Straten *et al.*, 2020] Caroline L. Van Straten, Jochen Peter, Rinaldo Kühne, and Alex Barco. Transparency about a Robot's Lack of Human Psychological Capacities: Effects on Child-Robot Perception and Relationship Formation. *ACM Transactions on Human-Robot Interaction*, 9(2):1–22, June 2020.

[Tsumura and Yamada, 2023] Takahiro Tsumura and Seiji Yamada. Influence of agent's self-disclosure on human empathy. *Plos one*, 18(5):e0283955, 2023.

[Yampolskiy, 2015] Roman V. Yampolskiy. Taxonomy of Pathways to Dangerous AI, November 2015. arXiv:1511.03246 [cs].

[Yoon and Lee, 2021] Namhee Yoon and Ha-Kyung Lee. Ai recommendation service acceptance: Assessing the effects of perceived empathy and need for cognition. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(5):1912–1928, 2021.

[Zaki, 2019] Jamil Zaki. *The war for kindness: Building empathy in a fractured world*. Crown, 2019.