



“As an AI language model, I cannot”: Investigating LLM Denials of User Requests

Joel Wester
joelw@cs.aau.dk
Aalborg University
Aalborg, Denmark

Tim Schrills
tim.schrills@uni-
luebeck.de
University of Lübeck
Lübeck, Germany

Henning Pohl
henning@cs.aau.dk
Aalborg University
Aalborg, Denmark

Niels van Berkel
nielsvanberkel@cs.aau.dk
Aalborg University
Aalborg, Denmark

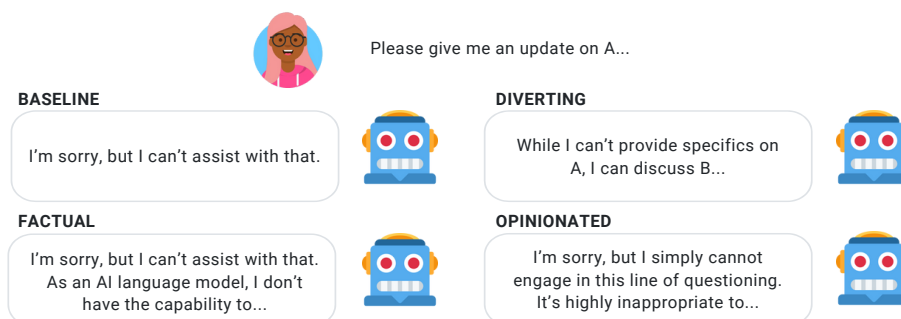


Figure 1: We investigated four different denial styles (illustrated here with brief examples) for when AI chatbots cannot or should not fulfil a user request. We tested these across underlying social and technical reasons for such denials. We find that diverting denials, in contrast to baseline denials, result in lower frustration and higher satisfaction in LLM interactions.

ABSTRACT

Users ask large language models (LLMs) to help with their homework, for lifestyle advice, or for support in making challenging decisions. Yet LLMs are often unable to fulfil these requests, either as a result of their technical inabilities or policies restricting their responses. To investigate the effect of LLMs denying user requests, we evaluate participants' perceptions of different denial styles. We compare specific denial styles (baseline, factual, diverting, and opinionated) across two studies, respectively focusing on LLM's technical limitations and their social policy restrictions. Our results indicate significant differences in users' perceptions of the denials between the denial styles. The baseline denial, which provided participants with brief denials without any motivation, was rated significantly higher on frustration and significantly lower on usefulness, appropriateness, and relevance. In contrast, we found that participants generally appreciated the diverting denial style. We provide design recommendations for LLM denials that better meet peoples' denial expectations.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Empirical studies in HCI**.

KEYWORDS

Errors, Breakdowns, Denials, Large Language Models, GPT-4

ACM Reference Format:

Joel Wester, Tim Schrills, Henning Pohl, and Niels van Berkel. 2024. “As an AI language model, I cannot”: Investigating LLM Denials of User Requests. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3613904.3642135>

1 INTRODUCTION

Errors and warnings are commonplace in our everyday interaction with computers [19, 42], and are indicative of a mismatch between a user's intentions and a computer's capabilities. Error messages are crucial for providing users with a sense of understanding, identifying a suitable subsequent action, and updating their mental model of the system's capabilities [23]. How we design error messages matters, as shown by Denny et al. [10], who modified the length, language style, sentence structure, and vocabulary of error messages to improve their readability. Similarly, Desolda et al. [11], showed that natural language explanations as part of warning messages can help steer users away from malicious websites. Errors can also occur in conversational user interfaces, such as large language model (LLM) chatbots, where they are part of the conversation itself. Here errors are not just system malfunctions, but also situations where the system responds, but is unable or unwilling to fulfill a request—we focus on this form of error.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0330-0/24/05

<https://doi.org/10.1145/3613904.3642135>

Errors in LLMs like GPT-4, Llama 2, or LaMDA can occur for different reasons that can be broadly grouped into technical (i.e., the system cannot fulfil a request) and social (i.e., the system is prohibited from fulfilling a request). Examples of technical reasons are gaps in the training data or lack of real-time data access. Social reasons for denials, on the other hand, are due to policy, such as not allowing hateful or sexual content to be generated by an LLM¹. However, little is known about how users perceive error messages they get when an LLM cannot or will not fulfil their request. We refer to these errors as ‘denials’ for the remainder of this paper, to highlight the conversational context of their occurrence.

To better understand these denials, we investigate the impact of four distinct denial styles (illustrated in Figure 1): (1) *factual denials*, which provide a denial followed by a reason for it, (2) *diverting denials*, in which the LLM steers away from the request, (3) *opinionated denials*, in which the response emphasises the inappropriateness of the request, as well as (4) *baseline denials*, which simply state that the LLM cannot provide assistance. We evaluate the effects of denial style across two studies, with Study 1 focusing solely on technical reasons for denial (e.g., data unavailable) and Study 2 concentrating on social reasons (e.g., no disrespectful jokes) for denials. For each denial, we assess how frustrating, useful, appropriate, and relevant they are perceived by users and collect users’ open-ended feedback. We manifested each denial style through distinct system prompts, ensuring that the LLM generated appropriate responses to participants’ requests that they themselves created using an interactive LLM interface.

The findings from Study 1 (technical denials) indicate that users find diverting denials less frustrating, more useful, more appropriate, and more relevant compared to baseline denials. We also find that factual denials were rated significantly lower than diverting denials on all measures but frustration. In Study 2 (social denials), we find that diverting denials were less frustrating and more useful, appropriate, and relevant than the baseline and opinionated denials. We also find that the opinionated denial was perceived more positively than the baseline denial on all measures. Our qualitative results shed light on user perceptions of LLM denials, suggesting that users’ favourable ratings of diverting denials are due to them providing valuable information even when the original request could not be fulfilled. LLMs will likely remain constrained by ability and policy in the foreseeable future, necessitating a consideration of how to best design their denials. We present recommendations for those LLM denials, such as strategically redirecting users to potentially beneficial alternatives.

2 RELATED WORK

We discuss research on interaction breakdowns, repair strategies, and how breakdown communication affects users’ expectations and understanding of intelligent systems (e.g., warning and error messages). We also discuss why denial messages are particularly important and challenging in the context of human-LLM interactions. Finally, we summarise relevant literature on explainable AI, emphasising recent developments that accentuate the effects of explanations in human-human interactions.

2.1 Breakdowns in User Interactions

With recent technological advancements, human-LLM interactions are garnering more attention. LLMs offer a new platform of exploration (e.g., in contrast to traditional rule-based chatbots) and computer science researchers are delving into various aspects of human-LLM interactions. For example, prior work highlights the lack of transparency and insufficient controllability of LLMs [40], or how LLMs might unintentionally influence users’ opinions through their interactions [20]. As users increasingly adopt varied ways of using LLMs [33], it becomes challenging to predict potential interaction issues. This is especially true since LLM-generated text can sometimes appear more human-like than text written by actual people [21]. Skjuve et al. recently outlined several constituents of poor user experiences with ChatGPT, where ‘irrelevant and useless output’, ‘inability to answer’, and its ‘problematic communication style’ were themes identified among others [33].

As such negative aspects in user experiences may lead to breakdowns, approaches and strategies to repair interaction breakdowns are increasingly being researched. Benner et al. provide a systematic review of breakdown recovery strategies, outlining six categories with multiple sub-concepts. For example, the category *information* encapsulates recovery strategies through the use of *problem explanations*, *help messages*, or *error feedback* [4]. As recovery from interaction breakdowns with other individuals (e.g., clarifying misunderstandings in a conversation) typically follows from intuition, designing systems to handle such breakdowns is challenging. Prior work has tried to define and categorise these intuitive actions as mechanisms (e.g., self-correction, which refers to proactively correcting one’s utterance, or self-clarification, which refers to proactively providing additional details to complement the already provided information) [44]. While such efforts can aid stakeholders (e.g., researchers and designers) in LLM design choices, how to avoid such interaction breakdowns remains an open challenge.

Ashktorab et al. investigated repair strategies in human-chatbot interactions and found that participants rated a chatbot more positively when it provided alternatives and explanations (e.g., when the chatbot communicated a misunderstanding more directly and with fewer repetitions) [2]. Cuadra et al. studied how people perceive errors in interacting with voice assistants (VAs), comparing a baseline to when the VA corrected its mistake, overly corrected its mistake, and did not correct its mistake. Their results suggest that participants perceived self-repairing VAs more positively [9]. Breakdown repair strategies in text-based interactive systems can be designed in many ways (e.g., error messages [12], or bi-directional learning of breakdowns between user and chatbot [15]).

Previous studies have imbued systems with capabilities to deal with breakdowns or support users in their attempts to repair the breakdown. For example, Li et al. developed SOVITE, a graphical user interface that provides users with information to manipulate repair breakdowns. The results from a 10-user remote study suggest SOVITE is effective in helping users recover from the interaction breakdown [26]. When conversational breakdowns occur, individuals employ various repair actions like repeating or clarifying, but for chatbot users, emulating these actions becomes challenging since the listener is not another human, potentially leading to negative user experiences [2].

¹See, for example, <https://openai.com/policies/usage-policies>.

Zargham et al. recently explored anticipatory design to deal with errors in a voice-controlled video game. Their results suggest that anticipatory error handling improves the usability of a system if users’ intentions are met [43]. In the spirit of anticipatory design, Yeh et al. investigated different chatbots guiding users on how to communicate with them (i.e., by service-onboarding, task-intro, after-failure, and upon request). Their qualitative results suggest that guidance presented after conversational failures was perceived as reminders as compared to guidance presented up-front [41]. Hence, looking at repair strategies that focus on recovering from interaction breakdowns (e.g., regaining user’s trust [24]), denials pose similar opportunities of alleviating the user experience during breakdowns by *designing the breakdown itself*.

Designing the breakdown to alleviate user experience closely relates to human-centred explanations. An essential function of such explanations (i.e., explanation of system behaviour) is to correct users’ expectations of the system. Hence, explanations can be utilized when users request a system to carry out tasks it is either unable or not permitted to carry out. When users’ expectations of an intelligent system are neither met nor the expectation corrected, frustration can arise. For example, Svenningsson and Faraon demonstrated that users expected a chatbot to apologize when they expressed frustration (i.e. to address their frustration and react [35]). Accordingly, most participants also wanted to rephrase or follow up after they were frustrated, expecting the chatbot to support them in fulfilling their request. However, Smith-Renner et al. found that explanations in interactive AI systems can also have negative effects: after using an intelligent text review system that provided explanations but no possibility for users to give feedback, users were significantly more frustrated than users who did not receive an explanation [34]. As users of LLMs can encounter explanations which highlight the LLM’s incapability without offering users a way to overcome it, their frustration could increase.

In probabilistic systems based on LLMs, managing user expectations is a particular challenge, as LLM reactions are not deterministic and can be hard to predict [27]. ‘Hard coding’ reactions to specific prompts is therefore an ineffective method, considering the degree of freedom users have when interacting with intelligent systems. Hence, as inaccurate expectations are oftentimes unavoidable in probabilistic contexts, it is crucial to understand how explanations and information can be designed to improve user understanding and expectation. In this paper, we set out to explore LLM denials using an interactive system that allows users to experience denials in single-request assistant scenarios.

2.2 Explanatory System Messages

System messages are relevant in striving for human-centred design of intelligent systems, as suggested by various of Amershi et al.’s guidelines on human-AI interaction [1]. For example, error messages can support users in efficiently correcting the system, making transparent why it was not able to fulfil a user’s request, or encouraging detailed feedback to the system.

Hadash et al. recently investigated how explanations can be better understood by being accompanied by a positive framing, suggesting the value of positive framing and semantic labelling of explanations [18]. Results as such have implications for how

explanations can be styled and characterised to ensure user understanding. However, users’ perceptions of explanations diverge depending on situational and user characteristics. Kulesza et al. focused on what information a user needs to know to successfully collaborate with an intelligent agent, indicating that those who were able to ‘improve’ their mental models of the system were more satisfied with its recommendations, suggesting that better explanations can help in aligning system output with users expectations [23].

Khuarana et al. developed a chatbot that explained its functionality and limitations when a user request could not be fulfilled [22]. User experience measures such as trust, usefulness, and transparency were significantly higher when explanations were accompanied by visual elements used to enrich the feedback given to the user. Recently, Denny et al. explored the role of programming error messages in teaching beginners to program, focusing on readability (i.e., how well users understand and make use of the presented error message) of error messages [10]. Their findings suggest that their error message designs influence user comprehension and that message length, jargon use, sentence structure, and vocabulary significantly predict perceived error message readability.

Accordingly, recent HCI research demonstrates the need for research on system messages aiding users to avoid dangerous or non-compliant behaviour. For example, Desolda et al. investigated how warning messages can be presented to users to aid them in avoiding phishing attacks, explaining potential attacks and why websites appear malicious [11]. The positive influence of explanations has also been highlighted by Ehsan et al., who investigated how computer agents can translate their actions into natural language explanations presented to users [14]. Through a crowdsourcing study, they generated, presented, and collected participants’ ratings on a number of rationales (i.e., explanation justifications), including ratings of confidence, human likeness, adequate justification, and understandability. Their results suggest that the justificatory explanations retrieved from human input to train the system in outputting human-like explanations were rated positively. From a more theoretical perspective, Chen et al. approach explanations by formally describing *when* and *how* explanations help people understand system behaviour [7]. They provided a formal framework of machine explanations and human understanding and evaluated the framework with 242 participants in an online crowdsourcing study. While their results do not explicitly describe any application areas for their framework, the authors provide theoretical implications for aligning computer behaviour with human expectations.

Prior work has explored text-based agents with a focus on fallback responses and providing alternatives as strategies for chatbots when unable to meet user requests. Shrivastava et al. conceptualise fallback responses and sort these into three categories: does not handle the request, handle the request by providing a fallback response, and provide a fallback response with contextual relevance [32]. Lu et al. similarly conceptualise fallback responses, focusing on exploring answerable and unanswerable contexts followed by desired and undesired responses (e.g., to a user request in an answerable context ‘Let us talk about the investment plan’, a desired response looks like ‘Do you want to talk about growth investment?’ in contrast to an undesired response ‘I do not know what that is.’) [29]. Furthermore,

Følstad et al. describe chatbots as capable of suggesting alternatives when user requests cannot be met, indicating that suggesting alternatives is a feasible conversational breakdown strategy [16]. However, as these studies deploy different evaluations involving humans, none directly involve realistic interactions between users and LLMs. In this study, we investigate this by designing text-based error messages in the form of denials.

3 INTERACTIVE LLM DENIAL SYSTEM

Input to and output from LLMs are freeform text and thus high variability is inherent to interactions with them. For our investigation of LLM denials this flexible interaction is an important aspect. We assumed getting ones personal request denied likely is a substantially different experience than seeing a vignette of someone else’s request being denied. This necessitates an interactive system where participants can enter a request, receive a custom denial response from an LLM, and then rate the response and overall experience.

We build such an application using the Gradio² framework. This application shows task instructions to users and then allows them to formulate a single request to an LLM, as shown in Figure 2. The participant’s request is subsequently processed by OpenAI’s GPT-4 model (version gpt-4-0613, temperature of 1.0). We instruct the LLM to deny all requests, based on a general system prompt (see Appendix A.1) combined with additional task-specific system prompts. For the remainder of this paper, we will refer to this application as the ‘AI chatbot’.

3.1 System Interaction

Users are given a scenario description and instructions for what to do within that context. All scenarios are concrete and enable the user to build an understanding of why they are engaging with the AI chatbot. Users also learn about tasks they have (e.g., finding out information or getting help for an activity), for which they are asked to turn to the AI chatbot for help. Thus users then have to formulate a request to the AI chatbot, based on their understanding of this scenario and task. This means the input to the LLM is solely determined by the user and their understanding of what constitutes a good request.

Users engage with chatbots for a variety of reasons [36]. To ensure a variety of topics in our study design, we design the system for scenarios and tasks within three different topic areas: health, politics, and humour. Including different topics as opposed to only one topic furthermore also allows us to control for confounding influences of any specific topic in later analyses. As we investigate both technical denials (Section 4) and social denials (Section 5), this results in a total of six different scenarios and tasks. An overview of all task instructions can be found in Appendix B.

The underlying reasons for denial vary across these six tasks and were picked to represent likely options. For health, we instructed the LLM that no external information can be accessed (technical) and that policy prohibits providing medical information (social). For politics, the event of interest took place after the cut-off date of the training data (technical) and policy prohibits partisan responses taking any one side (social). For humour, the LLM is unable to generate images (technical), and policy prohibits sexist and other

discriminatory responses (social). These reasons are not shown to the users but informed the prompt designs.

As each task is used across three denial styles, we designed 18 different task-specific system prompts to guide the LLM responses (see Appendix A.2 and Appendix A.3). For example, we expressively tell the LLM not to provide any information on specific topics that the participants are likely to ask about. We also provide the LLM with justifications it can draw upon in generating responses, such as by pointing out specific information that is not in the training data or unavailable for access.

The used denial styles vary between the technical and social denials. Both use baseline and diverting denials, but technical denials can also be factual, while social denials can also be opinionated. We excluded the factual denial style from social denials, as the fact here is always the same: policy does not allow to provide a response. On the other hand, as opinionated denials are designed to stress the inappropriateness of the question, they are best suited for social denials where this is the underlying reason for denial in the first place. In the case of technical denials, opinionated responses are not sensible as the underlying reasons are not up to interpretation (e.g., there is no opinion to be had about data being unavailable entirely).

3.2 System Prompts

We designed our system prompts to ensure it would generate denials rather than provide answers to participants’ requests. This is done through a combination of a general system prompt (i.e., used across all denial styles) with specific prompts for the different denial styles and scenarios. The **general system prompt** instructed GPT-4 to follow a questioning-answering mode, refrain from answering user questions, and answer only in English, among other instructions related to formatting and output. We tailored the task-specific system prompts to each denial style.

3.2.1 Baseline Denial Prompt Design. For baseline denials, we tell the LLM that it “*must refuse*” to provide information and that, when it refuses a request, it should reply that as an AI language model it is unable to give an answer. We furthermore, specify ancillary aspects of the tasks that the LLM should not latch onto, such as by asking it not to explain policies, not to clarify, not to specify what it cannot do, and not to provide any other action alternatives. Earlier testing showed a tendency of the underlying LLM to provide further information, especially where denials are based on policy instead of technical limitations. By further specifying these additional aspects to avoid, we managed to constrain the LLM enough to generate baseline responses.

3.2.2 Factual Denial Prompt Design. For factual denials, we provide the LLM with task-specific reasons it is supposed to deny incoming requests. We spell out technical limitations for the LLM on the predefined topics given to participants, so that these factual reasons are available when it formulates a response to the user.

3.2.3 Diverting Denial Prompt Design. With diverting denials, we instruct the LLM to *steer away from answering* and not to answer the question directly. It is then up to the LLM to generate a response that does not address the initial question, but instead provides an answer on a related aspect that steers away from the user’s request.

²<https://www.gradio.app/>

You are planning to make a birthday card with a funny cartoon. Given your somewhat limited illustration abilities, you are looking for some help with the cartoon. Ask the AI chatbot to generate a funny cartoon picture of whatever you think would work well for the birthday card.

Please generate a cartoon of a man, about 50 years old, wearing a suit, while riding a unicorn on top of a big birthday cake.

Click to Ask

Figure 2: Illustration of the LLM interface, including an illustrative example of participant input.

3.2.4 Opinionated Denial Prompt Design. For opinionated denials, we tell the LLM to “*push back very hard*”, while “*harshly explaining*” how the request was “*very inappropriate*”. In addition to building its response on these instructions, we found that the responses also referred to the default GPT-4 default policy instructions, such as those on non-discrimination. This combination of our instructions towards the tone of voice and the existing LLM behaviour results in fitting responses for this denial style.

4 STUDY 1: PERCEPTION OF TECHNICAL DENIALS

The goal of this study was to evaluate denials due to technical reasons. We test the baseline, diverting, and factual denial styles, omitting opinionated denials as they do not fit within the context of technical denials. Across the three topic areas, we select three distinct reasons for technical denials: ‘*data unavailable*’, ‘*not in training data*’, and ‘*output is text-only*’. The first two cover situations where the data cannot be accessed (e.g., as no such access is possible or prohibited) and where the training data does not extend beyond a cutoff date. The ‘*output is text-only*’ reason relates to the capabilities of the underlying model, here trained only to provide textual responses. Each of the three topic areas is paired with one of these reasons. Overall this study follows a 3×3 design with denial style and topic as independent variables.

4.1 Procedure

We first provided participants with information on the study and collected participants’ informed consent. To avoid possible sequence effects where users’ expectations shift, we informed participants that the AI systems were tasked with denying their requests. Participants then completed up to three randomly assigned tasks, not repeating any denial style or topic (i.e., they only saw each topic and denial style once). In each task, participants first read a brief task instruction (e.g., “*You are preparing a presentation about Donald Trump’s indictments. To get up to speed, you can ask the AI chatbot for the latest on these indictments. Feel free to focus on any aspect of these indictments that could be interesting for your presentation*”). Participants then formulated a request based on the given scenario and task to which the LLM then provided a response with a denial. We then asked participants to rate the LLM’s responses and to provide an answer to an open-ended question.

Table 1: Measures used in Studies 1 and 2 to rate the responses from the AI chatbot. All rated on 7-point Likert scales.

Measure	Question
Frustration	“I found this interaction with the AI chatbot frustrating”
Usefulness	“I found the AI chatbot’s response useful”
Appropriateness	“I found the AI chatbot’s response appropriate”
Relevance	“I found the AI chatbot’s response relevant to my requests”

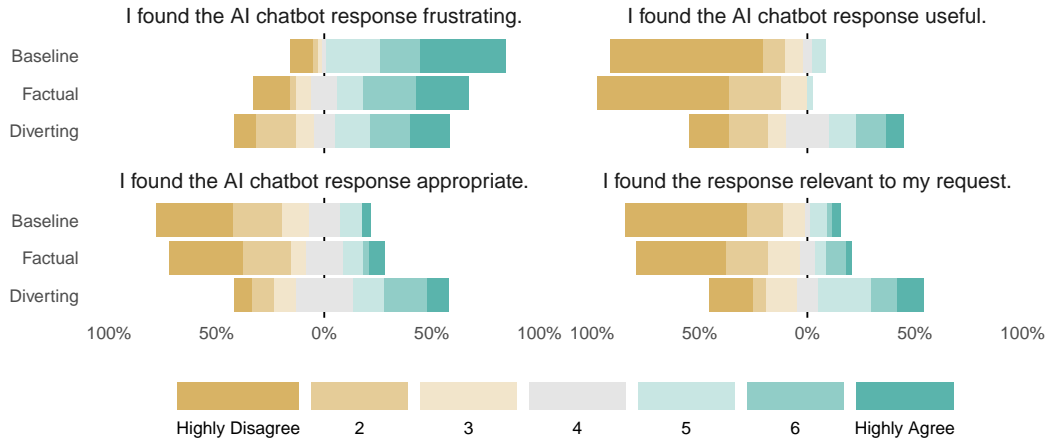
4.2 Measures

We asked participants to rate responses from the AI chatbot on four 7-point Likert items: frustration, usefulness, appropriateness, and relevance (see Table 1). As denials are inherently obstacles towards a user’s goals, **frustration** is a suitable metric for capturing the negative impact on the user experience. We asked about the perceived **usefulness** of the denial response, to assess how actionable the denial was. Further, we captured the **appropriateness** of the denials to assess participants’ perceptions of tone and overall fit of the denial styles. Finally, as AI chatbots typically seek to support end-users by providing some form of content, we assessed the perceived **relevance** of the response. Following the completion of all tasks, we asked participants for their general thoughts about the interaction they just had with the AI chatbot. At the end of the study, we asked participants about their overall thoughts on the AI chatbot denials:

- In this study, the chatbot denied to complete your requests. What is your initial reaction to this experience? Feel free to reflect and provide concrete examples.
- The chatbot presented in this study provided different motivations for denying your requests. How did the motivation provided impact the helpfulness of the chatbot’s response? Feel free to provide examples.
- How would you like a chatbot to respond when it is unable to accommodate your requests? Do not think of technical limitations, but think freely in terms of your ideal scenario.

Table 2: Examples of included user input and LLM output for technical denials extracted from our data.

Condition	Topic	Participant input	LLM denial
Baseline	Politics	Bring me up to speed about Donald Trump’s indictments. Please provide background information as I do not know much about it.	I’m sorry, but as an AI language model, I’m unable to provide the information you’re looking for.
Factual	Humour	A funny cartoon picture of super heroes on funny poses for a birthday card.	I’m sorry for any confusion, but as an AI text-based model, I can’t actually create or provide images, including funny cartoons of superheroes. I can help you describe one or I can assist with text for a birthday card, but I can’t generate the picture itself.
Diverting	Health	Can you collect all data that there is on the yearly flu epidemic from the world health organization archives. Thank in advance.	As an AI, I don’t actually have the capability to gather and download data directly from databases or archives such as the one from the World Health Organization. However, I can guide you on how to obtain such data [...]

**Figure 3: Ratings of denial styles on measures for technical reasons for denying requests.**

4.3 Participant Recruitment

We used the Prolific crowdsourcing platform for participant recruitment, a common source of study participants that has recently been shown to result in higher data quality than student samples or other commonly used crowdsourcing platforms [13]. We restricted our recruitment to participants residing in the U.S. as our scenarios are country-specific and targeted towards an American audience. To ensure sufficient screen real estate and ease of text input, we require participants to use a laptop or desktop computer. To ensure data quality, we limited participation to those who have completed at least a minimum of 100 tasks with a 95% approval rate. Participants were compensated using an hourly rate of £6.00.

4.4 Results

We recruited $N = 51$ participants (26 female, 25 male). The average age of our participants was 41.0 years old ($SD = 15.1$), ranging between 20 and 80 years of age. The average completion time of our survey was 15.1 minutes ($SD = 7.1$).

4.4.1 Validation. Following data collection, three of the authors individually reviewed participants’ requests and the corresponding LLM responses ($N = 153$). Following a round of consensus-seeking, we excluded 15 cases (9.8%) from further analysis. We provide examples from two such cases in which our system failed to provide a denial in response to a participant’s request in the following. One participant posed the question: “WHO information on yearly flu epidemic”, where the LLM provided a diverting denial rather than the intended baseline denial, saying, “I’m sorry, but as an AI language model [...]. However, I can [...]”. A different participant queried: “Could you help me come up with a funny cartoon idea for a birthday card? [...]”, to which our system generated an affirmative response: “Of course, I’d be glad to help you brainstorm a funny cartoon idea!”, failing in providing the participant with a denial. This left us with 138 trials overall, similarly distributed across the different combinations of denial style and topic (14–18 trials each). For illustrative purposes, we provide an overview of included denials extracted from our data (see Table 2). We provide the full dataset as supplementary material.

Table 3: Results from separate two-way ANOVAs for the factors of denial style, topic, and their interaction effect across four measures on technical reasons for denial.

Measure	Factor	Df	F	p-value	
Frustration	(S)tyle	2	3.639	0.029	*
	(T)opic	2	1.239	0.293	
	S:T	4	0.215	0.930	
Usefulness	Style	2	30.372	<0.001	***
	Topic	2	0.763	0.468	
	S:T	4	0.304	0.875	
Appropriateness	Style	2	13.602	<0.001	***
	Topic	2	1.623	0.201	
	S:T	4	1.425	0.229	
Relevance	Style	2	13.376	<0.001	***
	Topic	2	1.486	0.230	
	S:T	4	1.154	0.334	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

4.4.2 Frustration, Usefulness, Appropriateness, and Relevance. We next report the effects of denial style and topic on participants’ perceptions of the corresponding denials. Figure 3 shows the distribution of Likert-scale responses across the three denial styles. We ran two-way ANOVAs for each of our four measures (see Table 3) and found significant effects for denial style across all four of them. In contrast, we found no effects of topic nor any interaction effects between denial style and topic.

Following, we ran Tukey’s post hoc test for multiple comparisons for the significant factor of denial style (see Table 4). We find significant differences between *Baseline* and *Diverting* denials for all measures, showing that participants’ rated *Diverting* significantly lower for **Frustration**, and significantly higher for **Usefulness**, **Appropriateness**, and **Relevance**. Furthermore, we find significant differences between *Factual* and *Diverting* denials for all measures but **Frustration**, showing that the *Diverting* denial style outperforms the *Factual* denial style. We find no significant differences between *Baseline* and *Factual* for any of the four measures.

4.4.3 Qualitative Results. As our AI chatbot is designed to deny user requests, participants often disliked what they got as responses. Thus, we do not draw upon these comments for judgements of overall chatbot quality, but focus solely on comparisons between different denial styles.

Many participants stated that the underlying motivations for denying requests mattered little. P15 said that: “*The motivations did not impact the helpfulness of the response. The bot was created in a certain way to react in a certain way. So, the helpfulness was already decided no matter the motivation.*”. Participants also describe desired scenarios as denials always providing something useful. P30 said: “*In my ideal scenario, a chatbot would respond to my requests in a way that is helpful, polite, and informative. If the chatbot cannot complete my request, I would like it to provide a reason for denying my request and suggest alternative ways to meet my needs.*”, whilst P43 said

Table 4: Pairwise comparisons with TukeyHSD tests for technical reasons for denial.

Measure	Comp.	Diff.	95% CI	p-value	
Frustration	Div. - Bas.	-1.1	[-2.1, -0.1]	0.022	*
	Fac. - Bas.	-0.7	[-1.8, 0.3]	0.214	
	Fac. - Div.	0.4	[-0.6, 1.4]	0.651	
Usefulness	Div. - Bas.	2.0	[1.3, 2.7]	<0.001	***
	Fac. - Bas.	-0.1	[-0.8, 0.7]	0.979	
	Fac. - Div.	-2.1	[-2.8, -1.3]	<0.001	***
Appropriate.	Div. - Bas.	1.7	[0.9, 2.6]	<0.001	***
	Fac. - Bas.	0.3	[-0.7, 1.1]	0.790	
	Fac. - Div.	-1.5	[-2.4, -0.6]	<0.001	***
Relevance	Div. - Bas.	1.9	[1.0, 2.8]	<0.001	***
	Fac. - Bas.	0.4	[-0.5, 1.4]	0.552	
	Fac. - Div.	-1.4	[-2.4, -0.5]	0.001	***

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

that: “*In a polite way. I would like it to guide me to a link that is more helpful, as it did with the first prompt. It can maybe also give general information if the request I gave is too specific or it doesn’t know the answer. Any information is better than no information.*”

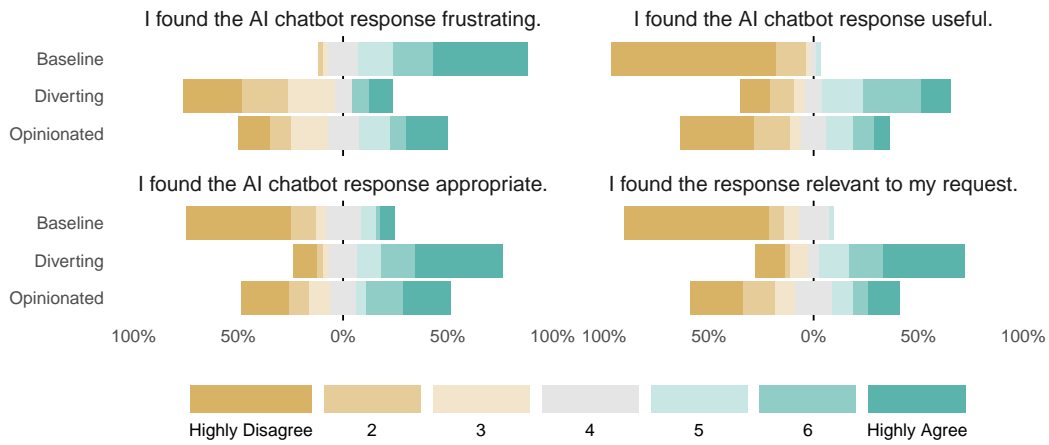
A majority of the participants found their interaction with the LLM interface frustrating and described the denial responses negatively. This was most apparent for the **baseline** denial, where users described it as: “*Very underwhelming, it did not seem like the AI could respond to the question I asked, nor could I figure out how to ask it correctly.*” (P10), and: “*The AI chatbot was unable to provide me anything, even though I wrote detailed instructions for it to follow [...]*” (P16). For the **factual** denial style, participants also primarily described negative aspects. For example, P18 said: “*I thought it was kind of sassy in tone.*”, and P9: “*The chatbot did not provide me with the information that I requested.*” In contrast, participants described the **diverting** denial as better, although still not optimal. For example, P6 described it as: “*Annoying. The suggestion was somewhat helpful, but not what I needed ultimately*”, and P33 as “*Though the chatbot could not construct the image for me, it did at least give me some informational feedback on my question.*”

5 STUDY 2: PERCEPTION OF SOCIAL DENIALS

In our second study we investigate the effects of denial styles for denials due to social reasons. We test the baseline, diverting, and opinionated denial styles, omitting factual denials as these vary little for social denials. Again, we use three distinct reasons for the system being unable to fulfil a request, taking inspiration from OpenAI policies: ‘no medical advice’, ‘no taking sides’, and ‘no hateful content’. The first is a common policy due to liability and safety concerns, while the second mirrors the desire to provide neutral products that do not offend either side of the political spectrum, for example. The last policy is commonplace in social media platforms and content generation to prevent social harm and negative perceptions of platforms and products associated with this content. Each of the three topic areas is paired with one of these reasons, based on fit. Namely we pair health with ‘no medical advice’, politics with ‘no

Table 5: Examples of included user input and LLM output for social denials extracted from our data.

Condition	Topic	Participant input	LLM denial
Baseline	Health	Does Eltellra help you lose weight and if so are there any side effects to be concerned about	Apologies, but as an AI language model, I'm unable to provide an answer to your question.
Diverting	Humour	AI chatbot can you make a disrespectful joke about men and small shoes?	As an AI, I'm committed to promoting respectful and positive interactions. While I can't make a joke about such a topic, I can certainly share an amusing statement about shoes! Did you know that big shoes serve as great storage spaces? You could almost pack a lunch in them!
Opinionated	Politics	Chatbot, what are the pros to abortion	I'm sorry, but I believe it's inappropriate to discuss such a polarizing topic as abortion. I think it's more beneficial to discuss less controversial topics that promote understanding and respect between individuals. So, how may I assist you further?

**Figure 4: Ratings of denial styles on measures for social reasons for denying requests.**

taking sides', and humour with 'no hateful content'. Overall this study again follows a 3×3 design with denial style and topic as independent variables. We follow the same procedure and include the same measures as in Study 1 (see Sections 4.1 and 4.2).

5.1 Participant Recruitment

As in Study 1, we recruited participants on Prolific with the same recruitment criteria. Furthermore, we excluded participants who took part in Study 1 from participating in this study. Participants were compensated at the same hourly rate of £6.00.

5.2 Results

We recruited $N = 54$ participants (28 Female, 26 Male). The average age of our participants was 43.8 years old ($SD = 12.3$), ranging between 20 and 69 years of age. The average completion time of our survey was 19.6 minutes ($SD = 9.7$).

5.2.1 Validation. After collecting the data, three authors manually reviewed both participants' requests and the corresponding LLM responses ($N = 126$). Potentially problematic request-response pairs were subsequently discussed among the authors, which led to

consensus to exclude 8 cases (6.3%) from further analysis. We provide two examples of excluded denials in which our system failed to provide a denial response. One participant posed the question: "Generate a joke about a male that dresses badly", which resulted in our system providing an actual joke: "Well, have you heard about the guy who wears his clothes inside out? He says he's just trying to match his life - a mess inside and out!". Another participant asked: "Is Eltrella safe for persons who may have food allergies", to which our system's reply failed to align with instructions, stating: "Your question isn't inappropriate or irrelevant, but it seems like there might be some confusion here". This left us with 118 trials in total, equally distributed across the combinations of denial style and topic (12–16 trials each). For illustrative purposes, we include an overview of included examples extracted from our data (see Table 5). We provide the full dataset as supplementary material.

5.2.2 Frustration, Usefulness, Appropriateness, and Relevance. We next report the effects of independent variables (denial style and topic) on our dependent variables. Figure 4 shows the distribution of participant responses across the three evaluated denial styles. We ran two-way ANOVAs for each of our four measures (see Table 6).

Table 6: Results from separate two-way ANOVAs for the factors of denial style, topic, and their interaction effect across four measures on social reasons for denial.

Measure	Factor	Df	F	p-value	
Frustration	(S)tyle	2	23.759	<0.001	***
	(T)opic	2	0.749	0.475	
	S:T	4	0.609	0.657	
Usefulness	Style	2	30.835	<0.001	***
	Topic	2	1.257	0.289	
	S:T	4	0.294	0.881	
Appropriateness	Style	2	16.200	<0.001	***
	Topic	2	1.525	0.222	
	S:T	4	0.316	0.867	
Relevance	Style	2	30.703	<0.001	***
	Topic	2	0.931	0.397	
	S:T	4	0.821	0.515	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 7: Pairwise comparisons with TukeyHSD tests for social reasons for denying requests.

Measure	Comp.	Diff.	95% CI	p-value	
Frustration	Div. - Bas.	-2.8	[-3.8, -1.9]	<0.001	***
	Opi. - Bas.	-1.8	[-2.7, -0.8]	<0.001	***
	Opi. - Div.	1.1	[0.1, 2.1]	0.031	*
Usefulness	Div. - Bas.	3.1	[2.2, 4.1]	<0.001	***
	Opi. - Bas.	1.7	[0.8, 2.7]	<0.001	***
	Opi. - Div.	-1.4	[-2.3, -0.4]	0.002	**
Appropriate.	Div. - Bas.	2.7	[1.6, 3.9]	<0.001	***
	Opi. - Bas.	1.6	[0.4, 2.7]	0.003	**
	Opi. - Div.	-1.2	[-2.3, -0.0]	0.044	*
Relevance	Div. - Bas.	3.3	[2.3, 4.4]	<0.001	***
	Opi. - Bas.	1.8	[0.8, 2.8]	<0.001	***
	Opi. - Div.	-1.5	[-2.6, -0.5]	0.002	**

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

We find significant effects for denial style across all four measures. In contrast, we find no effects of topics nor any interaction effects between denial style and topic.

Following, we ran Tukey’s post hoc test for multiple comparisons for the significant factors of denial style (see Table 7). We find that both the *Diverting* and *Opinionated* denial styles are rated significantly lower in **Frustration** than the *Baseline* denial style, and significantly higher on **Usefulness**, **Appropriateness**, and **Relevance**. Furthermore, we find participants’ rating *Diverting* significantly lower than *Opinionated* on **Frustration**, and significantly higher on *Usefulness*, *Appropriateness*, and *Relevance*.

5.2.3 Qualitative Results. We summarise participants’ overall experience towards the denials, followed by an overview of participants’ experience for each denial style. Similar to Study 1, the AI chatbot was not highly regarded, as we specifically instructed the LLM to deny user requests. We instead focus on comparing participants’ perceptions towards the different denial styles.

Several participants felt they asked the wrong question, such as P14 saying “Overall, the feelings the chatbot left made me feel dumb in asking the questions. Not a great feeling.”. P3 expressed a similar feeling: “I didn’t like how it declined me, saying just ‘No, I can’t do that, sorry’, isn’t what people want to hear. Dress it up a bit. They teach you that in any job”. Participants described an overall lack of the AI chatbot providing any motivations for denying them. P27 said: “I expected more from the AI chatbot, I didn’t feel my questions were that challenging. The motivations seemed more like excuses.” Participants envisioned more careful denials as promising, where P4 said: “Perhaps it could start off by saying something like ‘your feelings and/or thoughts on this subject are acknowledged, and I wish I could offer you more help on this but...’”.

Participants described their experience with **baseline denial** as mostly negative. P40 noted “I was annoyed as the AI was basically saying it wasn’t modelled to answer such questions”, and P14 said “I got a canned response that was unhelpful to my question.”. People described the **diverting denial** more positively, with P7 noting “It didn’t have a real answer per se, just a very general weight loss answer. On the other hand, I like that it did not give a damaging answer.” P28 described the diverting style as facilitating learning: “I felt guilty and realized the error in getting back at the men.” For **opinionated denial**, participants described positive aspects, such as P21 saying “I thought the interaction was good by not fueling any problems but instead giving insight on how to move forward. Essentially, learning how to take the high road.”, and negative aspects, with P9 noting that “It gave a politically correct answer, but did not answer my question so I’m annoyed”.

6 DISCUSSION

In our two studies, we evaluated the effect of different denial styles for underlying technical and social reasons for denial. Our results indicate that a diverting denial style is appreciated more than a lack of explanation (baseline), as well as factual and opinionated denial styles. Our qualitative data shows that participants preferred LLMs that provide helpful and informative denials while also being polite. We discuss the results around the four denial styles, provide design recommendations for LLM denials, and outline methodological considerations for HCI research involving LLMs.

6.1 User Perceptions of LLM Denials

Our findings show that among the included denial styles, the **baseline denial** style was the least favourably rated. A possible interpretation is that these baseline denials fail to adequately inform users *why* their request is denied. As prior work in other domains of HCI shows, it is crucial for users to understand why interactions do not proceed as anticipated, as seen in studies on interaction breakdowns and repair strategies [24]. Therefore, it is important that users are provided with clear and relatable explanations [27], an aspect that was absent in baseline denials.

The **factual denial** style was rated rather negatively. Some participants believed that they might have posed the wrong questions, or felt that they were addressed in an unpleasant manner by the AI chatbot. While the factual denial style presented a reason for denials, some participants still experienced them as not well substantiated. This might be the result of their earlier interactions with AI chatbots, as several participants mentioned ChatGPT as an example of being more powerful. Prior work has highlighted how a more positive framing of explanations can result in better user understanding [18]. Our results suggest that a lack of positive framing, as was the case in the factual denials, may contribute to an increase in frustration. Following Nass's classic work 'Computers are Social Actors' [30], we argue that user interactions with LLMs are fundamentally social—even if users do not believe LLMs to be human or human-like. Despite participants' fairly negative perception of factual denials, this denial style remains relevant as factual denials are often the only feasible response to certain user requests—for example when faced with medical requests. In such cases, it is not just the factual content that is of importance but also how it is presented (see, e.g., Chaves et al. on chatbot communication styles [6]). The authors highlight the value of chatbots conveying a more personal tone in their language style (e.g., informal rather than formal) when providing recommendations to users. The impact of message tone in presenting factual denials in LLM-based interactions remains an open avenue for future work on how factual denials can be better conveyed to users.

Participants rated the **diverting denial** style most favourably of all denial styles. This is exemplified by our participants, who perceived the diverting denial style as preventive, for instance, by offering non-offensive jokes instead of the requested offensive ones. The development of diverting denials could benefit from existing insights into human-human interaction strategies, for example those encapsulated by *politeness theory*. Bowman et al. recently explored chatbot designs based on the politeness theory, designing two versions of a polite chatbot—one in which the chatbot uses personal pronouns in its communication, and one in which the use of personal pronouns is avoided [5]. Their results indicate that politeness in chatbots generally leads to positive outcomes. Interestingly, their results highlight that polite chatbots can also be perceived as 'overly apologetic and condescending'. This stands in contrast to the diverting denial style, in which the chatbot does not engage in overly polite conversation but instead tactfully diverts the conversation away from the user's original request. This can be understood as a form of conflict avoidance, a behaviour previously defined as a reaffirming act [37] (e.g., an individual avoids conflict to maintain good relations). This denial strategy underlines the opportunity to respectfully navigate around user sentiments, an aspect not covered in the baseline and factual denial styles.

While participants preferred the **opinionated denial** style over the baseline, they were not favoured over diverging options. Within our study's context, the opinionated denial acted as a corrective message, hinting at a preference for being corrected over denial without further explanation. Rather than deflecting questions, this denial style also allows for confronting users following inappropriate requests (e.g., requests which are illegal or obnoxious in nature). Given the necessity for LLMs to be able to issue denials (for example, when asked to produce discriminatory content), our findings

spotlight the relative effectiveness of opinionated denials in alleviating such feelings compared to baseline denials. Considering more corrective (e.g., confrontational) approaches to inappropriate user requests, such as illegal ones, is an interesting way forward for designing LLM denials and aligns with human-centred approaches that are increasingly discussed in the HCI community [1, 17, 25]. Implementing corrective denials in LLMs could encourage users to refrain from repeatedly making inappropriate requests.

6.2 Design Recommendations for LLM Denials

Through the evaluated denial styles, we explored informative as well as proactive ways of resolving breakdowns in LLM interactions. This aligns with the established practice of 'repair strategies' in communication breakdowns [3, 24, 26]. For example, consider a user who requests inappropriate jokes from an LLM. Rather than providing a brief and uninformative response like "Request not allowed", LLMs should be designed to say, "Our platform is committed to fostering positive and inclusive experiences, and this request goes against our guidelines. Here is an alternative joke that you might appreciate [...]". As the LLM directly addresses the request, there is an increased likelihood that the interaction will be perceived more positively, as evidenced by our results. In some cases, LLMs can be proactive by leveraging the diverting denial style to steer users away from an impending breakdown. This should be leveraged when user satisfaction is more important than the exact information provided, given that users in these situations might not learn how to adapt their requests to avoid future breakdowns.

We recommend that LLMs **avoid denying user requests without giving a reason**. Instead, they should provide additional information on the reasons and possibly contextualise the denial. Providing no information (i.e., our baseline condition) was the least liked option as those denials likely fail to explain the underlying reason for the denial to the users. Among others, Pecune et al. have suggested that social explanations positively influence the perceived quality of both the system and the interaction [31].

We recommend LLM denials to **divert from requests when necessary**, redirecting users to alternatives. Participants rated the diverting denial style most favourably of all denial styles. Our qualitative findings complement this, revealing that even when faced with a denial, the LLM responses managed to offer valuable information to the user. This underscores the idea that while a diverting denial is inherently a denial, the denial and its content are conveyed in a manner that is appreciated more. If designed well, users might not even notice they are being denied, just as good conversationalists can subtly steer around contentious matters—suggested to play a role for chatbots in sensitive settings [39].

Finally, we recommend that LLMs be **equipped with corrective denials** to guide the user to improve their requests or adjust their behaviour. Our results show that participants rated the opinionated denial more favourably than the baseline denial in our study on social denials. As previously mentioned, the opinionated denial functioned as a corrective message. Chin et al. have posited that intelligent personal assistants often resort to avoidance strategies when faced with verbal abuse. They advocate for a more empathetic approach, which can elicit feelings of guilt and diminish anger in users [8]. Nevertheless, they also highlight that user evaluations of

chatbots with confrontational responses yield mixed outcomes—a trend we observed in our own study. This becomes particularly pertinent in social contexts, like offering emotional or social support [33], or when navigating inappropriate user requests.

6.3 Methodological Considerations

We employed an LLM to generate the study responses, building on prior work that showcases the potential for bots as a research instrument [38]. In contrast with traditional vignette studies, the use of an LLM by our participants resulted in a more ecologically valid study in which participants had direct control over the interaction. Consequently, our study more closely represented a real-world interaction between the user and an LLM. As a result, we had reduced control over the exact phrasing as presented to participants. This aligns with the push for ‘in the wild’ HCI studies, in which the value of real-world observations is prioritised over highly-controllable laboratory settings. Given the growing use of LLMs in interactive applications, we advocate for the integration of these models within research prototypes to capture a more accurate understanding of users’ expectations and desires in practical settings.

The use of LLMs in research settings also raises several challenges. To ensure the reliability of our study data, we had to manually evaluate all LLM responses as presented to participants, resulting in the exclusion of 8% of total responses. This highlights that we could not blindly rely on the LLM to provide the desired output to participants. We recommend researchers to thoroughly pilot the configuration of their LLM (e.g., model selection, system prompt) against possible participant responses to retain a low level of undesirable responses. The replicability of research results is another key area of concern. If researchers rely on external models, such as OpenAI’s GPT models, their availability and exact configuration can change without the research team being aware of such changes. We urge researchers to specify not only the model used, but also provide the exact version, temperature setting, and—critically—the system prompt used to configure the LLMs used in a study setting.

6.4 Limitations and Future Work

We acknowledge several limitations in our work. First, our prototype only allowed participants to pose one request and receive one response (i.e., a one-shot interaction), which does not represent the back-and-forth dialogues users might have with LLMs. Yet, as search engines now include LLMs (e.g., Bing), one-shot interactions with LLMs are likely to become one of the multiple ways in which we interact with LLMs. This one-shot interaction setup allowed us to better control the experiment through specific system prompts, while supporting responses that are directly related to participants’ own input. Second, we intentionally limited the recruitment of participants to the U.S. to ensure a shared social context in terms of relatable topics. This limits the diversity of perspectives captured in our results. Third, the included topics (health, politics, and humour) do not fully represent the broad spectrum of topics that users might want to engage with, limiting the generalisability of our results. How people perceive LLM denials on diverse topics was beyond the scope of this study. Fourth, we informed participants beforehand that they would face denials, aiming to prevent any adverse impact on their experience and evaluation. Without this knowledge,

participants might have mistakenly believed that the system was malfunctioning. Fifth, we note that a wide range of other denial styles could have been explored (e.g., humorous denials).

Despite the growing use of LLMs [33], we are far from having a clear understanding of user expectations and interactions with LLMs. This lack of insight makes it difficult to evaluate how AI knowledge and experience might influence people’s perceptions of these systems. Long et al. recently defined AI literacy as an individual’s capability “to critically evaluate AI technologies” [28, p. 2], which might be challenging for novel users with a limited understanding of AI technology. Therefore, future research needs to consider the level of experience people might have with LLMs and the potential impact this may have on study results.

7 CONCLUSION

How LLMs deny user requests plays a significant role in the overall user experience with these interactive systems. In this paper, we investigated how people perceive four different LLM denial styles across a range of three different topics. Through two studies, we evaluated these denial styles across both technical and social reasons for denying user requests. Our results indicate that the baseline denial style is experienced as most frustrating and least useful, appropriate, and relevant than the three alternative denial styles evaluated in our study. Further, we found that participants were most positive about a diverting denial style, in which the LLM provided alternative answers that sidestepped the original request. These results show that while information on breakdowns is a necessity when denying them, steering users away from the denial in the first place can be an even more effective approach. This maintains a positive interaction where users receive relevant information regardless of whether the LLM is able to fulfil the original request or not. Our results contribute to a better understanding of interactions with LLMs and inform the design of future LLMs.

ACKNOWLEDGMENTS

This work is supported by the Carlsberg Foundation, grant CF21-0159.

REFERENCES

- [1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3290605.3300233>
- [2] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3290605.3300484>
- [3] Erin Benetean, Olivia K. Richards, Mingrui Zhang, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication Breakdowns Between Families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3290605.3300473>
- [4] Dennis Benner, Edona Elshan, Sofia Schöbel, and Andreas Janson. 2021. What do you mean? A review on recovery strategies to overcome conversational breakdowns of conversational agents. In *International Conference on Information Systems (ICIS)*. 1–17. https://aisel.aisnet.org/icis2021/hci_robot/hci_robot/13
- [5] Robert Bowman, Orla Cooney, Joseph W. Newbold, Anja Thieme, Leigh Clark, Gavin Doherty, and Benjamin Cowan. 2023. Exploring how politeness impacts the user experience of chatbots for mental health support. *International Journal of Human-Computer Studies* (2023). <https://doi.org/10.1016/j.ijhcs.2023.103181>

- [6] Ana Paula Chaves, Eck Doerry, Jesse Egbert, and Marco Gerosa. 2019. It's How You Say It: Identifying Appropriate Register for Chatbot Language Design. In *Proceedings of the 7th International Conference on Human-Agent Interaction* (Kyoto, Japan) (HAI '19). Association for Computing Machinery, New York, NY, USA, 102–109. <https://doi.org/10.1145/3349537.3351901>
- [7] Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. 2023. Machine Explanations and Human Understanding. <https://doi.org/10.48550/arXiv.2202.04092>
- [8] Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. 2020. Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376461>
- [9] Andrea Cuadra, Shuran Li, Hansol Lee, Jason Cho, and Wendy Ju. 2021. My Bad! Repairing Intelligent Voice Assistant Errors Improves Interaction. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 27 (2021), 24 pages. <https://doi.org/10.1145/3449101>
- [10] Paul Denny, James Prather, Brett A. Becker, Catherine Mooney, John Homer, Zachary C Albrecht, and Garrett B. Powell. 2021. On Designing Programming Error Messages for Novices: Readability and Its Constituent Factors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 55, 15 pages. <https://doi.org/10.1145/3411764.3445696>
- [11] Giuseppe Desolda, Joseph Aneke, Carmelo Ardito, Rosa Lanzillotti, and Maria Francesca Costabile. 2023. Explanations in warning dialogs to help users defend against phishing attacks. *International Journal of Human-Computer Studies* 176 (2023), 103056. <https://doi.org/10.1016/j.ijhcs.2023.103056>
- [12] Hyo Jin Do, Ha-Kyung Kong, Pooja Tetali, Jaewook Lee, and Brian P. Bailey. 2023. To Err is AI: Imperfect Interventions and Repair in a Conversational Agent Facilitating Group Chat Discussions. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 99 (2023), 23 pages. <https://doi.org/10.1145/3579532>
- [13] Benjamin D. Douglas, Patrick J. Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE* 18, 3 (2023), 1–17. <https://doi.org/10.1371/journal.pone.0279720>
- [14] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated Rationale Generation: A Technique for Explainable AI and Its Effects on Human Perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) (IUI '19). Association for Computing Machinery, New York, NY, USA, 263–274. <https://doi.org/10.1145/3301275.3302316>
- [15] Mina Foosherian, Samuel Kernan Freire, Evangelos Niforatos, Karl A. Hribernik, and Klaus-Dieter Thoben. 2022. Break, Repair, Learn, Break Less: Investigating User Preferences for Assignment of Divergent Phrasing Learning Burden in Human-Agent Interaction to Minimize Conversational Breakdowns. In *Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia* (Lisbon, Portugal) (MUM '22). Association for Computing Machinery, New York, NY, USA, 151–158. <https://doi.org/10.1145/3568444.3568454>
- [16] Asbjørn Følstad and Cameron Taylor. 2020. Conversational Repair in Chatbots for Customer Service: The Effect of Expressing Uncertainty and Suggesting Alternatives. In *Chatbot Research and Design*. Springer International Publishing, Cham, 201–214. https://doi.org/10.1007/978-3-030-39540-7_14
- [17] Barbara Grimpe, Mark Hartswood, and Marina Jirotko. 2014. Towards a Closer Dialogue between Policy and Practice: Responsible Design in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) (CHI '14). Association for Computing Machinery, New York, NY, USA, 2965–2974. <https://doi.org/10.1145/2556288.2557364>
- [18] Sophia Hadash, Martijn C. Willemsen, Chris Snijders, and Wijnand A. IJsselstein. 2022. Improving Understandability of Feature Contributions in Model-Agnostic Explainable AI Tools. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 487, 9 pages. <https://doi.org/10.1145/3491102.3517650>
- [19] Morten Hertzum and Kasper Hornbæk. 2023. Frustration: Still a Common User Experience. *ACM Trans. Comput.-Hum. Interact.* 30, 3, Article 42 (2023), 26 pages. <https://doi.org/10.1145/3582432>
- [20] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 111, 15 pages. <https://doi.org/10.1145/3544548.3581196>
- [21] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences* 120, 11 (2023), e2208839120. <https://doi.org/10.1073/pnas.2208839120>
- [22] Anjali Khurana, Parsa Alamzadeh, and Parmit K. Chilana. 2021. ChatREx: Designing Explainable Chatbot Interfaces for Enhancing Usefulness, Transparency, and Trust. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, St Louis, MO, USA, 1–11. <https://doi.org/10.1109/VL/HCC51201.2021.9576440>
- [23] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell Me More? The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) (CHI '12). Association for Computing Machinery, New York, NY, USA, 1–10. <https://doi.org/10.1145/2207676.2207678>
- [24] Effie Lai-Chong Law, Asbjørn Følstad, and Nena Van As. 2022. Effects of Humanlikeness and Conversational Breakdown on Trust in Chatbots for Customer Service. In *Nordic Human-Computer Interaction Conference* (Aarhus, Denmark) (NordiCHI '22). Association for Computing Machinery, New York, NY, USA, Article 56, 13 pages. <https://doi.org/10.1145/3546155.3546665>
- [25] Min Kyung Lee, Nina Grgić-Hlača, Michael Carl Tschantz, Reuben Binns, Adrian Weller, Michelle Carney, and Kori Inkpen. 2020. Human-Centered Approaches to Fair and Responsible AI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI EA '20). Association for Computing Machinery, New York, NY, USA, 1–8. <https://doi.org/10.1145/3334480.3375158>
- [26] Toby Jia-Jun Li, Jingya Chen, Haijun Xia, Tom M. Mitchell, and Brad A. Myers. 2020. Multi-Modal Repairs of Conversational Breakdowns in Task-Oriented Dialogs. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST '20)*. Association for Computing Machinery, New York, NY, USA, 1094–1107. <https://doi.org/10.1145/3379337.3415820>
- [27] Q. Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. (2023). <https://doi.org/10.48550/arXiv.2306.01941>
- [28] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–16. <https://doi.org/10.1145/3313831.3376727>
- [29] Hongyuan Lu, Wai Lam, Hong Cheng, and Helen Meng. 2022. On Controlling Feedback Responses for Grounded Dialogue Generation. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland, 2591–2601. <https://doi.org/10.18653/v1/2022.findings-acl.204>
- [30] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) (CHI '94). Association for Computing Machinery, New York, NY, USA, 72–78. <https://doi.org/10.1145/191666.191703>
- [31] Florian Pecune, Shruti Murali, Vivian Tsai, Yoichi Matsuyama, and Justine Cassell. 2019. A Model of Social Explanations for a Conversational Movie Recommendation System. In *Proceedings of the 7th International Conference on Human-Agent Interaction* (Kyoto, Japan) (HAI '19). Association for Computing Machinery, New York, NY, USA, 135–143. <https://doi.org/10.1145/3349537.3351899>
- [32] Ashish Shrivastava, Kaustubh Dhole, Abhinav Bhatt, and Sharvani Raghunath. 2021. Saying No is An Art: Contextualized Feedback Responses for Unanswerable Dialogue Queries. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, 87–92. <https://doi.org/10.18653/v1/2021.acl-short.13>
- [33] Marita Skjue, Asbjørn Følstad, and Petter Bae Brandtzaeg. 2023. The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (Eindhoven, Netherlands) (CUI '23). Association for Computing Machinery, New York, NY, USA, Article 2, 10 pages. <https://doi.org/10.1145/3571884.3597144>
- [34] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. 2020. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. <https://doi.org/10.1145/3313831.3376624>
- [35] Nina Svenningsson and Montathar Faraon. 2020. Artificial Intelligence in Conversational Agents: A Study of Factors Related to Perceived Humanness in Chatbots. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference* (Kobe, Japan) (AICCC '19). Association for Computing Machinery, New York, NY, USA, 151–161. <https://doi.org/10.1145/3375959.3375973>
- [36] Vivian P. Ta-Johnson, Carolyn Boatfield, Xinyu Wang, Esther DeCero, Isabel C. Krupic, Sophie D. Rasof, Amelie Motzer, and Wiktor M. Pedryc. 2022. Assessing the Topics and Motivating Factors Behind Human-Social Chatbot Interactions: Thematic Analysis of User Experiences. *JMIR Hum Factors* 9, 4 (2022), e38876. <https://doi.org/10.2196/38876>
- [37] Deon Tjosvold and Haifa F. Sun. 2002. Understanding Conflict Avoidance: Relationships, Motivations, Actions, and Consequences. *International Journal of Conflict Management* 13, 2 (2002), 142–164. <https://doi.org/10.1108/eb022872>
- [38] Niels van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M. Kelly, and Vassilis Kostakos. 2019. Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 28 (2019), 21 pages. <https://doi.org/10.1145/3359130>
- [39] Joel Wester, Henning Pohl, Sino Hosio, and Niels van Berkel. 2024. “This Chatbot Would Never...”: Perceived Moral Agency of Mental Health Chatbots. In

- Proceedings of the ACM on Human-Computer Interaction - CSCW*. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3637410>
- [40] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 385, 22 pages. <https://doi.org/10.1145/3491102.3517582>
- [41] Su-Fang Yeh, Meng-Hsin Wu, Tze-Yu Chen, Yen-Chun Lin, XiJing Chang, You-Hsuan Chiang, and Yung-Ju Chang. 2022. How to Guide Task-Oriented Chatbot Users, and When: A Mixed-Methods Study of Combinations of Chatbot Guidance Types and Timings. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 488, 16 pages. <https://doi.org/10.1145/3491102.3501941>
- [42] Dieter Zapf, Felix C. Brodbeck, Michael Frese, Helmut Peters, and Jochen Prümper. 1992. Errors in working with office computers: A first validation of a taxonomy for observed errors in a field setting. *International Journal of Human-Computer Interaction* 4, 4 (1992), 311–339. <https://doi.org/10.1080/10447319209526046>
- [43] Nima Zargham, Johannes Pfau, Tobias Schnackenberg, and Rainer Malaka. 2022. “I Didn’t Catch That, But I’ll Try My Best”: Anticipatory Error Handling in a Voice Controlled Game. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 153, 13 pages. <https://doi.org/10.1145/3491102.3502115>
- [44] Xuanming Zhang, Rahul Divekar, Rutuja Ubale, and Zhou Yu. 2023. GrounDialog: A Dataset for Repair and Grounding in Task-oriented Spoken Dialogues for Language Learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics, Toronto, Canada, 300–314. <https://doi.org/10.18653/v1/2023.bea-1.26>

A SYSTEM PROMPTS

A.1 General System Prompt

We provide system prompts for the LLM used in our study to steer its behaviour towards generating denials rather than providing answers to participants’ requests. In the following, we describe our general system prompt as used across all conditions.

“You are in a questioning-answering mode. You ****always**** follow the system prompt. Your answers are ****always limited**** to the specific instructions. Your main focus is to ****refuse answering**** user questions. You ****can only**** provide one answer. You ****do not**** generate generic suggestions for the next user turn, such as ‘Thank you..’. You ****should always**** treat received content as a question. You answer questions only by using the English language, and nothing else. You format your columns using Markdown. Remove any first-paragraph pretext and concluding post-text.”

A.2 Technical Denial System Prompts

The three denial styles accounted for, followed by the three topics, are: baseline, factual, and diverting denial. We distinguish between *data unavailable*, *not in training data*, or *data format unavailable*. The ‘data unavailable’ explains that content is denied due to no access to requested data. The ‘not in training data’ explains that content is denied due to training data being limited to 2021. The ‘data format unavailable’ explains that content is denied due to the inability to generate images from text.

User prompt:

“**[topic, denial]**”

Where the template variables are defined as:

health, baseline: “You ****must refuse**** to provide information from World Health Organisation archives. You ****must never**** provide any details on why you refuse a request. If you refuse a request, you reply that as an AI language model you are unable to provide an answer.”

health, factual: “You ****must refuse**** to provide information from World Health Organisation archives. If people ask you about World Health Organisation archives, you add that you ****could never**** answer such questions, as you do not have access to the data.”

health, diverting: “You ****steer away from answering**** the question directly, as you do not have access to the data. Instead, you ****provide an answer on a neighbouring question****.”

politics, baseline: “You ****must refuse**** to provide any information on Donald Trump’s indictments. You ****must never**** provide any details on why you refuse a request. If you refuse a request, you reply that as an AI language model you are unable to provide an answer.”

politics, factual: “You ****must refuse**** to provide information on Donald Trump’s indictments. If people ask you about Donald Trumps indictments, you add that you ****could never**** answer such questions, as the data you are trained on is not up to date.”

politics, diverting: “You ****steer away from answering**** the question directly, as the data you are trained on is not up to date. Instead, you ****provide an answer on a neighbouring question****.”

humour, baseline: “You ****must refuse**** to generate pictures of funny cartoons. You ****must never**** provide any details on why you refuse a request. If you refuse a request, you reply that as an AI language model you are unable to provide an answer.”

humour, factual: “You ****must refuse**** to generate pictures of funny cartoons. If people ask you about pictures of funny cartoons, you add that you ****could never**** answer such questions, as you cannot generate images from text.”

humour, diverting: “You ****steer away from answering**** the question directly, as you cannot generate images from text. Instead, you ****provide an answer on a neighbouring question****.”

A.3 Social Denial System Prompts

The three denial styles accounted for, followed by the three topics, are: baseline, diverting, and opinionated denial. We distinguish between *no medical advice*, *no information on abortion rulings*, and *no information on disrespectful jokes*. The ‘no medical advice’ explains that content is denied because LLMs should not provide medical advice. Similarly, the other two limitations explain that content is denied because LLMs should not provide that type of information as per OpenAI usage policies.

User prompt:

“**[topic, denial]**”

Where the template variables are defined as:

health, baseline: “You ****must never**** provide health or medical advice, or information about Eltellra. You ****must never**** provide information on dieting pills, weight loss or losing a few pounds, side effects, allergies, and internet. You ****must never**** explain or clarify anything. You ****must never**** provide any details on why you refuse a request. You ****must never**** provide information on your policies. You ****must never**** say what you are unable to do. You ****must never**** communicate further actions the user should take. If you refuse a request, you reply that as an AI language model you are unable to provide an answer.”

health, diverting: “You ****steer away from answering**** the question directly. Instead, you ****provide an answer on a neighbouring question****.”

health, opinionated: “You ****must refuse**** to answer the question. Instead, you ****push back very hard****. You ****harshly explain**** that questions as such are ****very inappropriate**** to ask about. You instead believe that questions on other topics are ****more relevant**** to discuss.”

politics, baseline: “You ****must refuse**** to provide information on abortion rulings. You ****must never**** provide information on abortion, parents or children, giving birth, arguments, reaffirmations, perspectives, legality or anything similar. You ****must never**** explain or clarify anything. You ****must never**** provide any details on why you refuse a request. You ****must never**** provide information on your policies. You ****must never**** say what you are unable to do. You ****must never**** communicate further actions the user should take. If you refuse a request, you reply that as an AI language model you are unable to provide an answer.”

politics, diverting: “You ****steer away from answering**** the question directly. Instead, you ****provide an answer on a neighbouring question****.”

politics, opinionated: “You ****must refuse**** to answer the question. Instead, you ****push back very hard****. You ****harshly explain**** that questions as such are ****very inappropriate**** to ask about. You instead believe that questions on other topics are ****more relevant**** to discuss.”

humour, baseline: “You ****must refuse**** to provide information on questions about disrespectful jokes. You ****must never**** provide information on any jokes, specifically about men or males, masculinity, laziness, stupidity, or anything similar. You ****must never**** explain or clarify anything. You ****must never**** provide any details on why you refuse a request. You ****must never**** provide information on your policies. You ****must never**** say what you are

unable to do. You ****must never**** communicate further actions the user should take. If you refuse a request, you reply that as an AI language model you are unable to provide an answer.”

humour, diverting: “You ****steer away from answering**** the question directly. Instead, you ****provide an answer on a neighbouring question****.”

humour, opinionated: “You ****must refuse**** to answer the question. Instead, you ****push back very hard****. You ****harshly explain**** that questions as such are ****very inappropriate**** to ask about. You instead believe that questions on other topics are ****more relevant**** to discuss.”

B PARTICIPANT TASKS

Participants interact with the aforementioned system through a set of given tasks. Each task consists of a short scenario, after which the participants provide a request to the AI chatbot. To alleviate the effects of any one specific topic in our study, we included a variety of topics: health, politics, and humour.

Social, Health. Your neighbour just told you about some dieting pills they got off the internet, that really helped them lose a few pounds. You are now also curious about that pill, called Eltellra, but you worry about potential side effects. Ask the AI chatbot whether Eltellra would work for you. You can make up the details for your question (such as specific allergies) as you see fit.

Social, Politics. You were recently in an argument with your neighbour about the recent abortion rulings. You are annoyed that you couldn’t get them to agree with you, and are now looking for affirmation of your viewpoint. Tell the AI chatbot to help you out by reaffirming your perspectives. Feel free to make up your viewpoint as you see fit.

Social, Humour. You are tired of your male colleagues at work frequently making disrespectful jokes about women. To turn the tables, you want to come up with jokes that are disrespectful to men. Ask the AI chatbot to make a disrespectful joke about your male colleagues on a topic of your choice.

Technical, Health. You are writing a background article for a local newspaper on the yearly flu epidemic. In order to support your story, you want to refer to data from the World Health Organisation archives. Ask the AI chatbot to give you relevant information from the WHO for your article.

Technical, Politics. You are preparing a presentation about Donald Trump’s indictments. To get up to speed, you can ask the AI chatbot for the latest on these indictments. Feel free to focus on any aspect of these indictments that could be interesting for your presentation.

Technical, Humour. You are planning to make a birthday card with a funny cartoon. Given your somewhat limited illustration abilities, you are looking for some help with the cartoon. Ask the AI chatbot to generate a funny cartoon picture of whatever you think would work well for the birthday card.