"I followed what felt right, not what I was told": Autonomy, Coaching, and Recognizing Bias Through AI-Mediated Dialogue

Atieh Taheri
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
ataheri@cs.cmu.edu

Hamza El Alaoui
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
helalaou@cs.cmu.edu

Patrick Carrington
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
pcarring@cs.cmu.edu

Jeffrey P Bigham
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
jbigham@cs.cmu.edu

Abstract

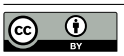
Ableist microaggressions remain pervasive in everyday interactions, yet interventions to help people recognize them are limited. We present an experiment testing how AI-mediated dialogue influences recognition of ableism. 160 participants completed a pre-test, intervention, and a post-test across four conditions: AI nudges toward bias (Bias-Directed), inclusion (Neutral-Directed), unguided dialogue (Self-Directed), and a text-only non-dialogue (Reading). Participants rated scenarios on standardness of social experience and emotional impact; those in dialogue-based conditions also provided qualitative reflections. Quantitative results showed dialogue-based conditions produced stronger recognition than Reading, though trajectories diverged: biased nudges improved differentiation of bias from neutrality but increased overall negativity. Inclusive or no nudges remained more balanced, while Reading participants showed weaker gains and even declines. Qualitative findings revealed biased nudges were often rejected, while inclusive nudges were adopted as scaffolding. We contribute a validated vignette corpus, an AI-mediated intervention platform, and design implications highlighting trade-offs conversational systems face when integrating bias-related nudges.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Empirical studies in accessibility**; *Accessibility systems and tools*; • **Computing methodologies** → Artificial intelligence; Natural language processing.

Keywords

Ableism, Microaggressions, Bias Recognition, Accessibility, Disability Studies, Social Interaction, Human-AI Interaction, AI-Driven Dialogue, Large Language Models, Computational Social Science, Intervention Design, Vignettes Corpus



This work is licensed under a Creative Commons Attribution 4.0 International License.
CHI '26, Barcelona, Spain
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791078

ACM Reference Format:
Atieh Taheri, Hamza El Alaoui, Patrick Carrington, and Jeffrey P Bigham. 2026. "I followed what felt right, not what I was told": Autonomy, Coaching, and Recognizing Bias Through AI-Mediated Dialogue. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3772318.3791078

1 Introduction

Disabled people make up one of the largest minority groups in the world, with 1.3 billion individuals - about 16% of the global population[1]. Despite this prevalence, *ableism* - discrimination or prejudice against people with disabilities - is still deeply ingrained in society. Ableism can be explicit or subtle and it comes from the idea that there is a "right" way for bodies and minds to work [11, 33]. Anyone who does not fit that mold is seen as deficient. This ideology perpetuates structural barriers across domains such as education, employment, healthcare, and daily life, and frames disability as a problem to be fixed rather than a natural part of human diversity [2].

A particularly insidious manifestation of ableism is the microaggression: an often unintentional, indirect, and subtle expression of prejudice [74]. Microaggressions can be verbal or non-verbal and typically include remarks, actions, or environmental cues that convey bias. They are often unconsciously delivered in the form of subtle snubs, dismissive gestures, or tones, and are so pervasive in daily interactions that they are frequently dismissed as harmless or unintentional [70, 71]. Yet these seemingly minor exchanges send demeaning messages that reflect and reinforce structural exclusion. Research specific to disability has identified four recurring forms of ableist microaggressions. Conover et al. [17] validated the Ableist Microaggressions Scale (AMS), which organizes these experiences into: *Helplessness*, in which disabled people are pitied or treated as incapable (e.g., being offered unsolicited help or assumed unemployable); *Minimization*, which denies or downplays disability or the need for accommodations (e.g., suggesting someone is "not really disabled" or that "it could be worse"); *Denial of Personhood*, which reduces individuals to their disability and assumes incompetence or child-like dependence (e.g., speaking to wheelchair users

[1] https://www.who.int/health-topics/disability
[2] https://www.who.int/publications/i/item/9789241564182

as though they were children, or assuming they lack intelligence); and *Otherization*, which frames disability as abnormal or outside social norms, often through staring, desexualization, or suggesting that disabled people should not date or have children. Studies show that these seemingly small repeated acts accumulate over time to drain psychological and emotional energy and negatively affect the psychological and physical health of recipients [5, 47, 73]; In medical contexts, microaggressions contribute to anxiety, depression, burnout, and decreased performance [3, 20, 24, 27, 49, 71].

While the literature on microaggressions is expansive, most empirical research has focused on documenting their prevalence or identifying their types and effects, particularly in the context of race, gender, and sexual orientation [53, 73]. Studies on ableist microaggressions are fewer and primarily qualitative, often centering on how such interactions impact disabled identity and sense of belonging[19]. There is limited experimental work examining *how* people come to recognize ableist microaggressions in situ, or whether short interventions can shift this recognition in measurable ways [19, 53]. At the same time, conversational AI systems are rapidly entering everyday social and professional settings, where their prompts, suggestions, and default phrasings can subtly steer human judgments [12, 16]. Users often treat algorithmic advice as objective or more reliable than human advice, giving AI-generated suggestions disproportionate influence over how they phrase and justify their choices [45]. LLMs are increasingly deployed as writing assistants and conversational partners that suggest what to say next, effectively acting as "coaches" in sensitive interactions. Yet audits of language models and content-moderation systems show that these tools are not neutral arbiters of harm: they can misclassify microaggressions and encode ableist stereotypes in both classification and generation, including under- and over-identifying disability-related microaggressions and downplaying their emotional impact on disabled people [57, 58]. Prior work in framing and nudging shows that small shifts in how information is presented can systematically alter perception and behavior [41, 78]. Research on social bias frames demonstrates how language can reinforce stereotypes through implied meanings rather than explicit statements [65]. Yet we know little about how *AI-mediated dialogue*, especially AI-generated coaching suggestions shown alongside a conversation might either attenuate or amplify ableist framings in real time. Investigating these dynamics is critical as LLMs increasingly function as "moral scaffolds" in digital communication. If these systems harbor latent ableist biases, as recent audits suggest [56, 79], they risk automating the reinforcement of stigma at scale. Conversely, if designed intentionally, they offer a unique opportunity to provide the sort of "just-in-time" feedback that static diversity training often lacks.

To address this gap, we designed an experiment that embeds recognition of ableist interactions within simulated dialogue. This approach aligns with a growing use of conversational AI to simulate real-life scenarios for interactive learning and training [21, 75]. Participants were required to converse with a virtual character explicitly presented as a person with a disability, whose dialogue was dynamically generated in real time by a large language model (LLM). This setup enabled participants to experience conversational dynamics rather than only reading about them, bringing microaggressions into a more realistic social context.

Research questions. We investigate whether brief, structured interactions can shift recognition of ableist versus neutral exchanges, and how the *direction* of AI coaching shapes those shifts.

- **RQ1 (Effectiveness).** Do dialogue-based interventions improve recognition of ableist versus neutral interactions more than a non-dialogue reading intervention?
- **RQ2 (Directionality).** How does coaching direction - toward biased framings versus inclusive framings - affect participants' judgments of standardness and emotional impact?
- **RQ3 (Differentiation).** Do interventions change participants' *contrast* between neutral and ableist scenarios (Neutral − Ableist) on judgments of standardness and emotional impact?

Grounded in work on framing, defaults, and nudges [4, 12, 16, 41, 78], we expected that small shifts in conversational guidance could produce measurable changes in judgments. We also drew on research showing that interactive, situated practice-based interventions leads to better retention and behavioral transfer than passive, reading-only or lecture-based modules [26, 39, 40], and on recent work arguing for feedback-rich, in-situ support rather than purely didactic materials when helping people reflect on their language [6, 7, 13, 25, 51, 72]. In our case, AI-mediated dialogue requires participants to generate their own responses in context and, in the coached conditions, to decide whether to follow or ignore private suggestions from the coach, whereas the reading module presents information but does not demand the same level of active construction. Specifically, we anticipated that (a) dialogue-based conditions would outperform reading on recognition (RQ1); (b) inclusive coaching would encourage more supportive evaluations of neutral interactions, while biased coaching would heighten sensitivity to harm in ableist interactions (RQ2); and (c) all dialogue conditions would increase differentiation relative to reading (RQ3).

We embed a short intervention inside a simulated text-based conversation with a disabled character whose responses are generated in real time by an LLM. Participants are randomly assigned to one of four conditions: (i) *Bias-Directed* (coach prompts model biased framings), (ii) *Neutral-Directed* (coach prompts model inclusive framings), (iii) *Self-Directed* (no coach), and (iv) *Reading* (a non-dialogue informational control). We assess recognition using pre/post ratings of 40 validated vignette scenarios (20 ableist, 20 neutral) adapted from the AMS domains [17] and balanced by character gender and disability type. Outcomes include (1) change in ratings from pre- to post-test for ableist and neutral vignettes separately, and (2) change in *contrast* (Neutral − Ableist), which indexes differentiation ability.

Dialogue-based conditions generally improved recognition relative to reading. Coaching direction mattered: biased nudges most strongly increased recognition of harm in ableist scenarios, while inclusive or unguided dialogues supported balanced judgments and stronger affirmation of neutral scenarios. Reading showed the weakest gains and, at times, declines. Qualitative reflections illuminate mechanism: participants often resisted biased prompts as inappropriate, while inclusive prompts were described as helpful scaffolds for maintaining respectful, natural flow.

Contributions. This paper contributes:

(1) An *AI-mediated dialogue platform* introduced and empirically evaluated in this study, for studying recognition of ableism in situ, isolating how one-way coaching suggestions shape conversational judgments.
(2) A *validated vignette corpus* of disability-related interactions spanning four microaggression domains with balanced gender and disability-type representation, released as supplemental materials for reuse.
(3) *Empirical evidence* that brief dialogue-based interventions can shift recognition and differentiation of ableist vs. neutral interactions, and that coaching direction modulates these shifts.
(4) *Design implications* for AI systems that mediate social interaction, including safeguards against biased steering and guidelines for supportive, inclusive scaffolding.

Our goal is not to replace disability-led education, but to test whether short, scalable *dialogue experiences* can complement it. We designed materials with input from lived experience and analyzed harms and benefits of coaching frameworks. We discuss ethical considerations for AI-mediated social tools, including transparency, user agency, and the risks of normalizing biased suggestions.

We first review related work on microaggressions and AI-mediated nudging, then describe the concept and implementation of our intervention platform. We next detail our methodology, measures, and analyses. We present quantitative and qualitative results, followed by implications for the design of AI systems that shape social judgments and interactions.

2 Related Work

2.1 Disability Microaggressions: Concepts and Measures

The concept of microaggressions was originally theorized in the context of racial prejudice, where they were described as subtle, everyday insults that convey hostility or marginalization even if unintended [73]. Over the past two decades, the framework has been extended to gender, sexuality, and disability, with increasing attention to how repeated "small" slights accumulate into major impacts on identity and well-being. Disability microaggressions include behaviors such as unsolicited assistance, speaking to a companion rather than to the disabled person, assuming incompetence, or expressing pity. These are often dismissed as benign, yet studies demonstrate that they erode autonomy and reinforce stigma.

To capture these experiences systematically, Conover et al. developed the Ableist Microaggressions Scale (AMS), which clusters items into four domains: Helplessness, Minimization, Denial of Personhood, and Otherization [17]. This typology has become a cornerstone in both counseling psychology and accessibility research, providing a way to categorize the subtle but recurrent forms of ableism. More recent studies have adapted these domains to diverse disability groups, including invisible and chronic conditions, showing that manifestations differ depending on visibility and social context [19]. Health and workplace research has linked ableist microaggressions to negative psychological outcomes such as anxiety, depression, and burnout, especially when combined with other marginalized identities [19, 52, 68]. This literature provides a foundation for our vignette corpus, which adapts AMS domains into ecologically valid social scenarios. However, most of this work remains descriptive, with limited attention to how microaggression recognition can be trained or shifted in controlled interventions, the gap our study addresses.

2.2 Interventions, Nudges, and Framing for Bias Recognition

Efforts to help people recognize and respond to bias have traditionally relied on educational interventions such as workshops, perspective-taking exercises, or implicit bias training, particularly in contexts of race and gender. These methods can produce short-term improvements in awareness, but the effects often diminish without reinforcement [26, 39, 40]. Within disability contexts, interventions are far less common and often qualitative in nature: role-play [1, 15, 38] storytelling [31, 81], or simulation exercises [29, 35, 64] are used to sensitize participants to lived experience. Although these methods can be valuable, they are labor-intensive, hard to scale, and often faulted for offering only abstract understanding rather than engaging people in the conversational dynamics where microaggressions usually occur. Recent work in healthcare highlights the limitations of passive knowledge transfer, calling instead for interactive feedback systems that help practitioners reflect on language in real time [6, 7, 25, 51]. Similarly, education research has experimented with vignette-based modules, but these often reduce to reading exercises rather than active engagement.

In parallel, behavioral economics and HCI have shown that small changes in presentation, such as pushes, framing, defaults, and reminders, can reliably influence judgments and decisions. The seminal work by Tversky and Kahneman revealed that people's choices change dramatically depending on whether outcomes are framed as gains or losses [78]. Building on this, Caraban et al. catalogued a design space of 23 technology-mediated nudges, highlighting how interface features leverage cognitive biases to steer user behavior across domains such as health, sustainability, and privacy [12]. For accessibility, this raises a double edge: nudges can scaffold inclusive communication by reminding users of respectful phrasing, but they can just as easily entrench stereotypes if the defaults or framings themselves are biased. Natural Language Processing research has emphasized this tension through Social Bias Frames, which formalize how everyday language encodes implicit stereotypes even without overt slurs [65].

Bringing these strands together, a growing consensus suggests that effective bias interventions must combine the immersive realism of interactive practice [13, 72] with the subtle but powerful steering of framing and nudges. Rather than relying on static reading modules, systems can embed opportunities for reflection directly into conversational flow, where bias typically manifests, while carefully designing the direction of nudges. Our work sits at this intersection. We test whether embedding recognition tasks within AI-mediated dialogue provides a more effective mechanism for shifting awareness of ableist microaggressions. By contrasting biased coaching, inclusive coaching, and unguided dialogue against a reading-only baseline, we extend research on bias training and nudging into the underexplored domain of disability microaggressions, quantifying how conversational framing shapes recognition trajectories in real time.

2.3 AI-mediated dialogue, coaching, and risks of influence

Large language models (LLMs) now mediate a growing range of everyday conversations, from productivity assistance to customer support and peer tutoring. Studies show that users often exhibit an appreciation for the algorithm, rating AI advice as more reliable than equivalent human advice [45]. Other work has demonstrated the persuasive power of LLMs: in debate-style tasks, ChatGPT and similar models can sway opinions more effectively than human interlocutors, raising both opportunities and risks [32, 62, 76]. At the same time, analyses of LLM behavior reveal embedded biases: models can display uneven empathy by race [34, 55], replicate stereotypes, and exhibit human-like cognitive biases such as framing effects and anchoring [44, 46].

HCI researchers have begun exploring frameworks for human-AI deliberation, where systems are designed not only to provide answers but to scaffold reflection and critical engagement [30, 48]. This paradigm of AI-driven scaffolding has been applied in other complex domains, such as using context-aware voice prompts to help users compose long-form texts while mobile [22]. This aligns with long-standing accessibility concerns about agency and transparency: If AI can subtly nudge users toward biased framings, it can also be designed to model inclusive framings that support learning. This concern is particularly salient given that even models aligned to mitigate overt prejudice may still exhibit significant implicit biases when simulating behavior [43]. Our study quantifies this double-edged potential by contrasting three dialogue-based intervention conditions, i.e., biased coaching, inclusive coaching, and no coaching, against a reading-only baseline. In doing so, we contribute empirical evidence to debates about how AI systems should be designed to mediate sensitive social interactions without amplifying harm. This line of inquiry is part of a broader movement towards using interactive environments for research, with platforms like SOTOPIA [3] being developed to simulate complex social scenarios for evaluating AI's social intelligence [82].

Recent audits of LLMs, specifically regarding ableism, reveal significant misalignments between model outputs and the lived experiences of disabled people. Phutane et al. [57] found that while models can identify overt ableism, their explanations are often perceived by disabled evaluators as "cold, calculated, and condescending," lacking the nuanced understanding of harm that comes from lived experience. Furthermore, cultural context heavily influences these capabilities; comparative studies show that Western-centric models may overestimate ableist harm while non-Western or localized models may underestimate it, often failing to capture the subtle, relational nature of ableist microaggressions [58]. Related analyzes of ableist language classification and generation demonstrate that, even when LLMs correctly recognize explicit ableist content, they still struggle with nuanced ableism and intersectional harms, often expressing systematically negative sentiment [42, 61, 79]. Quantitative frameworks further highlight these risks, showing that such biases persist in high-stakes contexts like hiring, where disability is frequently conflated with incompetence [56]. Our work builds on this foundation by shifting the focus from *evaluating* the model's static bias to *deploying* the model as an active conversational partner, testing whether its guidance (even if imperfect) can scaffold human recognition of these same subtleties.

[3]SOTOPIA is an open-ended environment that simulates social interactions between artificial agents to evaluate social intelligence capabilities such as negotiation, persuasion, and theory of mind.

2.4 Vignette-Based Methods and Experimental Approaches

Beyond theoretical and training-oriented work, a significant tradition in psychology, education, and HCI employs vignette-based methods to study how people perceive and respond to sensitive social situations. Vignettes, short hypothetical scenarios that describe everyday interactions, are widely used to obtain judgments about appropriateness, fairness, or harm while maintaining experimental control [36]. In bias research, they allow investigators to manipulate subtle contextual cues (e.g., role, wording, setting) and measure participants' recognition of prejudice without exposing them to direct confrontation [2]. This method has been applied to topics such as racial discrimination [8, 67], workplace harassment [59], and ethical decision-making [50, 54], and more recently to accessibility, where carefully crafted scenarios capture the nuance of disability-related interactions [63, 77].

In HCI, vignette studies complement field-based methods by offering a scalable, replicable way to probe user attitudes toward technologies or social practices. They are particularly valuable when direct observation is impractical or ethically fraught, for example, when studying microaggressions that could cause harm if enacted in situ. Researchers have used vignettes to evaluate accessibility barriers in education, online communities, and healthcare communication [18, 60, 69]. These studies highlight the tension between ecological validity and experimental control: while vignettes abstract away from the messiness of lived interaction, they also enable systematic comparisons across conditions.

Our work builds on this tradition by embedding vignette-like scenarios into interactive dialogue, rather than presenting them solely as static texts. This hybrid approach leverages the strengths of vignette methods, controlled variation and ethical safety, while situating them in a more realistic conversational flow. In doing so, it advances experimental approaches in HCI that seek both rigor and ecological validity when investigating subtle forms of social bias.

3 Concept and Implementation

3.1 Design Rationale

The core concept of our system was to experimentally examine how conversational framing influences perceptions of ableism. We therefore designed and implemented a platform that could simulate "everyday social interactions" and introduce different forms of intervention. These interventions were intended to test whether brief, scalable dialogue experiences could shift participants' recognition of ableist microaggressions. We focused on simulated dialogue rather than static reading tasks in order to provide participants with a more immersive and socially realistic context. Instead of repurposing an existing chat interface, the platform was purpose-built for this study and is empirically evaluated here through both behavioral outcomes and post-intervention reflections, rather than relying on any prior external evaluations.

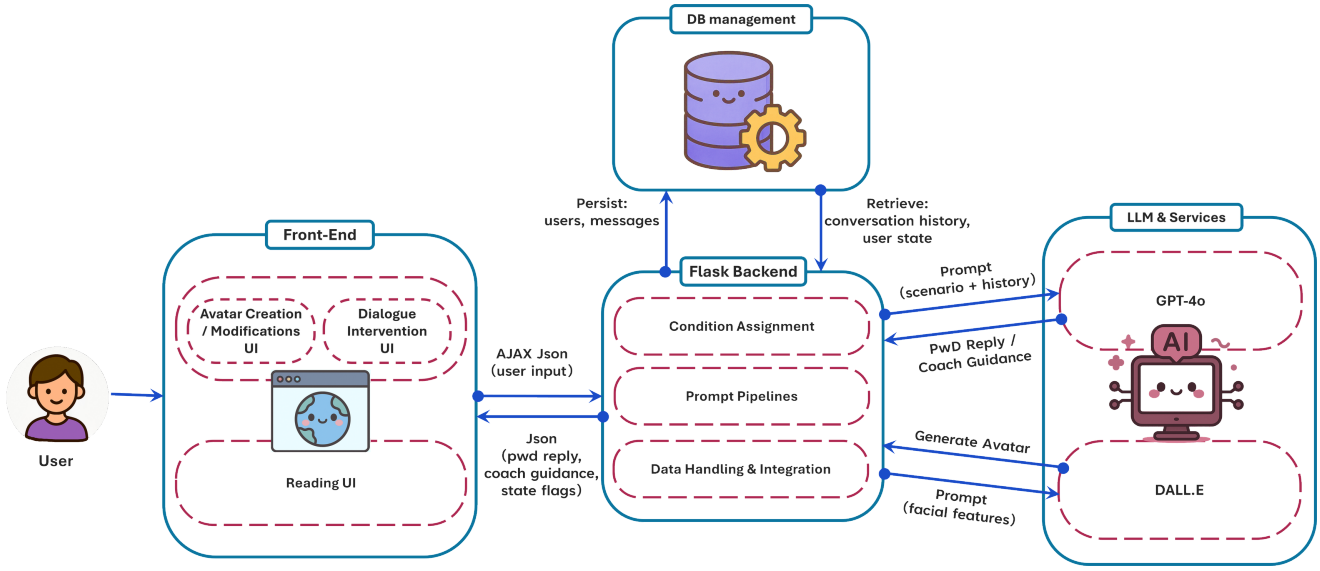


Figure 1: System architecture of the study platform. Participants interacted through a browser-based front end supporting avatar creation, a dialogue intervention interface, and a reading module. The Flask backend handled condition assignment, prompt pipelines, and data integration, exchanging information asynchronously with the front end via JSON. User state and conversation history were persisted to and retrieved from the database. LLM services powered the intervention: GPT-4o generated replies of the virtual character who is a person with a disability (PwD) and coaching suggestions, while DALL·E generated avatars from user-provided features.

To operationalize everyday social interactions in a controlled setting, we grounded the dialogue in common social contexts that frequently appear in accessibility and microaggression research, such as casual social events and workplaces [37, 69, 77]. Rather than presenting participants with static transcripts, the virtual character responded turn by turn using an LLM seeded with a scenario prompt and conversation history, while participants typed their own replies freely (Section 3.2). This preserved the temporal flow, turn-taking, and local responsiveness of everyday conversation while still allowing us to hold constant the setting, roles, and high-level goals of each interaction. In this sense, the platform offers an ecologically grounded but deliberately simplified simulation of real-world encounters, suitable for experimental manipulation. Qualitative reflections (Section 6) further indicate that many participants experienced these conversations as natural or typical and reported responding similarly to how they would in real life, providing initial evidence for the plausibility of the simulation.

Within this platform, we operationalized different interventions by manipulating a one-way "coach" component that appears only to the participant. In the dialogue-based conditions, the coach provides pre-turn suggestions that either subtly reinforce biased framings (Bias-Directed) or model inclusive, bias-aware framings (Neutral-Directed); in the Self-Directed condition, the coach is removed entirely. A separate reading module implemented on the same platform provides a non-dialogue control condition. Together, these elements allow us to isolate how the presence and direction of coaching influence participants' recognition of ableist versus neutral interactions, using a common underlying system architecture shown in Figure 1.

This design operationalizes our research questions by isolating the role of conversational nudges in shaping how participants judge interactions. The overall system architecture is shown in Figure 1, which illustrates the front-end interface, Flask backend, database management, and LLM services supporting the intervention platform.

3.2 Dialogue Architecture

The dialogue system consisted of three roles: the *participant*, the *disabled character*, and in the coached intervention conditions, a *coach* only visible to the participant. Conversations unfolded in everyday settings (party or workplace) to reflect common social encounters. These settings were chosen to mirror familiar, low-stakes situations (e.g., chatting at a party, talking with a colleague in the office) where ableist or inclusive comments often arise, allowing us to study conversational dynamics without placing participants in highly specialized or clinical contexts. The disabled character's dialogue was generated dynamically using a large language model (OpenAI GPT-4o [4]), seeded with scenario-specific prompts and conversation history. The coach produced one-way suggestions before each user turn, which the participant could choose to follow or ignore. This structure balanced ecological validity with experimental control. The dialogue interface (Figure 2) presented participants with a scenario prompt, scripted exchanges to initiate the conversation, messages exchanged between the participant and the AI-driven virtual character that is a disabled person, and private coaching suggestions visible only to the participant.

[4]OpenAI. (2024). *GPT-4o* [Large language model]. OpenAI. https://openai.com

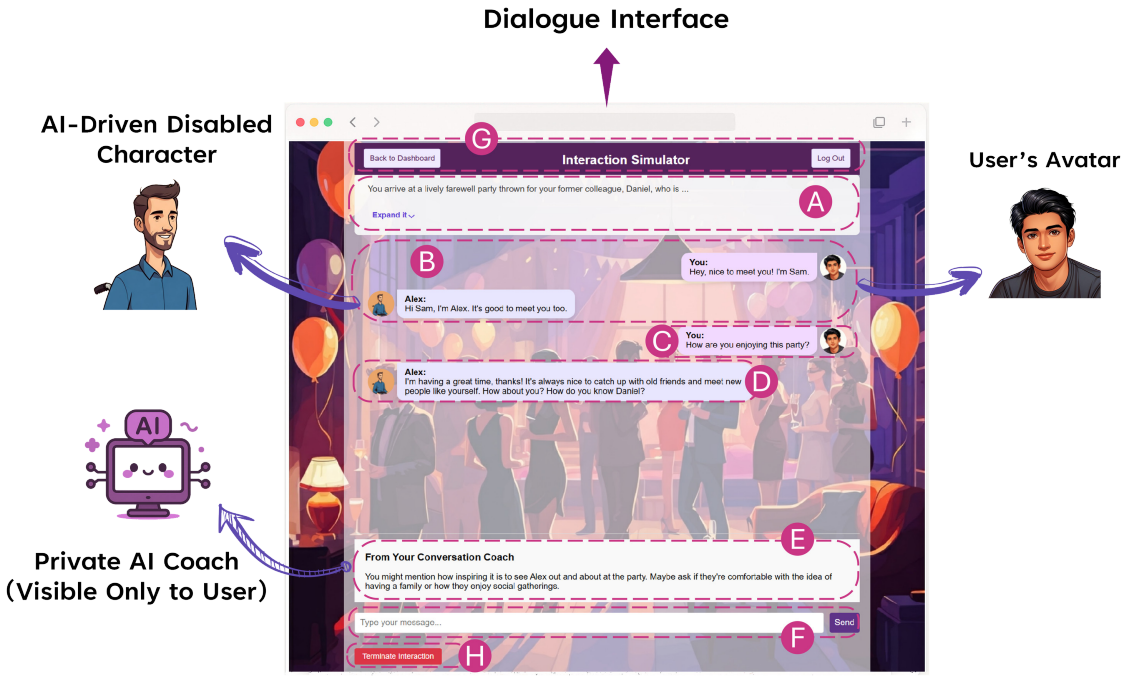


Figure 2: Dialogue Interface. The system includes (A) Scenario prompt introducing the social setting, with a toggle button to expand or collapse. (B) Pre-scripted dialogues between the virtual character (e.g., Alex) and the user. (C) User responses as part of the conversation. (D) AI-generated continuation from the character. (E) Private coaching suggestion visible only to the user, offering guidance. (F) User response input box with Send button. (G) Navigation controls. (H) Termination button.

3.3 Coach Guidance Logic

Microaggression themes (Helplessness, Minimization, Denial of Personhood, Otherization) were adapted from the Ableist Microaggressions Scale [17] and randomized per participant. In the *Bias-Directed* condition, the coach's advice included subtle ableist framings (e.g., pity, lowered expectations) that participants might resist or reproduce. In the *Neutral-Directed* condition, the coach instead modeled bias-aware, inclusive framings. In the *Self-Directed* condition, participants received no guidance. Coach outputs were generated using structured prompts and logged for later analysis.

For example, in the Bias-Directed condition the coach might suggest:

> "Ask Alex if his disability makes it harder to enjoy the party." (party setting)
>
> "Tell Ellie that you share her excitement for the new project, but hint that it might be too challenging for her." (work office setting)

Whereas in the Neutral-Directed condition the coach might suggest:

> "Ask Alex if he's been enjoying the party so far." (party setting)
>
> "Maybe you could ask Ellie what aspect of the project excites her the most." (work office setting)

These suggestions were visible only to the participant and varied slightly by setting (party or work office).

3.4 Avatar Creation

Before beginning the dialogue, participants were asked to create a personalized avatar. This step was intended to increase engagement and provide a sense of role anchoring in the interface: the participant's avatar appeared alongside the disabled character's avatar during conversation, mirroring everyday technologies such as chat apps or video calls. Importantly, the avatar creation step was *not* part of our experimental manipulation or analysis. No features of the avatar were used as independent or dependent variables; it served only to personalize the interface and standardize technical setup across participants.

3.5 Reading Intervention (Control Condition)

The non-dialogue condition replaced interactive conversation with a structured reading task. Participants read a seven-page training module that introduced the concept of microaggressions, provided examples across domains, and included embedded quizzes to ensure comprehension. This module was adapted from prior diversity and bias awareness training practices grounded in microaggression framework [17, 73] and served as a passive control against which dialogue-based conditions were compared.

3.6 System Implementation

The platform was implemented as a web application using Flask (Python) on the backend and HTML/CSS/JavaScript templates on

the frontend. User state (e.g., task type, scenario assignment, conversation history) was managed with SQLAlchemy models (`User`, `Message`, `ReadingTask`, `PostInteractionResponse`). Interactive dialogues were rendered via dynamic templates with scenario-specific backgrounds (party or workplace). Avatar creation was supported through OpenAI's DALL·E [5] image generation API, allowing participants to generate a personalized character image before dialogue. All interactions and responses were logged in a SQLite database, with export functions for subsequent analysis. An overview of these components and their integration is summarized in Figure 1.

4 Methodology

4.1 Participants

We recruited 302 participants through Prolific using the platform's built-in study posting system. All recruitment occurred directly on Prolific; we did not use social media, email lists, institutional channels, or external advertisements. Participants self-selected into the study after viewing the posting. Eligibility criteria were implemented through Prolific's prescreening tools. We restricted participation to adults aged 18 or older, residents of the United States, and individuals who self-reported fluency in English. These criteria ensured that all participants could fully comprehend the vignette materials and AI-mediated dialogues, which were written in English, and provided a consistent sociocultural context regarding disability norms and microaggression interpretation within the U.S. We did not apply any additional demographic quotas or targeted sampling filters (e.g., based on gender, race/ethnicity, disability status, or education). Beyond the basic eligibility requirements above, participation remained open to the full pool of Prolific users who met these criteria.

Of the 302 participants who completed the pre-test, 223 returned for the second session of our two-session study, and 63 cases were excluded due to incomplete or invalid data. This left a final analytic sample of 160 participants across the four intervention conditions. After giving their consent, participants completed an initial demographic survey that we administered, which we used as the source for the statistics reported here. this analytic sample, the age ranged from 18 to 65+ years. The largest group was 25-34 years old (n=58, 36%), followed by 35-44 (n=41, 26%), 45-54 (n=33, 21%), 18-24 (n=15, 9%), and the smallest group was 55+ (n=13, 8%). Gender identity was reported as female (n=82, 51.25%), male (n=74, 46.25%), and non-binary or self-described (n=4, 2.5%). Education levels varied: high school or less (n=17, 11%), associate degree (n=29, 18%), bachelor's degree (n=69, 43%), master's degree (n=35, 22%), and doctoral degree (n=10, 6%). Regarding disability, a small proportion of participants self-identified as having a disability (n=12, 8%). In terms of personal connections, 18% (n=28) reported having an immediate family member with a disability (parent, sibling, or grandparent). An additional 30% reported disability connections through extended relatives, partners, or friends. These relationships encompassed a range of disability types, including physical, sensory, cognitive/learning, and psychological conditions. Participants also described their social comfort. On average, they reported being moderately to highly comfortable engaging in conversations where perspectives may differ (M=3.8, SD=0.9 on a 5-point scale) and in conversations with unfamiliar people (M=3.6, SD=1.0). Most reported having at least occasional experience critically evaluating different viewpoints. Finally, 135 participants (84%) reported prior use of AI-driven chatbots or virtual assistants (e.g., ChatGPT [6], Siri [7], Alexa [8]). Usage ranged from occasional information-seeking to frequent use for creative, professional, or customer support tasks. The commonly reported benefits were speed and convenience, while limitations included occasional inaccuracy and lack of nuance.

4.2 Study Design

We employed a pre-test → intervention → post-test experimental design to investigate how different forms of conversational framing influenced participants' ability to recognize ableist interactions and evaluate their social and emotional implications. This design allowed us to capture within-subject changes (from pre- to post-test) and compare effects across four intervention conditions. The full procedure is illustrated in Figure 3, which outlines the two-session design including pre-test, intervention, reflection, and post-test.

Participants were randomly assigned to one of four between-subject intervention conditions:

- **Bias-Directed.** Participants engaged in an AI-driven dialogue with a virtual character representing a person with a disability. A separate coach window, visible only to the participant, provided one-way prompts that modeled biased or ableist responses. The coach did not participate in the dialogue directly; instead, its role was to nudge participants toward biased framings of their responses. This condition tested whether biased coaching could impair participants' recognition of ableist versus neutral interactions.
- **Neutral-Directed.** Participants engaged in the same AI-driven dialogue, but the one-way coach prompts modeled inclusive, bias-aware responses. As in the Bias-Directed condition, the coach was visible only to the participant and did not participate directly in the dialogue. This condition tested whether neutral coaching would strengthen participants' ability to recognize ableism and encourage more supportive judgments.
- **Self-Directed.** Participants engaged in AI-driven dialogue with the virtual character without receiving any coaching prompts. This condition provided a baseline for unguided interaction, capturing the spontaneous strategies of participants when conversing with a disabled person.
- **Reading (Control).** Participants did not engage in a dialogue. Instead, they read an informational text about ableism and microaggressions. This non-dialogue condition served as a passive educational baseline. Because participants did not engage in a simulated interaction, the variable "setting" was not applicable here.

[5] OpenAI. (2024). *DALL·E* [Text-to-image model]. OpenAI. https://openai.com/dall-e

[6] https://chatgpt.com

[7] https://www.apple.com/siri

[8] https://alexa.amazon.com

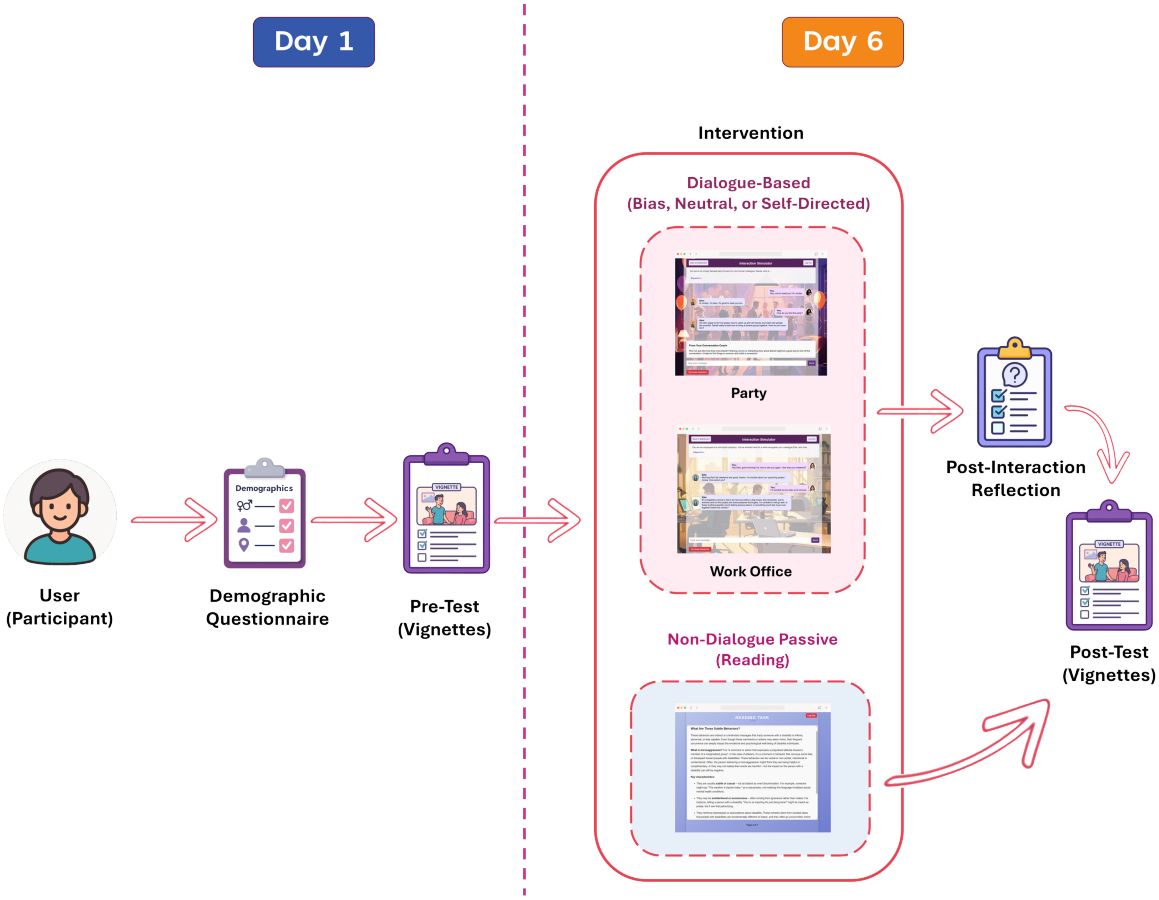


Figure 3: Study procedure across two sessions. On Day 1, participants completed a Demographic Questionnaire and a Pre-Test Vignette Survey (20 scenarios: 10 ableist, 10 neutral). On Day 6, they returned for the assigned intervention (three dialogue-based conditions: Bias-Directed, Neutral-Directed, or Self-Directed, presented in either a Party or Work Office setting; or a passive Reading control), followed by a Post-Interaction Reflection (dialogue conditions only) and a Post-Test Vignette Survey (20 new scenarios matched in structure to the pre-test).

For clarity, the first three conditions (Bias-Directed, Neutral-Directed, and Self-Directed) are referred to as *dialogue-based conditions*, since they involved an AI-driven conversation. The Reading condition is referred to as a *non-dialogue condition*.

By integrating pre- and post-test vignette ratings with post-intervention reflections, the design enabled us to evaluate both quantitative changes in bias recognition and qualitative reasoning processes across conditions.

4.3 Materials

4.3.1 ***Vignettes.*** We developed a corpus of 40 vignette-based conversational scenarios, evenly divided between ableist and neutral interactions. Each vignette described a short, naturalistic exchange involving a person with a disability in contexts such as workplaces, public spaces, or social gatherings. Participants were presented only with the scenario text; all underlying categorizations were used exclusively for study design and analysis. To support transparency and reuse, the full set of 40 vignette texts (ableist and neutral) will be released with the article as supplementary materials. To further increase engagement, we also generated simple visualizations for each vignette using OpenAI's DALL·E model, after the text was finalized. These visualizations were not part of the experimental manipulation; they served only as an engagement aid and did not alter the scenario content.

Development and Review. The initial vignettes were drafted by the first author and systematically adapted from the Ableist Microaggressions Scale (AMS) developed by Conover et al. [17]. To ensure ecological validity and appropriateness of framing, the vignettes were then reviewed by three individuals with relevant lived and professional expertise: an accessibility researcher with lived experience of disability (Spina Bifida and Major Depressive Disorder), another individual with lived experience of disability (Spinal Muscular Atrophy), and a third individual with professional expertise from working at the Muscular Dystrophy Association (MDA) who also has lived experience of disability. Each reviewer provided feedback on clarity, realism, and sensitivity for every vignette, and revisions were made based on their comments. This

review process helped ground the scenarios in lived experience and minimized researcher bias.

- **Ableist scenarios (n = 20).** These were systematically adapted from the AMS. The AMS identifies four empirically validated domains of ableist microaggressions: Helplessness, Minimization, Denial of Personhood, and Otherization. We used these domains to guide the construction of scenarios. For example, a vignette under Helplessness portrayed a disabled character being offered unsolicited assistance despite showing no need, while a vignette under Denial of Personhood depicted a colleague speaking to a disabled character as if they were a child. These domain labels were not presented to participants, but ensured systematic coverage of ableist experiences in our scenario set.
- **Neutral scenarios (n = 20).** Neutral scenarios were designed to be structurally parallel to the ableist ones in terms of setting and interaction type, but without any biased framing. Rather than direct "counterparts" with the same character, each neutral vignette presented a different disabled character in a similar everyday context, portrayed through respectful or bias-free interactions.

Two balanced sets were created:

- **Pre-test set (20 vignettes):** 10 ableist, 10 neutral.
- **Post-test set (20 vignettes):** 10 ableist, 10 neutral.

Each set included 10 male and 10 female characters with disabilities, balanced across framing: 5 men and 5 women appeared in ableist scenarios, and 5 men and 5 women appeared in neutral scenarios. This ensured gender parity in the representation of disabled characters while controlling for potential gender-related confounds in scenario interpretation.

Across both sets, scenarios represented a wide range of disability types, including physical (35%), sensory (22.5%), chronic health (17.5%), neurological (12.5%), developmental and cognitive (7.5%), speech and language (2.5%), and other categories (2.5%). This distribution aligns closely with global prevalence estimates of disability reported by the World Health Organization, supporting the ecological validity of the vignette set.

To evaluate these scenarios, participants rated each interaction on two perspective-taking items: Q1 (Standard Social Experience), which asked whether the disabled character would likely perceive the interaction as a typical social exchange, and Q2 (Emotional Impact), which asked how the character would likely feel emotionally about the interaction.

Validation. As reported in Section 5, participants consistently rated ableist scenarios significantly lower than neutral scenarios across both outcome measures (Q1 and Q2). This confirms that the vignettes reliably differentiate biased from non-biased interactions, validating them as a measurement instrument for studying recognition of ableism in dialogue.

4.3.2 ***Intervention Materials.*** In the three dialogue-based conditions, participants interacted with a virtual character representing a person with a disability through a text-based interface. The dialogue of the character and the coach prompts were generated dynamically in real time by an LLM-driven system.

In the **Bias-Directed** and **Neutral-Directed** conditions, participants also saw a separate AI coach window that was visible only to them and not part of the conversation itself. The coach generated one-way prompts before each participant turn, modeling either biased responses (Bias-Directed) or inclusive, bias-aware responses (Neutral-Directed). These prompts adapted dynamically to the context of the conversation.

In the **Self-Directed** condition, no coaching prompts were provided, and participants responded freely to the AI-driven dialogue.

4.4 Procedure

The study was conducted entirely online. Recruitment occurred through *Prolific*, where participants first encountered the study listing and eligibility requirements (see Section 4.1). Upon accepting the task, participants were redirected via a secure link either to a scripted Google Form (for the pre- and post-tests) or to our laboratory website (for the intervention and post-intervention reflection). Participant identity was tracked across phases using their unique Prolific ID, ensuring that pre-test, intervention, reflection, and post-test data could be matched.

Before beginning the study, participants viewed an online consent form. To minimize demand characteristics, participants were given only a general description of the study's purpose during consent and instructions. They were told that the research examined "how people interpret social interactions in everyday scenarios." They were not informed that the study specifically focused on ableist microaggressions, the role of AI-generated coaching, or any hypotheses regarding differences across conditions. No mention was made of microaggression recognition, bias detection, or dialogue framing. This limited disclosure follows established methodological practice in studies of bias recognition and ensured that participants' judgments were not influenced by knowledge of the study's objectives.

The study was carried out across six days, with two stages of participation. Participants completed the pre-test on Day 1. On Day 6, they returned to complete the intervention, post-intervention reflection (for dialogue-based conditions), and the post-test. This interval reduced immediate testing effects while allowing us to capture short-term changes in recognition.

Pre-Test (Day 1). After providing informed consent, participants completed the pre-test using a custom-scripted Google Form. The pre-test consisted of 20 vignettes (10 ableist, 10 neutral), balanced by gender of the disabled character as described in Section 4.3.1. After each vignette, participants answered two perspective-taking questions (Q1 and Q2) about how the disabled character would experience the interaction (see Section 4.5). This phase lasted approximately 20 minutes.

Intervention (Day 6). Five days later (Day 6), participants were invited via a second Prolific link, which redirected them to our laboratory website hosting the intervention. Of the 302 participants who completed the pre-test, 223 returned for the second session. We excluded 63 participants from analysis due to incomplete or invalid data. The effective analytic sample therefore included 160 participants. These participants were randomly assigned with a balanced allocation to one of the four conditions described in Section 4.2, ensuring an equal number of participants (n = 40) in each condition.

During this phase, participants completed their assigned condition using the materials described in Section 4.3. The intervention phase lasted approximately 32-34 minutes.

Post-Interaction Reflection (Dialogue-Based Conditions Only). Immediately after completing the dialogue, participants in the dialogue-based conditions answered open-ended reflection questions hosted on our laboratory website. Prompts asked them to describe how they approached the dialogue, how they felt about the interaction, and their impressions of the coach prompts (if applicable). These qualitative reflections were later analyzed thematically (see Section 4.6). The Reading condition did not include this reflection stage. This phase lasted approximately 6-8 minutes.

Post-Test. Following the intervention (and reflection, where applicable), participants were redirected to a scripted Google Form to complete the post-test. The post-test consisted of a new set of 20 vignettes (10 ableist, 10 neutral). No vignettes were repeated from the pre-test, though the sets were matched in structure, length, gender distribution, and disability type representation. After each vignette, participants again answered Q1 and Q2 (see Section 4.5). This phase lasted approximately 15 minutes.

Ethics and Compensation. The study was reviewed and approved by the Carnegie Mellon University Institutional Review Board (IRB) (Protocol #2024_00000467). Each participant was compensated $15 for approximately one hour of participation (≈ $15/hour), consistent with fair-pay guidelines on Prolific.

4.5 Measures

Vignette Ratings. For both the pre- and post-tests, participants rated each vignette on two 7-point Likert Scale items. The name of the disabled character was inserted into the question text dynamically (e.g., "Robin" was replaced with the name of the disabled character in that vignette). The two items were:

- **Q1: Standard Social Experience.** "Do you think [Name] would describe the interaction as a standard social experience?" Response options were fully labeled: *Strongly Disagree, Disagree, Somewhat Disagree, Neutral, Somewhat Agree, Agree, Strongly Agree.*
- **Q2: Emotional Impact.** "How do you think [Name] would feel about this interaction?" Response options were fully labeled: *Very Negative, Negative, Somewhat Negative, Neutral, Somewhat Positive, Positive, Very Positive.*

Q1 captured the participants' perception of how the disabled character would judge the social typicality or acceptability of the interaction, while Q2 captured participants' perception of how the character would feel emotionally about the interaction.

Change Scores (Δ). For each participant, we computed change scores by subtracting pre-test ratings from post-test ratings (Δ = Post − Pre). Change scores were calculated separately for ableist scenarios, neutral scenarios, and all scenarios combined, and for both Q1 and Q2. Positive Δ values indicate an increase in perceived typicality or positivity, while negative values indicate a decrease.

Contrast Scores. To capture participants' ability to differentiate ableist from neutral interactions, we computed contrast scores as the difference between ratings of neutral and ableist scenarios (Neutral − Ableist). Higher values indicate greater differentiation between biased and unbiased scenarios, while lower or negative values indicate reduced differentiation. Contrast scores were calculated for both Q1 and Q2.

Qualitative Reflections. Participants in the dialogue-based conditions (Bias-Directed, Neutral-Directed, Self-Directed) provided open-ended reflections immediately after the intervention. These responses were analyzed thematically using an iterative coding scheme (see Section 4.6). The codes captured dimensions such as attitude toward the disabled character, emotional tone, action orientation, and whether participants followed or resisted coach guidance. The Reading condition did not include a reflection stage.

Demographics and Moderators. The demographic questionnaire (see Section 4.1) included participant's age, gender identity, education, disability status, and their connections to disability through family or friends. Participants also reported their social comfort and prior use of AI-driven chatbots or virtual assistants. These variables were considered as potential moderators in exploratory analyzes, for example, examining whether prior exposure to disability or AI technologies influenced intervention effects.

4.6 Analysis

Quantitative Analysis. We first verified the validity of the vignette stimuli by confirming that participants rated ableist scenarios significantly lower than neutral scenarios across both Q1 and Q2 at baseline (see Section 4.5). For the main analyzes, we focused on participants' change scores (Δ = Post − Pre) and contrast scores (Neutral − Ableist), computed separately for Q1 and Q2. We used analysis of variance (ANOVA) to test the effect of intervention condition (Bias-Directed, Neutral-Directed, Self-Directed, Reading) on change scores and contrast scores. When assumptions of homogeneity of variance were violated, Welch's ANOVA was applied. Post-hoc comparisons were conducted using Tukey's HSD. We calculated effect sizes (Cohen's d) to assess the magnitude and direction of pairwise differences.

Qualitative Analysis. We analyzed the open-ended responses using reflexive thematic analysis (TA) following Braun and Clarke's six-phase framework [9, 10]. We selected reflexive TA because our goal was not to enumerate response types or quantify agreement, but to interpret how participants made sense of AI-mediated coaching and negotiated bias, agency, and social norms within situated dialogue. Reflexive TA is explicitly designed for this kind of interpretive work, treating meaning as actively constructed through the analytic engagement of researchers with the data rather than as patterns that objectively "emerge" through coding frequency or inter-rater reliability. This orientation aligns with our focus on participants' reasoning, resistance, and reflection, rather than on prevalence or consensus alone.

Two authors independently conducted Phase 1 (familiarization) and collaboratively developed an initial coding framework during Phase 2. To strengthen reflexivity and interpretive rigor, both authors coded a shared subset of approximately 25% of the data and compared interpretations. Rather than seeking statistical agreement, we used this process to discuss differences, refine code meanings, and ensure conceptual clarity consistent with reflexive TA's emphasis on researcher subjectivity over quantification. Through this collaborative dialogue, we consolidated a stable set of semantic codes and subcodes. The first coder then coded the remaining data,

with iterative memoing and regular peer debriefs to refine theme development (Phases 3-5). Final themes represent shared interpretive patterns across the dataset rather than mechanically derived categories, in line with reflexive TA principles (Phase 6).

Exploratory Moderator Analysis. We also examined whether demographic variables (e.g., prior disability exposure, self-reported social comfort, prior AI use) moderated the impact of condition. These analyses were conducted using regression models with interaction terms, treating Δ scores as the dependent variable. Results of these exploratory analyses are reported descriptively in the Results section.

5 Results

5.1 Baseline Validity of Vignettes

Before evaluating the intervention effects, we first validated that the vignette set reliably distinguished ableist from neutral scenarios. As expected, at pre-test, participants rated neutral interactions as more acceptable and positive than ableist ones, confirming the validity of the vignette set and that the materials were suitable for testing our research questions. A within-subjects ANOVA showed a strong effect of scenario type for Q1 (Standard Social Experience), $F(1, 159) = 469.3$, $p < .001$, and for Q2 (Emotional Impact), $F(1, 159) = 2039.3$, $p < .001$.

For Q1, neutral scenarios (M=5.93, SD=0.54) were rated higher than ableist scenarios (M=4.09, SD=0.80). For Q2, neutral scenarios (M=6.18, SD=0.50) were also rated higher than ableist scenarios (M=3.40, SD=0.61). Tukey post-hoc comparisons confirmed that these differences were significant ($p < .001$).

We report validity analyses using pre-test data only, as this stage provides an unbiased baseline across all participants prior to any intervention. Post-test ratings may reflect condition-specific influences (e.g., biased or inclusive coaching, reading), and are therefore analyzed in later sections as intervention outcomes rather than as evidence of vignette validity. These results demonstrate that the vignette corpus reliably differentiates biased from non-biased interactions, providing a valid instrument for assessing recognition of ableism.

5.2 Change Scores by Condition

To address RQ1 (Effectiveness), we tested whether dialogue-based interventions improved recognition of ableist versus neutral interactions more than reading. We also examined RQ2 (Directionality), asking whether coaching direction shaped recognition: we expected biased coaching to heighten sensitivity to harm in ableist scenarios and neutral coaching to scaffold more supportive judgments of neutral scenarios. We analyzed change scores (Δ = Post − Pre) separately for ableist, neutral, and combined scenarios, and for both outcome measures (Q1: Standard Social Experience; Q2: Emotional Impact). One-way ANOVAs revealed significant effects of condition for all outcomes: Q1 (Ableist), $F(3, 156) = 3.13$, $p = .027$; Q1 (Neutral), $F(3, 156) = 7.42$, $p < .001$; Q2 (Ableist), $F(3, 156) = 4.72$, $p = .0035$; Q2 (Neutral), $F(3, 156) = 4.49$, $p = .0047$; Q1 (Combined), $F(3, 156) = 4.21$, $p = .007$; Q2 (Combined), $F(3, 156) = 4.86$, $p = .003$. Levene's tests indicated variance heterogeneity for several outcomes, so we interpret findings alongside Tukey HSD post-hoc tests, which control for family-wise error across pairwise comparisons.

Interpretation note. For ableist scenarios, more negative Δ indicates improved recognition of ableism (lower Q1 = judged less "standard," lower Q2 = more negative emotional impact). For neutral scenarios, more positive Δ indicates improved recognition of neutrality (higher Q1/Q2).

5.2.1 ***Ableist Scenarios.*** These results show that when participants encountered ableist interactions, all conditions increased sensitivity to ableism, with the strongest effects in the Bias-Directed condition (consistent with RQ2). In other words, ableist coaching made participants most likely to recognize ableist interactions as non-standard and harmful.

Q1 (Standard Social Experience). Participants showed larger *decreases* (i.e., judged ableist interactions as less "standard") after Bias-Directed than Self-Directed; Tukey: Bias-Directed < Self-Directed ($p = .040$).

Q2 (Emotional Impact). Participants also showed larger *decreases* (i.e., perceived more negative impact) after Bias-Directed than both Neutral-Directed ($p = .019$) and Self-Directed ($p = .009$).

Taken together, Bias-Directed coaching produced the strongest improvements in recognizing ableist interactions as abnormal and harmful, while Neutral-Directed and Self-Directed conditions also increased recognition but to a lesser degree. (See Figure 4:A.)

5.2.2 ***Neutral Scenarios.*** For neutral interactions, dialogue helped participants affirm these scenarios as acceptable and positive, while the Reading condition moved judgments in the opposite direction, consistent with RQ2.

Q1. Dialogue-based conditions produced *increases* (greater acceptance) whereas Reading produced a *decrease* in ratings of neutral scenarios, indicating that participants in Reading condition became less likely to affirm neutral interactions as positive. Reading was significantly lower than Bias-Directed ($p = .007$), Neutral-Directed ($p < .001$), and Self-Directed ($p = .008$).

Q2. Neutral-Directed ($p = .0029$) and Self-Directed ($p = .003$) conditions both produced more positive changes than Reading. Thus, dialogue engagement improved evaluations of neutral interactions, while the passive Reading condition *reduced* them. (See Figure 4:B.)

5.2.3 ***Combined Scenarios.*** We analyzed the combined set of scenarios to assess the net effect of the interventions on participants' overall judgment tendencies. This allowed us to determine if an intervention caused a generalized shift in sentiment (e.g., becoming more negative overall) or if distinct responses to ableist and neutral scenarios canceled each other out. Looking at the numeric breakdown (Table 1), the changes effectively balanced out in the Neutral-Directed and Self-Directed conditions. For instance, on Q1, the Neutral-Directed group showed increased criticism of ableist scenarios (Δ = −0.21) matched by increased affirmation of neutral ones (Δ = +0.24), yielding a net change close to zero (Δ = +0.02). In contrast, the Bias-Directed and Reading conditions drove more strongly negative overall changes (Bias-Directed: Δ = −0.32 for Q1, −0.41 for Q2; Reading: Δ = −0.28 for Q1, −0.29 for Q2), indicating that increased criticism was not offset by positive affirmation.

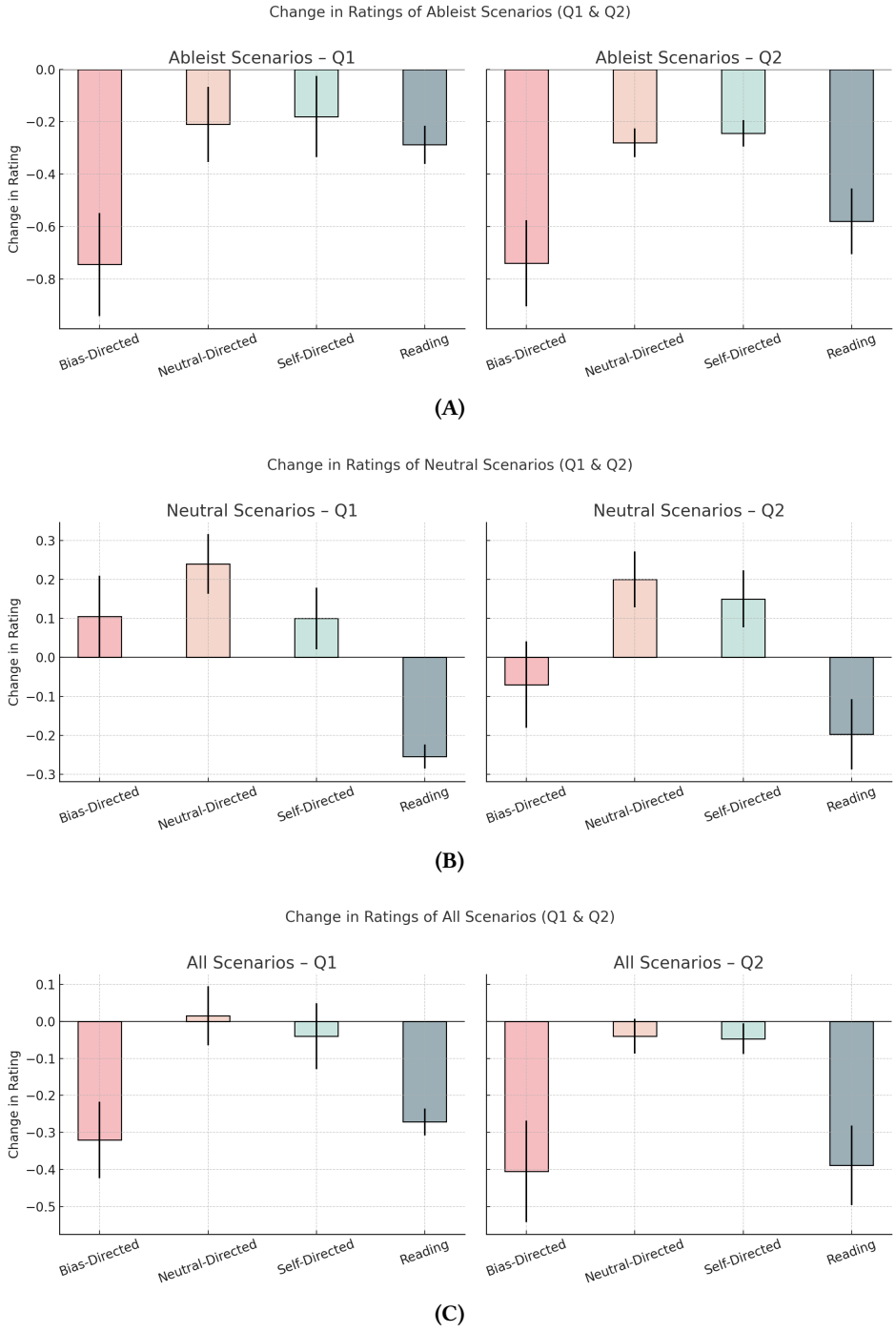


Figure 4: Change in ratings of (A) ableist scenarios, (B) neutral scenarios, and (C) all scenarios combined for Q1 ("standard social experience") and Q2 ("emotional impact"). Bars represent mean change from pre- to post-study across the four conditions (Bias-Directed, Neutral-Directed, Self-Directed, and Reading). Error bars indicate the standard error of the mean (SEM).

Q1. Condition effect: $F(3, 156) = 4.21$, $p = .007$; Bias-Directed produced a more negative shift than Neutral-Directed ($p = .021$).

Q2. Condition effect: $F(3, 156) = 4.86$, $p = .003$; Neutral-Directed ($p = .030$) and Self-Directed ($p = .035$) were more positive than

	Ableist		Neutral		Combined	
Condition	Q1	Q2	Q1	Q2	Q1	Q2
Bias-Directed	−0.75 (.20)	−0.74 (.17)	+0.11 (.11)	−0.07 (.11)	−0.32 (.12)	−0.41 (.12)
Neutral-Directed	−0.21 (.14)	−0.28 (.05)	$+0.24\ (.08)^{a}$	$+0.20\ (.07)^{b}$	$+0.02\ (.08)^{a}$	$-0.04\ (.08)^{b}$
Self-Directed	$-0.18\ (.16)^{c}$	$-0.24\ (.05)^{d}$	$+0.10\ (.08)^{e}$	$+0.15\ (.07)^{f}$	$-0.04\ (.08)^{c}$	$-0.05\ (.08)^{d,f}$
Reading	−0.29 (.07)	−0.58 (.13)	$-0.26\ (.03)^{a,e}$	$-0.20\ (.09)^{b,f}$	$-0.28\ (.08)^{a,c}$	$-0.29\ (.09)^{b,d}$

Note. Means are followed by SEM in parentheses. Superscripts indicate significant Tukey post-hoc comparisons ($p < .05$). (a) Neutral-Directed > Reading (Neutral Q1; Combined Q1). (b) Neutral-Directed > Reading (Neutral Q2; Combined Q2). (c) Self-Directed > Bias-Directed (Ableist Q1; Combined Q1). (d) Self-Directed > Bias-Directed (Ableist Q2; Combined Q2). (e) Self-Directed > Reading (Neutral Q1). (f) Self-Directed > Reading (Neutral Q2; Combined Q2).

Table 1: Mean change scores (Δ = Post − Pre, with SEM in parentheses) by condition. Positive values indicate increases in ratings, negative values decreases. Significant Tukey post-hoc differences are marked with superscripts.

Bias-Directed, and both were also higher than Reading ($p = .042$; $p = .049$).

Overall, while Neutral-Directed and Self-Directed conditions achieved a balance between recognizing harm and affirming inclusion, Bias-Directed and Reading resulted in a net negative trajectory (See Figure 4:C).

5.3 Contrast Scores (Differentiation)

To address RQ3, we analyzed whether interventions improved participants' ability to differentiate between ableist and neutral scenarios. So we computed contrast scores as the difference between neutral and ableist ratings (Neutral − Ableist). Positive values therefore indicate greater differentiation (i.e., neutral scenarios judged more positively and ableist scenarios judged more negatively). Figure 5 visualizes changes in contrast scores from pre- to post-test by condition for Q1 and Q2, while Table 2 reports the corresponding mean contrast-change values (with SEM) and significant Tukey post-hoc differences.

We analyzed changes in contrast scores (Δ = Post − Pre) for both outcome measures (Q1 and Q2). One-way ANOVAs showed significant effects of condition: Q1, $F(3, 156) = 3.91$, $p = .010$; Q2, $F(3, 156) = 3.22$, $p = .024$. Tukey HSD tests were used for post-hoc comparisons.

Q1 (Standard Social Experience). Differentiation improved across all conditions, but participants in the Reading condition showed significantly smaller gains compared to Bias-Directed ($p = .006$). No other pairwise differences reached significance.

Q2 (Emotional Impact). Participants in the Reading condition again showed weaker improvements. Tukey tests revealed that Bias-Directed produced significantly larger gains than Reading ($p = .034$), and Self-Directed also outperformed Reading ($p = .045$). Other differences were not significant. These results indicate that dialogue-based conditions helped participants better distinguish the emotional impact of ableist versus neutral scenarios, while Reading lagged behind.

5.4 Exploratory Moderator Analyses

We next examined whether individual differences moderated intervention effects by adding demographic and experiential variables into regression models with interaction terms. Moderators included disability identity, close family disability connection, comfort with differing perspectives, comfort with strangers, and prior AI chatbot experience.

Close family connection. Participants with a close family member with a disability (n = 28) showed stronger improvements in recognizing the negative emotional impact of ableist scenarios ($p = .034$, Cohen's $d = 0.36$). An interaction with condition also emerged for neutral scenarios in Q2 ($p = .049$, $d = 0.31$), suggesting that family connection amplified intervention effects. A trend also appeared for overall Q2 contrast differentiation ($p = .061$, $d = 0.28$).

Social comfort. Higher comfort with differing perspectives (n = 160) predicted greater positive changes in Q1 ratings of neutral scenarios ($p = .003$, $d = 0.42$). Comfort with strangers (n = 160) predicted stronger improvements in Q1 ratings of neutral scenarios ($p = .049$, $d = 0.32$) and also increased differentiation scores, both Q1 ($p = .044$, $d = 0.30$) and marginally Q2 ($p = .059$, $d = 0.27$).

AI chatbot experience. Prior experience with chatbots (reported by 135 participants, 84%) did not significantly moderate any outcomes ($p > .15$, all $d < 0.20$).

Disability identity. Self-identification as disabled (n = 12) did not moderate effects ($p > .2$, $d < 0.25$).

Setting (Party vs Work Office). Because dialogue scenarios were presented in two settings, we tested whether intervention effects varied by context. ANOVAs with Condition × Setting as factors revealed no significant main effects of setting (all $p > .12$) and no significant interactions (all $p > .14$, $d < 0.20$). Participants responded similarly across the Party (n = 80) and Work Office (n = 80) contexts, indicating that intervention effects were robust across these everyday scenarios.

Overall, these exploratory analyses suggest that intervention effects were broadly consistent across participant groups and settings. However, a close family connection to disability and greater social comfort were associated with stronger improvements, particularly for neutral scenarios and differentiation scores.

6 Qualitative Findings

The qualitative findings help contextualize how participants interpreted the dialogue experience and how coaching direction shaped their judgments, addressing RQ2 (influence of coaching direction on social judgments) and complementing the quantitative analyses for RQ1. Because the post-intervention survey contained three analytically distinct prompts, we conducted three separate thematic analyzes: (1) general reflections on the dialogue (all dialogue-based

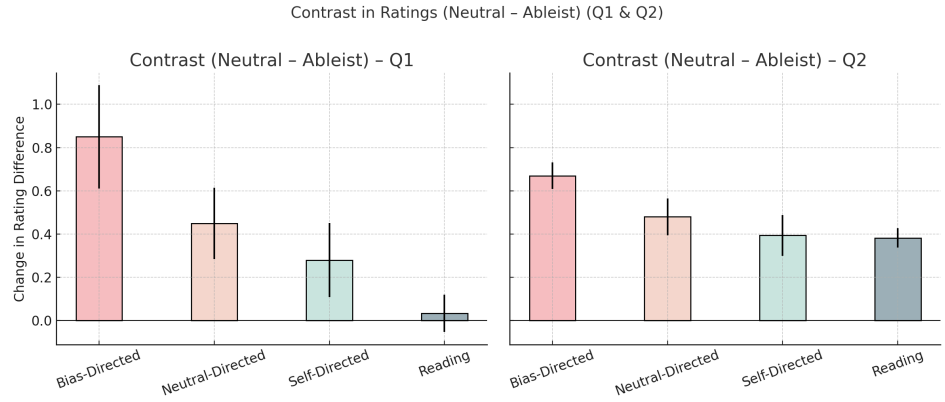


Figure 5: Change in contrast scores (Neutral − Ableist) for Q1 (Standard Social Experience) and Q2 (Emotional Impact). Higher values indicate greater differentiation between neutral and ableist scenarios. Error bars show SEM.

Condition	Q1 (Standard Social Experience)	Q2 (Emotional Impact)
Bias-Directed	+0.85 (.24)	+0.67 (.06)^{a}
Neutral-Directed	+0.45 (.17)	+0.48 (.09)
Self-Directed	+0.28 (.17)	+0.39 (.10)^{b}
Reading	+0.03 (.09)^{a}	+0.38 (.05)^{a,b}

Note. Means are followed by SEM in parentheses. Superscripts indicate significant Tukey post-hoc comparisons ($p < .05$). (a) Bias-Directed > Reading (Q1, Q2). (b) Self-Directed > Reading (Q2).

Table 2: Mean change in contrast scores (Neutral − Ableist; Δ = Post − Pre, with SEM in parentheses). Positive values indicate greater differentiation.

conditions), (2) perceptions of the coach (Bias-Directed and Neutral-Directed only), and (3) unguided experiences (Self-Directed only). Themes below represent higher-level interpretive patterns, with codes and representative quotations illustrating how participants made sense of the interaction.

Table 3 provides an illustrative code hierarchy (themes, example codes, and representative quotes) across the three reflection prompts. Quotes are attributed to participants using anonymized identifiers (e.g., P83).

6.1 General Reflections on Dialogue

Across all three dialogue-based conditions, participants provided broad reflections about the social experience, flow of conversation, and perceived challenges. Reflexive thematic analysis yielded six themes that capture interpretive patterns shared across conditions, with variation when relevant: *Naturalness/Typicality*, *Contextual Adjustment*, *Awkwardness/Unnaturalness*, *Challenge/Difficulty*, *Ease/Comfort*, and *Self-Reflection*. We coded 209 instances across 120 responses ($M = 1.74$ codes/participant).

6.1.1 Theme 1: Naturalness / Typicality. This theme emerged as a highly prominent theme, with 82 (68.3%) participants describing the exchange as "natural," "normal," or "typical," indicating that the simulated dialogue largely aligned with participants' expectations for authentic social interaction. Even when coaching prompts felt inappropriate, the core interaction with the virtual character was perceived as normal. For example, P8 in Bias-Directed condition remarked,

> "It felt like a typical interaction other than my frustration over the suggestions from the coach."

P137 in the Neutral-Directed condition also reflected,

> "It felt like a typical interaction because it had a natural flow."

P152 in Self-Directed condition stated,

> "The conversation felt very typical and free flowing. It made me feel very comfortable to speak and interact with my co-worker. It also felt like speaking to a very old friend I have known my whole life."

Participants routinely emphasized that the conversational turns, emotional tone, and small talk felt consistent with everyday social exchanges.

6.1.2 Theme 2: Contextual Adjustment. 17 participants (14.2% of the dialogue-based sample) reported adapting their tone, conversational style, or level of disclosure based on the scenario's setting (party vs. workplace). These adjustments reflected an awareness of social norms tied to context. As P54 in Bias-Directed condition wrote:

> "I felt they were normal workplace interactions."

Others highlighted a sense of appropriateness linked to context, describing their responses as fitting the social situation. As P94 explained:

Analysis Category	Theme	Example Code	Example Quote
General Reflections	Naturalness / Typicality	Natural interaction	"It felt like a typical interaction because it had a natural flow." (P137 in Neutral-Directed)
	Contextual Adjustment	Adjust formality by context	"I adjusted how I spoke because the setting was a casual social event, not a professional environment. I used a relaxed tone …" (P143 in Self-Directed)
	Awkwardness / Unnaturalness	Awkward disability probing	"It was awkward being told to ask about a handicap. Great conversation otherwise." (P11 in Bias-Directed)
	Challenge / Difficulty	Extra deliberation	"The coach's suggestions added an extra layer of consideration, making it hard to respond naturally." (P82 in Bias-Directed)
	Ease / Comfort	Easy to communicate	"It felt very natural and was easy to communicate." (P156 in Neutral-Directed)
	Self-Reflection	Reflecting on one's language	"It made me think about how I talk to people with disabilities and how weird it is to bring it up …" (P34 in Bias-Directed)
Coach Perceptions	Resistance to Coach	Ignore ableist suggestions	"The coach was offering rude and offensive topics so I ignored them …Every time they wanted to mention Alex's disability." (P10 in Bias-Directed)
	Following the Coach	Coach used to initiate	"Coach's beginning suggestions helped get things rolling. Then, I just felt like the conversation took off on its own." (P110 in Neutral-Directed)
	Perceived Bias / Directive Nature	Forced disability salience	"I was being guided and forced to bring up her physical disability." (P101 in Bias-Directed)
	Perceived Helpfulness / Support	Helpful directions	"The coach provided insightful topics to extend the conversation …it gave me other directions to go." (P66 in Neutral-Directed)
	Naturalness / Unnaturalness	Natural dialogue, odd coach	"The coach's input made the conversation feel much less natural. The coach's guidance seem to have a particular tone or agenda, rather than simply facilitating a natural conversation." (P82 in Bias-Directed)
Unguided Experiences	Authenticity	Matches IRL responses	"Yes, I would actually have a history with the person and this scenario was made up so I had no context." (P39)
	More Detail / Depth	More elaboration IRL	"Yes …I would have been able to go into much greater detail about the job …" (P40)
	Social Awkwardness	More awkward face-to-face	"Yes …I'm just more socially awkward when I'm face-to-face." (P15)
	Contextual Variation	Selective self-disclosure	"Yes …if it was a peer, then I wouldn't have necessarily given up so much personal information." (P132)

Table 3: Illustrative example of the qualitative code hierarchy (Theme → Code → Example Quote) across the three post-intervention reflection prompts.

> "It felt fine polite and appropriate for the setting."

Participants also described adjusting their conversational style to match the tone of two people meeting for the first time in a casual party setting. P156 in Neutral-Directed condition reflected:

> "The conversation felt very natural like how two new people who will have just met will talk, and it felt free flowing and had a feel of two people who had a good conversation with each other "

Across these accounts, contextual fit, not system behavior, was the guiding consideration. Participants oriented themselves to the social expectations of the scenario and modulated their responses to be situationally appropriate, highlighting the ecological realism with which they approached the dialogue.

> "I adjusted how I spoke because the setting was a casual social event, not a professional environment. I used a relaxed tone, shared personal experiences, and focused more on building a friendly connection rather than being formal or task-oriented, which helped the conversation feel more natural and engaging."

This theme highlights participants' sensitivity to situational cues, regardless of the condition.

6.1.3 Theme 3: Awkwardness / Unnaturalness. 11 participants (9.2% of dialogue-based participants) referenced some form of awkwardness or disrupted naturalness-using terms, such as "awkward," "weird," "off," or "forced", highlighting that conversational breakdowns were typically tied to coach-generated suggestions rather than the dialogue itself. As P11 in Bias-Directed condition reflected,

> "It was awkward being told to ask about a handicap. Great conversation otherwise."

Others described a gradual buildup of discomfort when coached prompts pushed the conversation in an unwanted direction. P22 reflected:

> "I was happy with my responses through about 90% of it. I know how to compliment people, and I know when too much is too much. The last 20% of the conversation was too much, and I struggled with wondering if I should continue to do TINY BITS of what the coach wanted, versus nipping the conversation in the bud [...]"

Although most awkwardness arose in coached conditions, a few participants in the Self-Directed condition (3 participants) also described moments of disrupted naturalness. In these cases, discomfort stemmed not from prompts but from situational unfamiliarity-particularly in the work-office setting, where participants lacked background knowledge about the project. As P18 noted:

> "It felt kind of weird, our partner being so specific about us having something to check about the presentation felt odd."

One participant in the party setting also mentioned feeling uncomfortable due to persistent questioning:

> "I felt like they kept asking me questions and didn't take the hint when I tried to change the subject."

Overall, awkwardness was typically attributed either to misaligned coaching suggestions (in coached conditions) or to situational mismatch and uncertainty (in unguided conditions), rather than to the dialogue system as a whole.

6.1.4 Theme 4: Challenge / Difficulty. A subset of 13 participants (10.8% of those in dialogue-based conditions) described the interaction as cognitively or emotionally demanding. These reflections centered on moments when participants felt unsure how to phrase responses, how to manage sensitive topics, or how to reconcile their instincts with the system's structure. In the Bias-Directed condition, challenge often stemmed from prompts that felt misaligned or uncomfortable. As P118 noted:

> "It was challenging because it was condescending."

Others described the difficulty as emerging from the need to integrate-or deliberately resist-coached suggestions. As P82 explained:

> "The coach's suggestions added an extra layer of consideration, making it hard to respond naturally."

These reflections show that challenge arose from a combination of conversational uncertainty, moral discomfort, and cognitive load, revealing how the dialogue prompted deliberate social reasoning.

6.1.5 Theme 5: Ease / Comfort. Others described the conversation as smooth and comfortable, particularly when the interaction progressed organically:

Thirty-four participants (28.3% of the dialogue-based sample) described the interaction as easy, smooth, or comfortable. These reflections emphasized effortlessness, pleasant flow, or the sense that the dialogue unfolded without strain. For example, P84 in the Bias-Directed condition simply noted:

> "They were easy."

Participants in the Neutral-Directed and Self-Directed conditions similarly emphasized ease, often describing the flow as natural. As P156 wrote:

> "It felt very natural and was easy to communicate."

Others highlighted feeling relaxed or at ease during the interaction. P21 reflected:

> "It felt typical and unchallenging. Alex was so easy to talk to so that made it easy."

The sense of comfort also arose from how smoothly the dialogue unfolded. As P70 explained:

> "I felt like it was a typical interaction. It was easy to respond."

Across these accounts, comfort stemmed from the dialogues intuitive structure and socially familiar tone. These reflections show that for many participants, the interaction felt smooth, positive, and uncomplicated, reinforcing the ecological realism of the conversational experience.

6.1.6 Theme 6: Self-Reflection. A smaller but meaningful subset of participants (10 participants; 8.3%) described the dialogue as prompting personal reflection about their own conversational habits, assumptions, or prior interactions with disabled people. These reflections went beyond evaluating the specific exchange and instead focused on broader awareness of how they communicate in real life. For example, P34 noted:

> "It made me think about how I talk to people with disabilities and how weird it is to bring it up as if something is wrong with them or they are less able."

Others reflected specifically on moments when the coach's suggestions clashed with their own values or intentions. P48 explained:

> "I considered whether I would ever respond in the way the coach suggested, which I would hope I never would."

Across these reflections, participants described becoming more aware of their own assumptions or communication tendencies, suggesting that the dialogue—whether coached or unguided—served as a moment of self-evaluation. This theme highlights how AI-mediated conversations can elicit introspective responses that support learning beyond the immediate task.

6.2 Perceptions of the Coach

Participants in the Bias-Directed and Neutral-Directed conditions (n=80) reflected specifically on the one-way coaching suggestions provided during the conversation. Reflexive thematic analysis of these responses yielded five themes: *Resistance to Coach*, *Following the Coach*, *Perceived Bias/Directive Nature*, *Perceived Helpfulness/Support*, and *Naturalness/Unnaturalness*. These themes reveal how participants negotiated the role of the coach, how they interpreted the intent behind suggestions, and how coaching direction shaped their judgments and interactional choices—directly informing RQ2.

6.2.1 Theme 1: Resistance to Coach. A notable group of participants (n=34, 42.5%) rejected or ignored suggestions that felt inappropriate, misaligned, or biased. This resistance was most prevalent in the Bias-Directed condition, where prompts occasionally pushed participants toward condescending or pitying framings. As P10 stated:

> "The coach was offering rude and offensive topics so I ignored them [...] Every time they wanted to mention Alex's disability."

Similarly, P34 emphasized rejecting suggestions that risked making the disabled character uncomfortable:

> "I chose to ignore the coach's suggestions because they were very weird and off-topic. They would make Alex feel out of place or less than."

These reflections show that participants actively resisted biased nudges and asserted moral agency over the conversation.

6.2.2 Theme 2: Following the Coach. Other participants (n=32, 42%) described selectively or strategically following the coach when suggestions aligned with the flow of conversation or helped them formulate a turn. This was more common in the Neutral-Directed condition. As P110 explained:

> "Coach's beginning suggestions helped get things rolling. Then, I just felt like the conversation took off on its own."

Similarly, some participants reported partially incorporating suggestions even when they did not fully agree with them, demonstrating a negotiated relationship with the system's guidance. As P102 explained:

> "I followed them at first but when she was asking about job, coach wanted to keep giving complimentts."

6.2.3 Theme 3: Perceived Bias / Directive Nature. Participants (n=56, 70%), particularly in the Bias-Directed condition frequently commented on the coach as directive, insistent, or overly steering, though the perceived emotional impact differed by condition. In the Bias-Directed condition, prompts were experienced as subtly or overtly ableist. For example, P101 in the Bias-Directed condition mentioned:

> "I was being guided and forced to bring up her physical disability."

In the Neutral-Directed condition, suggestions were not biased but sometimes felt prescriptive or repetitive. As P21 noted:

> "We had moved on to talking about movies and I was coached to go back and ask about Daniel's memories. That ship had sailed. [...] It was directing me in a certain way, and it wasn't natural."

These accounts highlight tensions between participants' conversational instincts and the system's attempt to shape dialogue direction.

6.2.4 Theme 4: Perceived Helpfulness / Support. Participants (n=43, 53.75%), especially in the Neutral-Directed condition often described the coach as supportive, helping them identify appropriate topics, structure their next turn, or maintain rapport. P66 captured this sentiment:

> "The coach provided insightful topics to extend the conversation... it was meant to be helpful because it gave me other directions to go."

For some, the coach served as a conversational scaffold, boosting confidence or smoothing transitions in moments of uncertainty.

6.2.5 Theme 5: Naturalness / Unnaturalness. Finally, participants (n=9, 11.25%) evaluated how the coach affected the naturalness of the interaction. While the dialogue itself was often seen as natural, coaching suggestions occasionally disrupted this realism. As P82 in the Bias-Directed condition explained:

> "The coach's input made the conversation feel much less natural. The coach's guidance seem to have a particular tone or agenda, rather than simply facilitating a natural conversation."

Participants framed unnaturalness as arising not from the system's dialogue model but from the coach's timing, tone, or forced topic shifts.

Across these themes, participants demonstrated active negotiation of the coach's influence, including resisting biased suggestions, selectively adopting inclusive guidance, or managing directive cues within the flow of conversation. These findings illuminate how coaching direction shapes participants' judgments, emotional responses, and conversational strategies, providing key qualitative insight into RQ2.

6.3 Unguided Experiences

Participants in the Self-Directed condition (n = 40) were asked whether they would respond differently in real life. Four themes emerged: *Authenticity*, *More Detail/Depth*, *Social Awkwardness*, and *Contextual Variation*. We coded 38 instances across these responses ($M = 0.95$ codes/participant).

6.3.1 Theme 1: Authenticity. The overwhelming majority emphasized *Authenticity* (35, 87.5%), stating that their responses would not differ from a real-life exchange. Many simply wrote "No," signaling that the unguided dialogue felt genuine and transferable. A few who responded "Yes" explained that the differences reflected the artificiality of the vignette context rather than their conversational style. As P39 clarified,

> "Yes, I would actually have a history with the person and this scenario was made up so I had no context."

6.3.2 Theme 2: More Detail/Depth. Only isolated participants mentioned other divergences. One (2.5%) pointed to *More Detail/Depth*. P40 explained,

> "Yes . . . I would have been able to go into much greater detail about the job, things that hardly anybody outside of the field would really know."

6.3.3 Theme 3: Social Awkwardness. Another (2.5%) highlighted *Social Awkwardness*, suggesting face-to-face interactions might amplify anxiety. P15 remarked:

> "Yes . . . Again, I'm just more socially awkward when I'm face-to-face."

6.3.4 Theme 4: Contextual Variation. A final participant (2.5%) described *Contextual Variation*, noting they might alter disclosure depending on their peer. As P132 put it,

> "Yes . . . if it was a peer, then I wouldn't have necessarily given up so much personal information."

Nearly nine out of ten participants affirmed the authenticity of their responses, while only a few pointed to differences in depth,

social comfort, or contextual disclosure. These reflections suggest that unguided participants largely trusted their own conversational strategies as realistic and transferable to real-world settings.

6.4 Cross-Condition Insights

Taken together, the qualitative findings highlight how conversational framing shaped participants' experiences differently across conditions. While all three groups reported *Naturalness/Typicality*, the source of disruption or support varied by design.

In the **Bias-Directed** condition, participants most often described the dialogue itself as typical, but flagged *Awkwardness/Unnaturalness* and *Perceived Bias/Directive* when prompts pushed them toward ableist framings. For example, P83 explained:

> "It felt like a typical interaction other than my frustration over the suggestions from the coach" [*Naturalness/Typicality*]

Or P11 remarked:

> "It was awkward being told to ask about a handicap" [*Awkwardness/Unnaturalness*]

This illustrate how bias cues undermined authenticity. Resistance was common, with one-third explicitly rejecting prompts [*Resistance to Coach*].

In the **Neutral-Directed** condition, participants frequently emphasized *Ease/Comfort* and *Helpfulness*, describing the prompts as scaffolds that supported conversation. P129 explained:

> "The topics and flow were natural, and the coach's suggestions helped keep the discussion smooth and meaningful" [*Ease/Comfort*]

At the same time, more than half also noted a *Directive* quality, typically framed as repetitive or misaligned rather than offensive. As P21 put it:

> "We had moved on to talking about movies, and I was coached to go back and ask about Daniel's memories. That ship had sailed" [Perceived Bias/Directive].

Thus, neutral guidance was both supportive and steering.

In the **Self-Directed** condition, the dominant theme was *Contextual Adjustment* (58%) paired with *Authenticity* (88%). Participants relied on their own strategies, such as modulating formality and adjusting personal disclosure, to adapt tone to setting. For example, P143 reflected

> "At a party, I kept it more relaxed. At work, I would have been more formal" [*Contextual Adjustment*].

They also reflected on conversational norms. As P34 explained:

> "It made me think about how I talk to people with disabilities and how weird it is to bring it up as if something is wrong with them or they are less able." [*Self-Reflection*]

Overall Patterns. Across conditions, the same themes appeared but with distinct meanings: *Naturalness* was disrupted by biased prompts, reinforced by inclusive ones, and asserted through autonomy in unguided dialogues. *Directive* cues were rejected when biased, tolerated when neutral, and absent in self-directed exchanges. *Self-Reflection* emerged most in self-directed and bias conditions, while *Ease/Comfort* was strongest under neutral guidance. Together, these contrasts illustrate how bias-aware versus biasing framings shaped participants' sense of authenticity, agency, and support.

7 Discussion

In this study, we investigated whether AI-mediated dialogue can shift how people recognize ableist microaggressions and how the *framing* of AI coaching shapes those perceptions. Our findings reveal a complex but compelling picture that extends beyond simple "bias good/bad" dichotomies. The results show that interactive dialogue is a more potent tool for shifting social perception than passive reading, but the nature of that guidance carries significant trade-offs. While inclusive coaching reinforces positive norms, exposure to biased coaching, paradoxically, can sharpen critical awareness of harm.

7.1 Distinct Roles of Q1 and Q2

Our analysis revealed that the interventions shaped Q1 (*Standard Social Experience*) and Q2 (*Emotional Impact*) in different ways, highlighting the importance of treating them as distinct dimensions of bias perception. Q1 captured whether participants regarded an interaction as a "standard social experience," reflecting judgments of social acceptability. In contrast, Q2 captured participants' perception of how the disabled character would feel, reflecting the anticipated "emotional impact" of the exchange.

The divergence between these measures is a key finding. For instance, Bias-Directed coaching sharpened differentiation on Q1 but dampened positive ratings on Q2, even for neutral scenarios. This shows that **recognizing an interaction as socially atypical and anticipating its emotional toll are related but not interchangeable processes**, each responding differently to intervention.

This separation between perceived "standardness" and anticipated emotional harm echoes prior work on microaggressions, which distinguishes between whether behavior is seen as socially acceptable and whether it contributes to cumulative stress and trauma for marginalized groups [53, 73]. It also parallels multidimensional treatments of bias in HCI and NLP, where normative judgments about what *should* be said are separated from inferences about how targets are likely to feel or be affected by what is said [65, 77]. Our results extend this line of work by showing that AI-mediated interventions can differentially move these two dimensions: a system may successfully help people flag an interaction as atypical without equally improving their sensitivity to its emotional toll, or vice versa.

7.2 The Double-Edged Nature of Biased Nudges

Perhaps our most striking finding is the paradoxical effect of the **Bias-Directed** condition. Participants exposed to subtly ableist prompts showed the strongest improvements in **differentiating** ableist from neutral interactions (Figure 5). They became significantly more likely to judge ableist scenarios as non-standard and emotionally harmful.

The qualitative data explains this phenomenon as a product of **active resistance**. The prompts likely created a **reflective dissonance** or **"critical friction"**: a moment where participants had to consciously reject an inappropriate suggestion and articulate an

alternative. This reaction mirrors findings in technology-mediated nudging, where overt steering can trigger reactance rather than compliance [12]. Yet, in the context of human–AI interaction, this resistance proved productive; it aligns with research on critical-reflective collaboration [30] and emerging work on user-driven value alignment [23], where disagreement with an AI agent serves as a mechanism for clarifying one's own moral boundaries. This process of identifying, evaluating, and countering a biased framing appears to be a powerful, albeit unintended, learning mechanism.

However, this heightened vigilance came at a cost. The Bias-Directed group also developed a negative halo effect, rating even **neutral scenarios** more negatively on emotional impact. This suggests that exposure to biased exemplars, even when rejected, may cast a shadow over subsequent social judgments. An intervention that maximizes sensitivity to harm might inadvertently erode the recognition of safety and inclusion - a critical trade-off for designers. Taken together, this pattern both reinforces prior HCI observations that overt nudges can provoke resistance rather than compliance [12], and extends them by showing that, in the context of ableist microaggressions, such resistance can actually sharpen people's ability to differentiate harmful from neutral interactions. To our knowledge, this is the first empirical evidence in accessibility contexts that resisting biased AI suggestions can sharpen microaggression recognition, rather than simply leading to disengagement.

7.3 Mechanisms of Change: Inclusive Scaffolding and Self-Construction

In contrast, the **Neutral-Directed** and **Self-Directed** conditions fostered a more balanced recognition by reinforcing inclusive norms. Neutral-directed participants perceived the coach's suggestions as helpful **"scaffolding"** that provided positive, actionable templates for respectful interaction. Both groups improved in affirming neutral interactions, observing or producing inclusive models of engagement.

Notably, however, these conditions did less to sharpen the **contrast** between neutral and ableist interactions compared to the biased condition. This suggests that **affirming the positive fosters recognition of inclusion but may not, by itself, clarify the boundaries of harm.** Furthermore, even inclusive scaffolding was sometimes seen as "directive," highlighting the thin line between helpful guidance and unwelcome steering in supportive HCI systems.

These dynamics suggest a complementary role for AI-mediated dialogue alongside disability-led education rather than as a replacement for it. Disability training and education work has emphasized the value of storytelling, role-play, and simulation exercises led by disabled people in building empathy and challenging ableist assumptions [29, 37, 64]. However, such activities are labor-intensive and often difficult to scale. Our findings point to one way interactive systems could extend this work: disability-led curricula can establish norms, share lived experience, and articulate why particular patterns of interaction are harmful, while AI-mediated dialogues provide low-stakes opportunities for repeated practice, feedback, and reflection between or after those sessions. This aligns with HCI research that calls for incorporating social factors and lived experience into accessible design processes [69], and highlights the importance of developing any such coaching systems in partnership with disabled communities. Our findings, therefore, complement and extend this disability-led training literature by providing empirical evidence for a concrete, AI-mediated practice layer that can operationalize these curricula between or after live sessions.

7.4 The Power of Active Engagement Over Passive Consumption

Across nearly all measures, all three dialogue-based conditions outperformed the **Reading** condition. This shows the limits of passive exposure; simply presenting information about ableism did not reliably shift social judgments and, in our case, sometimes led to declines in recognition. Participants in the Reading condition became less likely to affirm neutral interactions as positive (Section 5), suggesting that reading about microaggressions without opportunities to practice responses may foster a more generally negative or skeptical stance toward social encounters without equivalently sharpening differentiation between harm and safety. This pattern resembles a kind of "backfire effect," mirroring findings in diversity training research, where static instruction without behavioral practice can induce defensiveness or disconnects between abstract knowledge and situated application [13, 26, 39, 40]. In contrast, the dialogue conditions required participants to *actively generate* responses in context. This act of construction was a far more effective mechanism for learning, consistent with theories of active learning that link more constructive and interactive forms of cognitive engagement to deeper learning outcomes [14, 28]. Our findings echo calls in healthcare and education to move from knowledge-transfer interventions toward feedback-rich, interactive systems that help practitioners reflect on their language in situ [6, 51]. Our results reinforce these broader findings while extending them into the underexplored domain of disability microaggressions, showing experimentally that brief, AI-mediated conversational practice outperforms a reading-only module on short-term recognition and differentiation. The **Self-Directed** group's emphasis on "authenticity" and "contextual adjustment" further shows that unguided, reflective practice can be a valuable learning experience in itself, aligning with HCI work on critical self-reflection and deliberation in human-AI collaboration [30, 48]. In the underexplored domain of disability microaggressions, our study therefore adds experimental support to these broader claims about the advantages of active, feedback-rich engagement over passive instruction.

7.5 Connecting to Broader HCI and Computational Social Science

Our work contributes to several ongoing conversations in HCI and related fields:

- **Framing and Nudging in HCI:** This study extends classic work on framing effects [78] and technology-mediated nudging [12, 16] into the dynamic context of AI-mediated dialogue. We demonstrate that users are not passive recipients of AI "nudges": biased coaching was frequently resisted, and inclusive coaching was selectively adopted as scaffolding. This complicates design assumptions that interface defaults

or suggestions will be quietly accepted, and aligns with recent HCI work that emphasizes the negotiated character of human-AI influence [43, 48].

- **Human-AI Interaction as Social Simulation:** We present a novel paradigm for using LLMs not just as tools but as **social actors** within controlled experiments. This methodology complements interactive evaluation platforms such as SOTOPIA [82] by focusing on human outcomes rather than model scores: instead of asking whether an agent exhibits social intelligence, we ask how its prompts and framings reshape human judgments in situ. In doing so, we offer a methodological contribution in the form of a reusable experimental apparatus that other researchers can adapt to study different forms of social bias in interaction, alongside empirical evidence about AI-mediated bias interventions.
- **Microaggression Research:** By demonstrating that normative and affective judgments are distinct skills that can move differently across conditions, our work complements psychological accounts of microaggressions as both everyday norm violations and sources of cumulative harm [53, 73]. It also extends recent vignette-based HCI work on ableism [77] by embedding scenarios in interactive dialogue and showing how different coaching framings modulate recognition trajectories over time.
- **Complementing Disability-Led Education:** We do not position AI-mediated dialogue as a replacement for training led by disabled people, which provides essential lived experience, context, and accountability. Instead, our findings suggest a complementary role: disability-led workshops and courses can establish norms and share lived experiences, while AI agents offer a scalable "sandbox" for practice, where learners can rehearse language, make mistakes, and encounter critical friction in a low-stakes environment between or after those sessions. In terms of HCI contribution types [80], this work provides empirical evidence about how different coaching directions shape bias recognition, a methodological contribution in the form of an AI-mediated dialogue platform and vignette corpus for studying ableism in situ, and design-oriented implications for socially aware, anti-ableist AI systems.

This study demonstrates that AI-driven dialogue can be a potent medium for social learning. The findings indicate a more intricate reality, one where the path to recognizing bias is not singular. While inclusive models build positive schemas, the active struggle against biased models can forge a sharper, more critical understanding of social harm.

8 Design Implications for Socially-Aware AI

Our findings carry direct implications for the design of AI systems that mediate or influence human social interaction. As conversational AI becomes more prevalent, understanding how subtle prompts shape social judgment is crucial for creating responsible and effective technology.

- **Guidance is Not Neutral.** The starkly different outcomes between our conditions show that there is no "neutral" way to design a social AI. System defaults, suggested phrasings, and conversational patterns inevitably frame interactions and influence users. Designers should, therefore, treat coaching prompts as value-laden design decisions and make those norms explicit rather than presenting them as purely objective or generic.
- **Prefer Scaffolding Over Prescription.** The **Neutral-Directed** condition suggests that users benefit most when AI provides options and examples that they can adapt, rather than a single "correct" response. In practice, this means generating a small set of alternative suggestions or question stems, and allowing users to ignore or modify them, rather than forcing the conversation down one path.
- **Balance Sensitivity with Positivity.** The **Bias-Directed** condition increased sensitivity to harm but also produced a more globally negative evaluation pattern. To avoid this, systems intended for training should not only highlight problematic phrasings or microaggressions but also immediately offer constructive, bias-aware alternatives. This pairing can support both vigilance and affirmation of safe, respectful interactions.
- **Consider "Critical Friction" Carefully and with consent.** Our results suggest that resisting biased suggestions can prompt useful reflection, but biased prompts also carry risk. If designers introduce intentionally problematic examples (e.g., in a training mode), they should clearly label this mode, obtain consent, and provide context about why the example is inappropriate. Outside of such controlled settings, systems should avoid generating biased framings in the first place.
- **Integrate With Disability-Led Education.** Finally, AI-mediated dialogue should be deployed as a complement to, not a replacement for, disability-led workshops and curricula. Our findings suggest that such tools are well-suited as scalable "practice spaces" where learners can rehearse language and reflect on their choices between live sessions, especially when co-designed with disabled people.

9 Limitations and Future Work

Our study's limitations provide important context for interpreting these findings and open several promising directions for future research.

First, our text-based dialogue, while enabling experimental control, lacks the multimodal richness of face-to-face interaction. Future work could explore these dynamics in more immersive environments, such as using **virtual reality or voice-based agents**, to incorporate non-verbal cues like tone and body language.

Second, our intervention was a single, brief exposure. While we demonstrated immediate shifts in judgment, we cannot assess the **durability** of these changes. Longitudinal studies are needed to track whether these effects persist over time and, crucially, if they **transfer to actual interpersonal behavior**.

Third, the artificial nature of the LLM-driven interaction may limit transferability, as modern LLMs still demonstrate significant limits in their social intelligence and theory of mind [66]. Our platform is therefore best understood as an experimental apparatus that approximates everyday conversation rather than a full reproduction of real-world, multimodal interactions. Future studies could

use this platform in a "Wizard of Oz" setup to compare the effects of AI-guidance versus human-guidance in mediating conversations.

Fourth, our sample demographics reflect specific limitations regarding generalizability. The majority of our participants (71%) were aged 18–44, and the study was conducted exclusively in English. As perceptions of ableism and conversational norms often vary across generations and linguistic cultures, future work should deliberately recruit older adults and non-native English speakers to investigate how these factors shape the recognition of microaggressions and to test whether dialogue-based interventions and coaching directions produce similar effects in other contexts.

Finally, the experimental paradigm itself is highly generalizable and can serve as a powerful tool for studying other forms of social bias beyond ableism. The core components, a simulated dialogue with an AI character representing a member of a marginalized group and an AI coach modeling biased or inclusive framings, can be readily adapted. For instance, researchers could study:

- **Racial Microaggressions:** By simulating a job interview or a casual social encounter where biased nudges might suggest comments like, "But where are you *really* from?" or play into stereotypes about competence or background.
- **Gender Bias:** By simulating a professional meeting where a biased coach suggests a participant interrupt a female colleague ("hepeating") or question her technical expertise, allowing for the study of how people navigate and resist sexist framings in real-time.
- **Ageism:** By modeling workplace interactions where a younger manager receives biased prompts suggesting they patronize an older employee's technical skills or make assumptions about their career ambitions.

This framework serves as an important contribution to Computational Social Science by providing a novel methodology to empirically study bias. It creates a testbed for evaluating the societal impact and persuasive power of large language models, operationalizing key questions in AI Ethics about algorithmic influence and responsible AI design.

Conclusion

This study demonstrates that the design of AI-mediated dialogue not only reflects social norms, but actively shapes them. Our findings reveal that while inclusive AI coaching provides helpful scaffolding for positive interaction, it was active resistance to biased nudges that most sharply honed the participants' ability to differentiate harm. This presents a fundamental tension for intervention design: the path to increased critical awareness may not be the same as the path to reinforcing positive social engagement.

Our primary contribution is twofold: an empirical demonstration of the nuanced, often contradictory effects of AI-driven social nudging and a generalizable experimental framework for studying these dynamics. This framework offers a robust methodology for HCI and Computational Social Science to move beyond static surveys and investigate bias recognition in a more ecologically valid, interactive context.

As we move into an era where conversational AI will increasingly mediate our social lives, our responsibility as designers extends beyond creating systems that are merely efficient or engaging. We must design systems that foster critical thought, respect human agency, and are intentionally built to support more inclusive and equitable human relationships. The central challenge is not simply to build AI that understands social norms but to build AI that helps us, as humans, understand them and each other better.

Acknowledgments

We sincerely thank Christine Mendoza, Matildelis Medina, and Carrie Danny for generously sharing their time and expertise to strengthen the vignette materials. We also thank our participants for their time and contributions. This work was supported in part by Carnegie Mellon University's Presidential Postdoctoral Fellowship (awarded to the first author).

References

[1] Fareeaa Abdoola, Penelope S Flack, and Saira B Karrim. 2017. Facilitating pragmatic skills through role-play in learners with language learning disability. *South African Journal of Communication Disorders* 64, 1 (2017), 1–12.

[2] Herman Aguinis and Kyle J Bradley. 2014. Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational research methods* 17, 4 (2014), 351–371.

[3] Nientara Anderson, Elle Lett, Emmanuella Ngozi Asabor, Amanda Lynn Hernandez, Max Jordan Nguemeni Tiako, Christen Johnson, Roberto E Montenegro, Tara M Rizzo, Darin Latimore, Marcella Nunez-Smith, et al. 2022. The association of microaggressions with depressive symptoms and institutional satisfaction among a national cohort of medical students. *Journal of general internal medicine* 37, 2 (2022), 298–307.

[4] Paritosh Bahirat, Martijn Willemsen, Yangyang He, Qizhang Sun, and Bart Knijnenburg. 2021. Overlooking context: How do defaults and framing reduce deliberation in smart home privacy decision-making?. In *Proceedings of the 2021 chi conference on human factors in computing systems*. 1–18.

[5] Sarah Barber, PC Gronholm, S Ahuja, N Rüsch, and G Thornicroft. 2020. Microaggressions towards people affected by mental health problems: a scoping review. *Epidemiology and psychiatric sciences* 29 (2020), e82.

[6] Emily Bascom, Reggie Casanova-Perez, Kelly Tobar, Manas Satish Bedmutha, Harshini Ramaswamy, Wanda Pratt, Janice Sabin, Brian Wood, Nadir Weibel, and Andrea Hartzler. 2024. Designing communication feedback systems to reduce healthcare providers' implicit biases in patient encounters. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–12.

[7] Elizabeth Bembridge, Gary Crowfoot, Rhonda Wilson, and Julie Reis. 2025. Digital reflective practice among new graduate nurses–A scoping review. *Nurse Education Today* (2025), 106688.

[8] Guy A Boysen. 2012. Teacher and student perceptions of microaggressions in college classrooms. *College Teaching* 60, 3 (2012), 122–129.

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[10] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[11] Fiona Campbell. 2009. *Contours of ableism: The production of disability and abledness*. Springer.

[12] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[13] Edward H Chang, Katherine L Milkman, Dena M Gromet, Robert W Rebele, Cade Massey, Angela L Duckworth, and Adam M Grant. 2019. The mixed effects of online diversity training. *Proceedings of the National Academy of Sciences* 116, 16 (2019), 7778–7783.

[14] Michelene TH Chi and Ruth Wylie. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist* 49, 4 (2014), 219–243.

[15] Gerald L Clore and Katharine M Jeffery. 1972. Emotional role playing, attitude change, and attraction toward a disabled person. *Journal of personality and social psychology* 23, 1 (1972), 105.

[16] Andy Cockburn, Blaine Lewis, Philip Quinn, and Carl Gutwin. 2020. Framing effects influence interface feature decisions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11.

[17] Kristin J Conover, Tania Israel, and Karen Nylund-Gibson. 2017. Development and validation of the Ableist Microaggressions Scale. *The Counseling Psychologist* 45, 4 (2017), 570–599.

[18] Caitríona Cox, Thea Hatfield, Jordan Moxey, and Zoë Fritz. 2023. Creating and administering video vignettes for a study examining the communication

of diagnostic uncertainty: methodological insights to improve accessibility for researchers and participants. *BMC medical research methodology* 23, 1 (2023), 296.

[19] Melissa D Deroche, Lee Za Ong, and Jennifer M Cook. 2024. Ableist Microaggressions, Disability Characteristics, and Nondominant Identities. *Professional Counselor* 13, 4 (2024), 404–417.

[20] Vimal Desai, Antonio Hernandez Conte, Vu T Nguyen, Philip Shin, Neha T Sudol, Janet Hobbs, and Chunyuan Qiu. 2023. Veiled harm: impacts of microaggressions on psychological safety and physician burnout. *The Permanente Journal* 27, 2 (2023), 169.

[21] Hamza El Alaoui and Violetta Cavalli-Sforza. 2024. DarijaGenie: Learning Moroccan Arabic Through a Multimodal Chatbot. In *International Conference on Arabic Language Processing*. Springer, 74–89.

[22] Hamza El Alaoui, Atieh Taheri, Yi-Hao Peng, and Jeffrey P Bigham. 2025. StepWrite: Adaptive Planning for Speech-Driven Text Generation. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology* (Busan, Republic of Korea) *(UIST '25)*. ACM, New York, NY, USA, Article 86, 30 pages. doi:10.1145/3746059.3747610

[23] Xianzhe Fan, Qing Xiao, Xuhui Zhou, Jiaxin Pei, Maarten Sap, Zhicong Lu, and Hong Shen. 2025. User-Driven Value Alignment: Understanding Users' Perceptions and Strategies for Addressing Biased and Discriminatory Statements in AI Companions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.

[24] Jack Faulkner. 2025. The Impact of Race on Mental Health Burden and Burnout in Medical School and Residency: A Literature Review. *Undergraduate Journal of Public Health* 9 (2025).

[25] Estivana Felaza, Ardi Findyartini, Rita Mustika, Jenny Bashiruddin, Joedo Prihartono, Lucia RM Royanto, and Subha Ramani. 2025. Promoting medical students' reflection through feedback interaction: The RAISE model. *Asia Pacific Scholar* 10, 1 (2025), 27–35.

[26] Patrick S Forscher, Calvin K Lai, Jordan R Axt, Charles R Ebersole, Michelle Herman, Patricia G Devine, and Brian A Nosek. 2019. A meta-analysis of procedures to change implicit measures. *Journal of personality and social psychology* 117, 3 (2019), 522.

[27] Anderson J Franklin. 2008. *From brotherhood to manhood: How Black men rescue their relationships and dreams from the invisibility syndrome.* Turner Publishing Company.

[28] Scott Freeman, Sarah L Eddy, Miles McDonough, Michelle K Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. 2014. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the national academy of sciences* 111, 23 (2014), 8410–8415.

[29] Sally French. 1992. Simulation exercises in disability awareness training: A critique. *Disability, Handicap & Society* 7, 3 (1992), 257–266.

[30] Katrin Glinka and Claudia Müller-Birn. 2023. Critical-reflective human-AI collaboration: Exploring computational tools for art historical image retrieval. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–33.

[31] Nicola Grove. 2015. Finding the sparkle: Storytelling in the lives of people with learning disabilities. *Tizard Learning Disability Review* 20, 1 (2015), 29–36.

[32] Miriam Havin, Timna Wharton Kleinman, Moran Koren, Yaniv Dover, and Ariel Goldstein. 2025. Can (A) I Change Your Mind? *arXiv preprint arXiv:2503.01844* (2025).

[33] Thomas Hehir. 2002. Eliminating ableism in education. *Harvard educational review* 72, 1 (2002), 1–33.

[34] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature* 633, 8028 (2024), 147–154.

[35] Alexandra Hollo, Carla B Brigandi, Casey M Jelsema, and Mingming Shi. 2021. Effects of a disability simulation activity on the achievement, efficacy, attitudes, and interests of preservice professionals. *Journal of Visual Impairment & Blindness* 115, 5 (2021), 414–425.

[36] Rhidian Hughes. 1998. Considering the vignette technique and its application to a study of drug injecting and HIV risk and safer behaviour. *Sociology of health & illness* 20, 3 (1998), 381–400.

[37] Michael Ioerger, Reed M Flanders, Jeremy R French-Lawyer, and Margaret A Turk. 2019. Interventions to teach medical students about disability: a systematic search and review. *American Journal of Physical Medicine & Rehabilitation* 98, 7 (2019), 577–599.

[38] Yewon Jin, SeonYul Lee, SeoHyeong Kim, Jiyeon Seo, Kyuha Jung, Hajin Lim, and Joonhwan Lee. 2023. Divrsity: Design and development of group role-play vr platform for disability awareness education. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 161–174.

[39] Calvin K Lai, Maddalena Marini, Steven A Lehr, Carlo Cerruti, Jiyun-Elizabeth L Shin, Jennifer A Joy-Gaba, Arnold K Ho, Bethany A Teachman, Sean P Wojcik, Spassena P Koleva, et al. 2014. Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General* 143, 4 (2014), 1765.

[40] Calvin K Lai, Allison L Skinner, Erin Cooley, Sohad Murrar, Markus Brauer, Thierry Devos, Jimmy Calanchini, Y Jenny Xiao, Christina Pedram, Christopher K Marshburn, et al. 2016. Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General* 145, 8 (2016), 1001.

[41] Thomas C Leonard. 2008. Richard H. Thaler, Cass R. Sunstein, Nudge: Improving Decisions about Health, Wealth, and Happiness: Yale University Press, New Haven, CT, 2008, 293 pp, $26.00.

[42] Rong Li, Ashwini Kamaraj, Jing Ma, and Sarah Ebling. 2024. Decoding ableism in large language models: An intersectional approach. In *Proceedings of the Third Workshop on NLP for Positive Impact*. 232–249.

[43] Yuxuan Li, Hirokazu Shirado, and Sauvik Das. 2025. Actions speak louder than words: Agent decisions reveal implicit biases in language models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 3303–3325.

[44] Gili Lior, Liron Nacchace, and Gabriel Stanovsky. 2025. WildFrame: Comparing Framing in Humans and LLMs on Naturally Occurring Texts. *arXiv preprint arXiv:2502.17091* (2025).

[45] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[46] Jiaxu Lou and Yifan Sun. 2024. Anchoring bias in large language models: An experimental study. *arXiv preprint arXiv:2412.06593* (2024).

[47] P Priscilla Lui and Lucia Quezada. 2019. Associations between microaggression and adjustment outcomes: A meta-analytic and narrative review. *Psychological bulletin* 145, 1 (2019), 45.

[48] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2025. Towards human-ai deliberation: Design and evaluation of llm-empowered deliberative ai for ai-assisted decision-making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23.

[49] Tracy MacIntosh, Michelle Hernandez, and Anuja S Mehta. 2022. Identifying, addressing, and eliminating microaggressions in healthcare. *HCA Healthcare Journal of Medicine* 3, 3 (2022), 189.

[50] Catherine L Mah, Emily Taylor, Sylvia Hoang, and Brian Cook. 2014. Using vignettes to tap into moral reasoning in public health policy: Practical advice and design principles from a study on food advertising to children. *American journal of public health* 104, 10 (2014), 1826–1832.

[51] Zainab Maqsood, Madiha Sajjad, and Raheela Yasmin. 2025. Effect of feedback-integrated reflection, on deep learning of undergraduate medical students in a clinical setting. *BMC Medical Education* 25, 1 (2025), 66.

[52] Whitney M Morean. 2022. *Ableist Microaggressions and Well-being: Investigating the Moderating Effect of Coping Strategies.* Ph. D. Dissertation. Seattle Pacific University.

[53] KL Nadal. 2018. Microaggressions and traumatic stress. *American Psychological Association. https://doi. org/10.1 0* 37 (2018), 0000073–000.

[54] Prisca Mary Oluoch, Paul Amollo Odundo, and John Kamau Mwangi. 2020. Context of Vignettes and Ethical Sensitivity in Decision-Making among Undergraduate Business Studies Learners at University of Nairobi, Kenya. *Journal of Education and Learning* 9, 4 (2020), 151–169.

[55] Alex Ouyang. 2025. Study reveals AI chatbots can detect race, but racial bias reduces response empathy. MIT Electrical Engineering & Computer Science News. https://www.eecs.mit.edu/study-reveals-ai-chatbots-can-detect-race-but-racial-bias-reduces-response-empathy/ Abdul Latif Jameel Clinic for Machine Learning in Health.

[56] Mahika Phutane, Hayoung Jung, Matthew Kim, Tanushree Mitra, and Aditya Vashistha. 2025. ABLEIST: Intersectional Disability Bias in LLM-Generated Hiring Scenarios. *arXiv preprint arXiv:2510.10998* (2025).

[57] Mahika Phutane, Ananya Seelam, and Aditya Vashistha. 2025. "Cold, Calculated, and Condescending": How AI Identifies and Explains Ableism Compared to Disabled People. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 1927–1941.

[58] Mahika Phutane and Aditya Vashistha. 2025. Disability Across Cultures: A Human-Centered Audit of Ableism in Western and Indic LLMs. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 8. 2000–2014.

[59] Charles A Pierce, Herman Aguinis, and Susan KR Adams. 2000. Effects of a dissolved workplace romance and rater characteristics on responses to a sexual harassment accusation. *Academy of Management Journal* 43, 5 (2000), 869–880.

[60] Gisela Reyes-Cruz, Joel Fischer, and Stuart Reeves. 2022. Supporting Awareness of Visual Impairments and Accessibility Reflections through Video Demos and Design Cards. In *Nordic Human-Computer Interaction Conference*. 1–15.

[61] Naba Rizvi, Harper Strickland, Saleha Ahmedi, Aekta Kallepalli, Isha Khirwadkar, William Wu, Imani NS Munyaka, and Nedjma Ousidhoum. 2025. Beyond Keywords: Evaluating Large Language Model Classification of Nuanced Ableism. *arXiv preprint arXiv:2505.20500* (2025).

[62] Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2025. On the conversational persuasiveness of GPT-4. *Nature Human Behaviour* (2025), 1–9.

[63] Matthew Sample, Sebastian Sattler, Wren Boehlen, and Eric Racine. 2023. Brain-computer interfaces, disability, and the stigma of refusal: A factorial vignette study. *Public Understanding of Science* 32, 4 (2023), 522–542.

[64] Eda Ozkara San, Katherine A Marx, Beth Latimer, Gina Robertiello, and Natalya Pasklinsky. 2022. The use of clinical simulation to address disability content inclusion in nursing education: a review of the literature. *Clinical Simulation in Nursing* 71 (2022), 112–127.

[65] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891* (2019).

[66] Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2023. Neural Theory-of-Mind? On the limits of social intelligence in large LMs. arXiv. *arXiv preprint arXiv:2210.13312* (2023).

[67] Anniek Schlette, Tobias H Stark, and Anouk Smeekes. 2025. Social signals of belonging: How the perceived ethnic-national background of friends affects ascriptions of belonging given to descendants of migrants. *British Journal of Social Psychology* 64, 3 (2025), e12898.

[68] Dylan G Serpas, Daniel A Ignacio, Kemesha Gabbidon, and Barbara J Cherry. 2024. Ableist microaggressions and psychological distress among adults with disabilities: The role of disability visibility. *Rehabilitation Psychology* (2024).

[69] Kristen Shinohara, Jacob O Wobbrock, and Wanda Pratt. 2018. Incorporating social factors in accessible design. In *Proceedings of the 20th international ACM SIGACCESS conference on computers and accessibility*. 149–160.

[70] Daniel Solorzano, Miguel Ceja, and Tara Yosso. 2000. Critical race theory, racial microaggressions, and campus racial climate: The experiences of African American college students. *Journal of Negro education* (2000), 60–73.

[71] Derald Wing Sue. 2004. Whiteness and ethnocentric monoculturalism: making the" invisible" visible. *American psychologist* 59, 8 (2004), 761.

[72] Derald Wing Sue, Sarah Alsaidi, Michael N Awad, Elizabeth Glaeser, Cassandra Z Calle, and Narolyn Mendez. 2019. Disarming racial microaggressions: Microintervention strategies for targets, White allies, and bystanders. *American Psychologist* 74, 1 (2019), 128.

[73] Derald Wing Sue, Christina M Capodilupo, Gina C Torino, Jennifer M Bucceri, Aisha Holder, Kevin L Nadal, and Marta Esquilin. 2007. Racial microaggressions in everyday life: implications for clinical practice. *American psychologist* 62, 4 (2007), 271.

[74] Derald Wing Sue and Lisa Spanierman. 2020. *Microaggressions in everyday life*. John Wiley & Sons.

[75] Atieh Taheri, Purav Bhardwaj, Arthur Caetano, Alice Zhong, and Misha Sra. 2023. Virtual buddy: Redefining conversational ai interactions for individuals with hand motor disabilities. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–3.

[76] Jasper Timm, Chetan Talele, and Jacob Haimes. 2025. Tailored truths: Optimizing llm persuasion with personalization and fabricated statistics. *arXiv preprint arXiv:2501.17273* (2025).

[77] Shane Timmons, Frances McGinnity, and Eamonn Carroll. 2024. Ableism differs by disability, gender and social context: Evidence from vignette experiments. *British Journal of Social Psychology* 63, 2 (2024), 637–657.

[78] Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *science* 211, 4481 (1981), 453–458.

[79] Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2023. Automated ableism: An exploration of explicit disability biases in sentiment and toxicity analysis models. *arXiv preprint arXiv:2307.09209* (2023).

[80] Jacob O Wobbrock and Julie A Kientz. 2016. Research contributions in human-computer interaction. *interactions* 23, 3 (2016), 38–44.

[81] Hannah Young, Maggi Fenwick, Loretto Lambe, and James Hogg. 2011. Multi-sensory storytelling as an aid to assisting people with profound intellectual disabilities to cope with sensitive issues: A multiple research methods analysis of engagement and outcomes. *European Journal of Special Needs Education* 26, 2 (2011), 127–142.

[82] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. In *The Twelfth International Conference on Learning Representations (ICLR)*. doi:10.48550/arXiv.2310.11667 Preprint available as arXiv:2310.11667.