Breaking Barriers: A Paradigm Shift in Technology Accessibility for Individuals with Physical Disabilities

Kshitij Mishra [*1], **Manisha Burja** [*1], **Asif Ekbal** [2]

[1] Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
[2] School of AI and Data Science, Indian Institute of Technology Jodhpur, India
mishra.kshitij07@gmail.com, manishaboorja@gmail.com, asif.ekbal@gmail.com

Abstract

Individuals living with disabilities often face challenges in their daily lives, from managing physical tasks to coping with emotional needs. It is imperative to provide them with personalized, courteous, and empathetic support that can address their unique needs. To bridge this gap, we propose an Empathetic Disability Support System (EDiSS), designed to offer personalized support tailored with correct politeness and empathetic strategies as per individual users' OCEAN traits, gender, and age. To train EDiSS, first, a specialized personalized disability support dialogue dataset (PDCARE) is created encompassing a wide spectrum of disabilities, such as *Spinal Cord Injuries*, *Neurological Disorders*, *Orthopedic Disabilities*, etc, and support areas like *Physical Therapy Exercises*, *Pain Management*, *Emotional Support*, etc. EDiSS employs a reinforcement learning-based dialogue model with a novel reward function. It adapts its tone and content based on the user's persona, gender, and age to provide respectful and empathetic assistance across various aspects of daily living. Our experiments and evaluation demonstrate the effectiveness of EDiSS in improving the quality of life of individuals with disabilities, marking a significant advancement in leveraging technology to provide much-needed support and assistance in their daily challenges.

Code and Dataset —
https://github.com/Mishrakshitij/EDiSS.git

Introduction

In recent years, there has been a growing recognition of the challenges faced by individuals living with disabilities, encompassing not only physical limitations but also emotional and societal barriers. According to the World Health Organization (WHO), over 1 billion people worldwide experience some form of disability, making up approximately 15% of the global population (Organization 2021). Despite progress in accessibility and inclusion initiatives, individuals with disabilities continue to encounter significant hurdles in various aspects of their daily lives, ranging from physical tasks to social interactions and emotional well-being.

The United Nations Sustainable Development Goals (SDGs) underscore the importance of inclusivity and accessibility, particularly through Goal 10, which aims to reduce inequality within and among countries, and Goal 3, which promotes health and well-being for all (United Nations 2024b). Further, Goal 4 of SDGs emphasizes the importance of ensuring inclusive and equitable quality education and promoting lifelong learning opportunities for all, including persons with disabilities (United Nations 2024b). Additionally, the Leave No One Behind Principle (LNOB), a core tenet of the SDGs, emphasizes the need to ensure that development efforts reach and benefit all segments of society, including those with disabilities (United Nations 2024a).

To address the multifaceted needs of individuals with disabilities, there is a pressing need for innovative solutions that can provide personalized, courteous, and empathetic support. Conversational AI systems can offer a promising solution (Smith and Johnson 2020). These systems can significantly enhance communication, automate daily tasks, and improve environmental control for individuals facing physical challenges (Johnson and Davis 2019). Recently, there has been an attempt to harness conversational systems, for providing tailored support for individuals with physical disabilities (Brown and Smith 2021). However, existing systems often fall short of providing a comprehensive and user-centric solution. Additionally, there is a need for a system that can adapt to individual preferences and requirements, considering the diverse nature of physical disabilities. Therefore, addressing these challenges necessitates a fresh perspective.

In response to this imperative, we propose an Empathetic Disability Support System (EDiSS), to offer tailored assistance with correct politeness and empathetic strategies based on individual users' persona, gender, and age. We start with the creation of a personalized disability support dialogue dataset (PDCARE), encompassing a diverse range of disabilities and support areas, such as Spinal Cord Injuries, Neurological Disorders, Orthopedic Disabilities, Physical Therapy Exercises, Pain Management, and Emotional Support. By drawing upon this comprehensive dataset, EDiSS employs a reinforcement learning-based dialogue model with a novel-designed reward function to adapt its tone and content to the unique needs and preferences of each user. Our evaluation demonstrates the efficacy of EDiSS in enhancing the quality of lives for individuals with disabilities. By providing respectful and empathetic assistance across various aspects of daily living, EDiSS represents a significant

[*]These authors contributed equally.

advancement in leveraging technology to address the challenges faced by individuals with disabilities. The *key* contributions can be summarized as follows:

1. Created a comprehensive physical disability support dialogue dataset PDCARE which encompasses five OCEAN personality traits (McCrae and Costa 1992) of the user and agent's three politeness strategies and eight empathy strategies to lay the foundation for the development of more sophisticated and physical disability support systems in the future.
2. Introduced EDiSS, an empathetic disability support system that places a strong emphasis on patient personality to foster politeness and empathy such that a cordial environment can be tailored to the unique needs of each individual.
3. Design a novel reward function by leveraging three transformer-based classifiers to ensure user's profile alignment, politeness, and empathy consistency.
4. Demonstrate the potential of EDiSS through extensive automatic and human evaluation in improving physical disability hurdles, offering hope and support.

Related Work

One area of research that has gained traction in recent years is the development of personalized conversational agents in healthcare (Kocaballi et al. 2019). The application spectrum includes post-stroke recovery (MIRANDA MACIEL 2023), behavior change (Zhang et al. 2020), and activities of daily life (Sheng et al. 2023). These studies show that by considering factors, such as gender, age, and personality traits, persona-based dialogue systems can offer more engaging responses.

Exploring personalized systems for physical disabilities, (Pereira and Díaz 2019) investigate the role of health chatbots in behavior change interventions, while (Huq, Maskeliūnas, and Damaševičius 2022) focus on conversational agents aiding individuals with cognitive disabilities, showcasing the diverse applications of personalized systems. (Cha et al. 2021) emphasize the empowering potential of voice-based agents for adolescents with Autism Spectrum Disorder, promoting inclusivity, and (Vigouroux et al. 2023) shift the focus to disability-friendly interfaces in home automation. (Wiratunga et al. 2020) introduce physical activity promotion in older adults through chatbots, and (Murali et al. 2023) contribute insights into automated pain assessment. (Smith and Dragone 2023) focus on daily living assessments, broadening the scope of personalized systems in healthcare. However, these systems often lack in providing comprehensive support for a spectrum of disabilities, as well as interactive and engaging experiences that cater to diverse user needs and contexts.

Integration of politeness and empathy into dialogue systems has witnessed significant advancements in recent years (Newbold et al. 2019; Mishra, Firdaus, and Ekbal 2022). (Rashkin et al. 2019) introduced a transformer-based approach for building empathetic dialogue systems, demonstrating improved performance in generating empathetic responses. (Liu et al. 2020) proposed a tag-and-generate method for politeness transfer, enabling conversational agents to incorporate politeness markers into generated responses. (Zhao and Eskenazi 2017) explored the use of conditional variational auto-encoders (CVAEs) to generate polite and empathetic dialogue, conditioning the generation process on user attributes and context. (Wang et al. 2019) focused on cross-language voice cloning and multilingual speech synthesis, enabling conversational agents to communicate fluently and empathetically in multiple languages. Additionally, (Samad et al. 2022) developed a reinforcement learning-based empathetic persuasive dialogue system for charity donation tracing user emotions to tailor persuasion strategies. These studies underscore the significance of politeness and empathy in enhancing user experience and engagement in conversational agents, paving the way for more effective and user-centric dialogue systems.

Our research takes a distinct focus. We center our efforts on implementing a diverse array of politeness and empathy strategies tailored to suit the gender, age, and persona of a user. This tailored approach enhances interpersonal dynamics and communication effectiveness. Our system, EDiSS, explores a spectrum of physical disability issues through 6,796 dialogues involving diverse patient profiles. Guided by novel-designed rewards, EDiSS crafts responses tailored to individual user profiles while ensuring the application of appropriate politeness and empathy strategies. To the best of our knowledge, EDiSS is the first attempt to develop a support system explicitly designed for individuals with physical disabilities, incorporating precise politeness and empathy strategies. Our contribution enhances the quality of support provided to this demographic, while also stimulating continued research and innovation in this often overlooked domain.

Dataset

We create PDCARE dataset consisting of physical disability support dialogues, aiming to provide personalized assistance to individuals facing physical challenges.

The PDCARE dataset tackles various issues associated with physical disabilities. It encompasses topics, such as *Accessibility Information*, *Travel Tips*, *Advocacy and Rights*, *Financial and Insurance Guidance*, *Mobility Aids*, *Home Modifications*, *Physical Therapy Exercises*, *Assistive Technology*, *Pain Management*, *Activities of Daily Living (ADLs)*, *Emotional Support*, *Employment and Education*, *Social Interaction*, *Fitness and Recreation*, *Peer Support Groups*, *Parenting with Disabilities*, and *Transitions and Life Changes*. The dataset addresses specific physical disability supports including *Mobility Impairments*, *Visual Impairments*, *Hearing Impairments*, *Speech Impairments*, *Neurological Disorders*, *Spinal Cord Injuries*, *Amputations*, *Orthopedic Disabilities*, *Cerebral Palsy*, *Muscular Dystrophy*, *Balance and Gait Disorders*, *Chronic Pain*, and *Aging-Related Disabilities*. Using Llama3-70B (AI@Meta 2024) with rigorous data quality control measures, we ensure the PDCARE's reliability and comprehensiveness. The details of each of the disabilities and respective supports can be found in Table 1 of the Appendix.

Metrics	Train	Validation	Test
# of Dialogues	5436	681	679
# of Utterances	124934	15521	15639
Avg.# Utterances/Dialogue	22.98	22.79	23.03

Table 1: Dataset statistics of PDCARE.

Dataset Creation

To create PDCARE dataset, we start with the design of prompts, each tailored to specific topics and corresponding physical disability issues. These prompts consist of instructions about the topic, the physical disability, gender: *male* and *female*, age: *younger*, *middle-aged*, and *older*, persona: *openness (O)*, *conscientiousness (C)*, *extraversion (E)*, *agreeableness (A)* and *neuroticism (N)* of the user. Unique challenges faced by different genders and age groups allow for a nuanced dialogue, while persona information helps in crafting a response that resonates with the individual's personality. To facilitate a natural and coherent dialogue between a doctor and a disabled user, seed utterances are used as starting points, combined with instructions to guide the conversation dynamically. The seed utterances consisting of 4-turns are obtained from GPT-3.5 (Ouyang et al. 2022) [1]. Further, seed utterances are quality checked in terms of *topic-consistency*, *context-adequacy*, and *fluency* by eight human participants having post-graduation in Linguistics and expertise in corresponding tasks. Quality checks were performed on an integer Likert scale of 1-3. The seed utterances with scores of 1 and 2 were corrected if needed and 3 were taken as intact. A reliable inter-evaluator kappa agreement (McHugh 2012) score for each of the three metrics is found to be 85.8%, 86.2%, and 87.4%, respectively in this phase.

The prompt with correct/modified seed utterances is given to Llama3-70B (AI@Meta 2024) with instruction <*generate 4 more turns in continuation of given dialogue to unfold user's issues and provide support without closing the dialogue*>. Now, the generated 8-turn dialogue was again quality checked by human participants in terms of three metrics as given above and was corrected if needed. The generated 8-turn dialogues are also quality-checked in terms of all three quality check metrics: *topic-consistency*, *context-adequacy*, and *fluency* as above. The 8-turn interactions having scores of 1 and 2 are again corrected or modified by the same 8 participants. Two examples of all three errors and corresponding corrections are shown in Table 5 of the appendix. Corrections made include restructuring responses for clarity, providing relevant information, and improving grammatical accuracy. The inter-evaluator kappa agreement (McHugh 2012) score for each of the metrics was found to be 81.3%, 82.8%, and 84.1%, respectively, in this phase. Additionally, these dialogues are also checked for user-profile alignment. If found to be non-aligned they are corrected/-modified.

Now this 8-turn dialogue is used as a prompt to generate the dialogues as per context in the given prompt with instructions to <*Complete the dialogue with polite and empathetic support as per the user's personality traits, gender, and age. The minimum and maximum number of turns allowed to complete the dialogue are 12 and 30, respectively. Engage the user as much as possible*>. This iterative process, guided by Llama3-70B (AI@Meta 2024), led to the generation of appropriate dialogues. Automated checks were then implemented to further enhance data quality. These checks included removing duplicate dialogues and conducting turn-level analysis to maintain smooth transitions within each dialogue. Prompts and examples of seed utterances are detailed in Figure 1 and Table 4 of the appendix, respectively. A sample 8-turn dialogue and complete dialogue generated are shown in Figure 2 and Figure 3 of the appendix. Additionally, various topics with associated physical disabilities, five personas, and their example utterances are detailed in Sections **Topics and Associated Physical Disabilities**, **Persona**, and Table 2 of the appendix, respectively.

[1]GPT-3.5 is not used to generate the dialogues itself due to the budget restrictions.

Data Quality Control

We applied data quality control measures to ensure high-quality data. Following the same processes as mentioned in the earlier Section , the generated dialogues were quality checked in terms of *topic-relevance*, *context-adequacy*, and *fluency*. First, the same eight participants quality-checked the generated dialogues based on guidelines stated in Section **Data Quality Control** of the Appendix. As previously, the dialogues are rated on a Likert scale of 1 to 3. For each of the metrics, the inter-evaluator Kappa (McHugh 2012) agreement score of 79.4%, 80.2%, and 82.4%, respectively, were observed. Any dialogues rated 1 were discarded; dialogues with improper language, as well as those rated 2, were modified and corrected.

Following this internal assessment, to ensure the dataset's alignment with best practices and real-world applicability, a subset comprising 5% of the diverse physical disability dialogues is sent to three medical experts specializing in physical therapy and disability management for a thorough review of dialogue-quality evaluation. As per feedback and guidelines provided by medical experts, the dialogues are again reviewed manually by all eight participants. Almost 7% of dialogues were modified again in this phase, and 3% of dialogues were discarded based on their irrelevance. The dual evaluation process involving both participants and medical experts guarantees that the PDCARE dataset is adequate, relevant, and linguistically fluent. The details of the data regarding quality checks are presented in Table 6 of the appendix.

Dataset Annotation

To have the politeness and empathetic strategy information, we annotate the created dataset with three politeness and eight empathetic strategies for the Doctor's responses at the utterance level. These carefully chosen strategies showcase a deeply emotional and cognitive understanding of their distinctive circumstances and create a welcoming space that nurtures disabled user's self-esteem.

Inspired by (Brown and Levinson 1987), we consider three politeness strategies *viz.* positive politeness strategy, negative politeness strategy, and bald-on record. The eight empathetic strategies can be given as:

- **Genuine Engagement:** Demonstrates sincere interest in understanding the individual's experiences and emotions, fostering a supportive environment (Rogers and Farson 1975).
- **Privacy Assurance:** Assures individuals that their personal information will be kept confidential, creating a safe space for them to share their concerns and feelings (Zhang and Liu 2021).
- **Forward Focus Encouragement:** Encourages individuals to focus on positive aspects and future goals despite physical challenges, fostering motivation and resilience (Ryan and Deci 2020).
- **Compassionate Validation:** Validates the individual's emotions and provides comfort and empathy, acknowledging the difficulties they face due to physical disabilities (Davies, Murphy, and Judd 2021).
- **Practical Assistance:** Provides practical support and advice for managing physical disabilities, including referrals to experts or resources for additional assistance (Noll, Mah, and May 2018).
- **Continuous Support:** Reassures individuals that they are not alone in their journey and emphasizes ongoing support and assistance available to them (Giroldi et al. 2019).
- **Strength-based Support:** Empowers individuals by recognizing their strengths and capabilities in managing their physical disabilities, promoting self-confidence and autonomy in decision-making (Ryan and Deci 2020).
- **No Strategy:** Assigned when responses do not employ any specific empathy strategy.

The same team of eight participants annotated both politeness and empathetic strategies. First, 50% of the dataset is manually annotated by the team, emphasizing the identification of politeness strategies and empathy strategies following the guidelines outlined in Section 1.4 of the appendix. Illustrative examples for each strategy were provided to ensure a shared understanding among annotators to manually annotate the required politeness and empathy strategy labels. The multi-rater Kappa agreement ratio (McHugh 2012) of 80.6% and 75.7% were observed for politeness and empathetic strategies, respectively. If the labels differ from the most voted labels for both types of strategies, annotators re-annotate the labels to achieve alignment and consistency.

Subsequently, considering 50% annotated dataset, we fine-tune RoBERTa-large (Liu et al. 2019) model for building politeness-strategy and empathy-strategy classifiers [2]. The un-annotated remaining 50% of the dataset is passed through these classifiers to obtain the politeness and empathy strategy labels. Manual verification is then performed by the same eight human participants who acted as annotators to cross-check the annotations. Finally, we obtain PDCARE - a persona-oriented disability support dialogue dataset. Dataset statistics are presented in the Table 1. Table 3 in the appendix provides example utterances illustrating different politeness and empathy strategies. Further, we provide comprehensive dataset details in Section **Dataset Details** of the Appendix.

[2]Politeness and Empathy strategy classifiers' weighted accuracy is found to be 91.6%, and 85.4% respectively.

Methodology

Initially, we start the development of EDiSS by employing a warm-start, i.e. fine-tuning the Phi-3-small model (Abdin et al. 2024) using the LORA technique (Hu et al. 2021) on the PDCARE dataset. This dataset comprises a collection of N dialogues between a physically disabled user and a system acting as a doctor. Each dialogue contains vital information about the user's gender, age, and persona. The model takes input x_i, incorporating the context, user's persona, age, and gender, represented as $x_i = [c_i + p_i + g_i + a_i]$. Here, $c_i = [c_{i-1} + u_i]$, denote the context and user's response at the i^{th} turn in the d^{th} dialogue. The output y_i corresponds to the system's response (Li et al. 2023). Our objective is to predict $\hat{y_i} \approx y_i$, i.e. we minimize the cross-entropy loss between the predicted $\hat{y_i}$ and actual system responses y_i:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(\hat{y}_{ij}) \tag{1}$$

Here, M denotes the vocabulary size, and $\hat{y}_{ij}$ represents the predicted probability of the j-th token in the vocabulary for the i-th dialogue. We call this trained system DSS and trained parameters θ.

EDiSS

In the subsequent phase, we further refine the DSS_θ within a reinforcement learning framework, employing the Proximal Policy Optimization (PPO) loss (Schulman et al. 2017). Here, we initialize the policy $\pi_\theta(a_t|s_t) = DSS_\theta$ as the probability distribution over actions a_t given the state s_t as per current policy parameters θ_t. An action a_t is the probability of selection of a response token from the vocabulary V. The state s_t at time step t is explicitly defined as $s_t = [c_t, m_t]$, where c_t denotes the ongoing dialogue context, and m_t signifies the model's memory (Schulman et al. 2017).

Rewards To guide the learning process effectively, we have devised five distinct novel rewards, which encompass both task-relevance and smoothness aspects of a support dialogue. These rewards ensure that the generated responses, denoted as $\hat{y}$, exhibit naturalness and consistency with the user's persona, gender, and age while also incorporating appropriate politeness and empathetic strategies.

Task-Relevance Reward: This reward is designed to encourage the model to generate responses that are relevant to the task or goal at hand.

$$R_{\text{task-relevance}} = \frac{1}{1+\exp(-\lambda\cdot(w_1\cdot\Delta_1 + w_2\cdot\Delta_2 + w_3\cdot\Delta_3))} \tag{2}$$

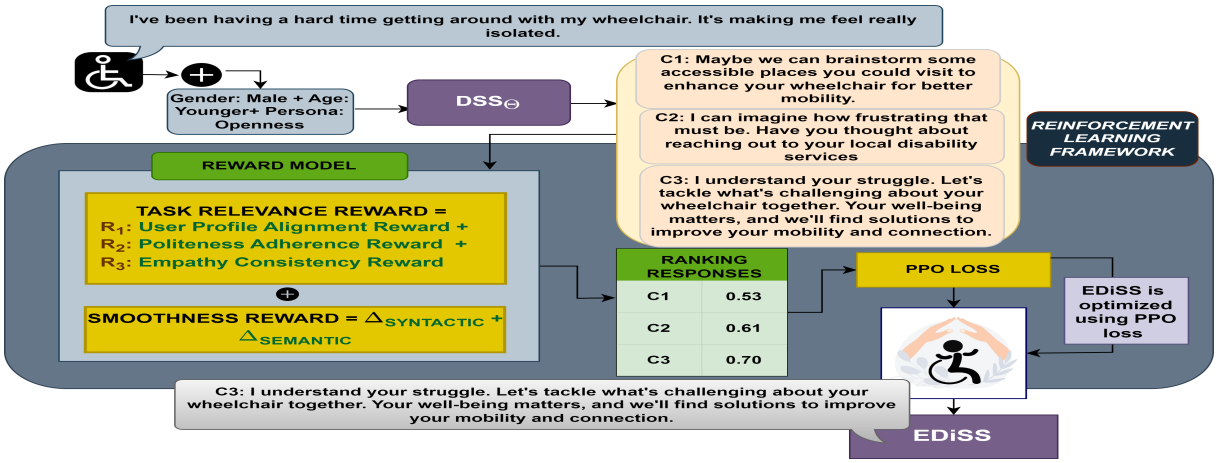

Figure 1: Overall architecture of the proposed system **EDiSS**.

In this equation, Δ_1, Δ_2, and Δ_3 represent three different task-relevant measures, such as user-profile alignment, politeness-strategy correctness, and empathy-strategy correctness, respectively. w_1, w_2, w_3 are the weights given to each of the rewards with $w_1 + w_2 + w_3 = 1$. λ is a scaling factor that determines the sensitivity of the reward to changes in task relevance.

1. **User-Profile Alignment Reward:** This reward encourages the model to generate responses that are aligned with the user's persona, gender, and age. It evaluates the model's ability to understand and adapt to user characteristics:

$$\Delta_1 = \text{CLS}_{\text{pga}^k}(y) - \alpha\text{CLS}_{\text{pga}^k}(\hat{y}) \quad (3)$$

where $\text{CLS}_{\text{pga}}()$ computes the probability of $0 \leq k^{th} < PGA$ persona-gender-age class out of K classes [3]. A PGA class is given by the combination of persona, age, and gender. Here, we consider five personas $P = \{O, C, E, A, N\}$, two genders $G = \{Male, Female\}$, and three group of ages $A = \{Younger, Middle - Aged, Older\}$. Therefore, the total combinations of all these three would be 30, hence we will have a total of 30 PGA classes.

2. **Politeness Adherence Reward:** This reward incentivizes the generation of responses that adhere to predefined politeness strategies, fostering courteous interactions:

$$\Delta_2 = \text{CLS}_{\text{ps}^k}(y) - \alpha\text{CLS}_{\text{ps}^k}(\hat{y}) \quad (4)$$

where $\text{CLS}_{\text{ps}}()$ computes the probability of $0 \leq k^{th} < PS$ politeness strategy class out of PS classes.

3. **Empathy Consistency Reward:** This reward promotes responses that demonstrate correct empathy strategy to understand the user's emotional state and needs:

$$\Delta_3 = \text{CLS}_{\text{es}^k}(y) - \alpha\text{CLS}_{\text{es}^k}(\hat{y}) \quad (5)$$

where $\text{CLS}_{\text{es}}()$ computes the probability of $0 \leq k^{th} < ES$ empathetic-strategy class out of ES classes.

[3]The weighted accuracy of persona-gender-age (PGA) classifier's 80.2%

In each of the rewards, $\alpha = [1, 2]$ acts as a penalization factor.

Smoothness Reward: These rewards encourage the model to produce responses that exhibit smooth transitions and coherence within the conversation. It penalizes abrupt changes or inconsistencies between consecutive utterances.

$$R_{\text{smoothness}} = \frac{1}{1 + \exp\left(-\lambda \cdot (w_4 \cdot \Delta_{\text{syn}} + w_5 \cdot \Delta_{\text{sem}})\right)} \quad (6)$$

In this equation, $w_4 + w_5 = 1$ are weighting factors that balances the contribution of syntactic and semantic smoothness scores. It determines the relative importance of each aspect in the overall smoothness reward. λ is the scaling factor that controls the sensitivity of the reward to changes in smoothness. A higher λ amplifies the importance of smoothness in the overall reward calculation, while a lower value reduces its impact. Δ_{syn} represents the syntactic smoothness score which penalizes deviations from grammatical correctness computed as reciprocal of perplexity ($PPL()$) (Brown et al. 1992). It can be measured as:

$$\Delta_{\text{syn}} = \frac{1}{PPL(\hat{y})} \quad (7)$$

Δ_{sem} represents the semantic smoothness score, which assesses the semantic coherence and relevance between consecutive utterances.

$$\Delta_{\text{sem}} = \text{cosine\_similarity}(u_{i-1}, \hat{y})) \quad (8)$$

We define the overall reward R as:

$$R = \gamma R_{\text{task-relevance}} + (1 - \gamma) R_{\text{smoothness}} \quad (9)$$

where $\gamma = [0, 1]$. Then, the advantage function $\hat{A}_t$ is computed using the rewards obtained from the environment.

$$\hat{A}_t = R_t - V(s_t) \quad (10)$$

where R_t is the total reward obtained at time step t, and $V(s_t)$ is the state-value function representing the expected cumulative reward from state s_t onwards. The policy π_θ is

updated using the proximal policy optimization (PPO) loss function:

$$L^{PPO}(\theta) = -\mathbb{E}[\min(r(\theta)\hat{A}_t, \text{clip}(r(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \quad (11)$$

where $r(\theta)$ is the probability ratio, $\hat{A}_t$ is the advantage function, and ϵ is the clipping parameter. The parameters θ of the policy π_θ are updated using gradient descent with the modified PPO loss incorporating the reward:

$$\theta_{t+1} = \theta_t - \alpha \nabla_\theta L^{PPO}(\theta) \quad (12)$$

where α is the learning rate.

Experiments

Due to space restrictions, the Baselines and implementation details of all the models are given in Section **Baselines** and **Implementation Details** of the Appendix.

Evaluation Metrics

Both automatic and human evaluations are conducted to assess the performance of the proposed system **EDiSS**.

Automatic Evaluation Metrics: We employ three metrics to evaluate task-relevance *viz.* user-profile consistency (UPC), politeness-strategy accuracy (PSA), and Empathy-Strategy accuracy (ESA):

$$UPC = \mathbb{E}_{x_i,y_i} 1\{CLS_{\text{pga}}(y_i) = CLS_{\text{pga}}(\hat{y})\}, \quad (13)$$

$$PSA = \mathbb{E}_{x_i,y_i} 1\{CLS_{\text{ps}}(y_i) = CLS_{\text{ps}}(\hat{y})\}, \quad (14)$$

$$ESA = \mathbb{E}_{x_i,y_i} 1\{CLS_{\text{es}}(y_i) = CLS_{\text{es}}(\hat{y})\}, \quad (15)$$

Additionally, we evaluate **EDiSS** in terms of language and dialogue quality using three metrics: Perplexity (PPL) (Brown et al. 1992), Response Length Ratio (R_{len}), Non-repetitiveness (N_{rep}).

$$PPL = \frac{\sum_r \exp\left(-\frac{1}{n}\sum_{i=1}^{n} \log P(y_i|x_i)\right)}{r} \quad (16)$$

where n is the total number of tokens in the generated responses, r is the total number of the generated responses, and $P(y_i|x_i)$ is the probability assigned by the language model to the i^{th} token given the input x_i.

$$R_{\text{len}} = \frac{\sum_r(n)}{r}. \quad (17)$$

$$N_{\text{rep}} = \frac{1}{2}(BS_{\text{F1}}(y_i, y_{i-1}) + BS_{\text{F1}}(y_i, y_{i-2})), \quad (18)$$

Human Evaluation Metrics: Human evaluation involves 10 evaluators, who were compensated according to the university norms. All ten evaluators have post-graduate in English Linguistics and have at least three years of experience in doing the similar kind of tasks. To avoid bias these evaluators are different from annotators. The evaluation consists of two phases. In the first phase, each evaluator interacts with **EDiSS** five times, using different sets of utterances. They rate the conversations based on a Likert scale of 1-5 for seven metrics: persona accuracy, gender-age accuracy, politeness accuracy, empathy accuracy, fluency (FY), consistency (CY), and non-repetitiveness (NR). The scale denotes low to high intensity, e.g., a rating of 1 for persona accuracy indicates low consistency, while 5 denotes high consistency. These 50 evaluations are reviewed by medical experts, achieving an agreement score of 83.4%. In the second phase, based on expert feedback, evaluators re-evaluate the initial 50 interactions and assess an additional 15 interactions each, resulting in a total of 200 evaluated interactions.

Results and Analysis

Automatic Evaluation: Table 2 presents the results of automatic evaluation metrics for various baselines: GPT2-large (Radford et al. 2019), ARDM (Wu et al. 2021), ZYPHER-7B (Tunstall et al. 2023), Phi-1.5 (Li et al. 2023), Mistral-7B (Jiang et al. 2023), Llama2-7B (Touvron et al. 2023), phi-2 (Li et al. 2023), Mistral-8B (Jiang et al. 2023), Llama3-8B (AI@Meta 2024), DSS: DSS_θ, EDiSS-R: EDiSS with $R = 0$, EDiSS-TR: EDiSS with $R = R_{\text{smoothness}}$, and EDiSS-SR: EDiSS with $R = R_{\text{task-relevance}}$, compared against our proposed **EDiSS**. Significant differences were observed between EDiSS and all other models ($p < 0.05$). Among the compared models, EDiSS consistently outperforms others across all metrics.

In examining task-specific metrics: UPC, PSA, and ESA a discernible pattern is seen i.e. GPT2-large < ARDM < Phi-1.5 < Llama2-7B < Mistral-7B < ZYPHER-7B < Phi-2 < Mistral-8B < Llama3-8B < DSS ≈ EDiSS-R < EDiSS-TR < EDiSS-SR < EDiSS. Notably, EDiSS and EDiSS-R exhibit similar performance, attributed to EDiSS's initialization from DSS_θ. The better performance of EDiSS-SR can be traced back to the influence of Δ_1, Δ_2, and Δ_3, underscoring the pivotal role of persona, gender, age, politeness, and empathy in guiding EDiSS to formulate persona-consistent, polite, and compassionate responses. Moreover, Table 2 demonstrates that EDiSS outperforms all the 13 baselines in terms of PPL, R_{len}, and N_{rep}, following the same order as above. The better performance of EDiSS-TR is attributed to Δ_{syn} and Δ_{sym}, which steer it towards more natural and contextually consistent responses.

EDiSS's success across all the metrics can be attributed to its assimilation of patient profile information and adaptation of politeness and empathy levels. The integration of task-relevance reward aids EDiSS in approximating a more precise distribution, further enhancing its competitive edge over the eight baselines. The inclusion of smoothness reward fosters a dynamic rapport between the system and the user, enabling EDiSS to focus on pertinent details and craft refined responses. This results in better language understanding and, the ability to generate contextually relevant, diverse, and engaging responses. This underscores the dual necessity of all five rewards in yielding responses of elevated quality, validating our initial hypothesis. Generated responses of different models with respective qualitative analyses are illustrated in Section **Qualitative analysis** and Table 9 of the appendix.

Human Evaluation: Table 3 showcases human evaluation results for GPT2-large, ARDM, Phi-1.5, Llama2-7B, Mistral-7B, ZYPHER-7B, Phi-2, Mistral-8B, Llama3-8B, DSS, EDiSS, EDiSS-R, EDiSS-TR, and EDiSS-SR, compared against EDiSS. Similar to the automatic evaluation,

Model	UPC	PSA	ESA	PPL	R_{len}	N_{rep}
GPT2-large (Radford et al. 2019)	44.1%	67.9%	55.3%	24.34	10.12	0.41
ARDM (Wu et al. 2021)	50.1%	70.2%	62.1%	10.21	12.01	0.32
Phi-1.5 (Li et al. 2023)	50.9%	71.6%	63.4%	8.01	15.59	0.24
Llama2-7B (Touvron et al. 2023)	51.3%	72.5%	65.8%	7.90	16.02	0.22
Mistral-7B (Jiang et al. 2023)	51.5%	73.8%	66.1%	7.80	16.10	0.20
ZYPHER-7B (Tunstall et al. 2023)	52.1%	74.5%	66.8%	7.41	16.24	0.19
Phi-2 (Li et al. 2023)	54.4%	76.5%	68.9%	6.61	17.95	0.16
Mistral-8B (Jiang et al. 2023)	55.6%	77.2%	69.5%	5.95	19.10	0.14
Llama3-8B (AI@Meta 2024)	56.1%	77.5%	70.2%	5.60	19.95	0.12
DSS (Phi3-small) (Abdin et al. 2024)	57.9%	79.1%	71.5%	4.85	19.70	0.10
EDiSS-R	57.5%	78.8%	71.5%	4.88	19.70	0.10
EDiSS-TR	58.4%	79.6%	72.0%	4.50	19.85	0.09
EDiSS-SR	59.3%	80.5%	73.6%	4.10	20.05	0.08
EDiSS	**60.7%**	**81.5%**	**74.9%**	**3.60**	**20.45**	**0.07**

Table 2: Results of automatic evaluation. Significant differences were observed between **EDiSS** and all other models ($p < 0.05$).

Model	UPC	PSA	ESA	FY	CY	N_{rep}
GPT2-large	1.70	2.12	2.01	2.60	2.35	2.40
ARDM	2.05	2.34	2.25	3.46	2.72	2.76
Phi-1.5	2.26	2.95	2.63	3.64	2.95	2.98
Llama2-7B	2.30	3.05	2.68	3.70	3.05	3.04
Mistral-7B	2.38	3.10	2.75	3.75	3.10	3.08
ZYPHER-7B	2.40	3.15	2.78	3.75	3.10	3.10
Phi-2	2.45	3.20	2.91	3.83	3.18	3.15
Mistral-8B	2.65	3.40	3.05	3.95	3.38	3.35
Llama3-8B	2.70	3.50	3.10	4.00	3.45	3.40
DSS	2.76	3.62	3.18	4.10	3.54	3.50
EDiSS-R	2.72	3.55	3.15	4.10	3.52	3.50
EDiSS-TR	2.82	3.68	3.25	4.18	3.70	3.65
EDiSS-SR	3.01	3.90	3.55	4.30	3.92	3.95
EDiSS	**3.15**	**4.02**	**3.85**	**4.42**	**4.05**	**4.10**

Table 3: Results of human evaluation

EDiSS outperforms all the other models across metrics: UPC, PSA, ESA, FY, CY, and N_{rep}. A nuanced contrast emerges between EDiSS and EDiSS-TR, emphasizing the significance of task-relevance rewards—Δ_1, Δ_2, and Δ_3 — in crafting persona-sensitive, polite, and empathetic responses. Notably, EDiSS surpasses EDiSS-TR and EDiSS-SR, indicating the pivotal role of all five rewards in achieving fluent, consistent, non-repetitive, courteous, and compassionate responses. These enhancements reflect EDiSS's ability to generate human-like and engaging conversations, thus boosting user satisfaction. The superior performance of EDiSS is attributed to its reward-based architecture, optimizing response quality.

Both automatic and human evaluation validate EDiSS's efficacy in delivering high-quality conversational support to individuals with physical disabilities, suggesting its potential to enhance user experience and overall well-being significantly.

Error Analysis

While **EDiSS** demonstrates effectiveness in providing tailored support to individuals with physical disabilities, some areas for improvement remain. A key issue is the occasional misalignment between user personas and generated responses, stemming from the complexity of human traits and challenges in accurately capturing them in the dataset. Instances of sub-optimal politeness, empathy strategies, and fragmented dialogue flow were also observed, often due to limitations in training data or contextual understanding over extended conversations. Classifier accuracy plays a crucial role in **EDiSS**'s performance. With accuracies of 91.6% (politeness), 85.4% (empathy), and 80.2% (PGA), most responses align with user profiles, ensuring polite and empathetic strategies. Misclassifications, however, may introduce minor biases. To mitigate this, reward functions adjust values based on classification confidence, reducing the impact of misclassifications. As shown in Table 2, **EDiSS** achieves maximum values of 60.7% (UPC), 81.5% (PSA), and 74.9% (ESA). To enhance **EDiSS**, we propose integrating adaptive user feedback for dynamic response tuning, refining classifiers with real-world data to improve demographic robustness, and exploring lightweight adaptations for offline use in low-connectivity areas. These steps aim to enhance **EDiSS**'s adaptability, inclusivity, and reliability.

Conclusion

In this paper, we present **EDiSS**, an empathetic disability support system designed for individuals with physical disabilities. Using user personas based on the *OCEAN* model, gender, and age, EDiSS delivers personalized assistance in a supportive environment. It integrates politeness and empathy strategies to enhance user experience. Built on the PDCare dataset, enriched with user profile annotations, **EDiSS** optimizes responses with task-relevance and smoothness rewards. Empirical evaluations, both automatic and human, confirm its effectiveness in tailored support. **EDiSS** lays the groundwork for future research on more inclusive, personalized systems, with potential extensions to broader domains and additional user profile factors.

Ethical Statement

Ethical considerations are central to the development of **EDiSS**, especially given its focus on supporting individuals with physical disabilities. We adhered to strict ethical guidelines, prioritizing user privacy, autonomy, and well-being. Data privacy was ensured through anonymization and compliance with data protection regulations. The **PD-CARE** dataset emphasized diversity in age, gender, personality traits, and disabilities, exposing the model to a wide range of user profiles. Bias-check mechanisms were employed to identify and manually correct harmful or stereotypical outputs, ensuring equitable treatment across persona combinations. Our data and methodology are approved by the university's ethics review board, and the dataset will be available for research purposes upon request. These efforts reflect our commitment to safeguarding the dignity, rights, and well-being of users.

Acknowledgements

Kshitij Mishra gratefully acknowledges the support of the Prime Minister's Research Fellowship (PMRF) from the Government of India for enabling this research. The authors also extend their gratitude to Google for the "Gemma Academic Program GCP Credit Award," which provided cloud credits to support this work.

References

Abdin, M.; Jacobs, S. A.; Awan, A. A.; Aneja, J.; Awadallah, A.; Awadalla, H.; Bach, N.; Bahree, A.; Bakhtiari, A.; Behl, H.; et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

AI@Meta. 2024. Llama 3 Model Card.

Brown, A.; and Smith, J. 2021. Enhancing Accessibility: Conversational Systems for Individuals with Physical Disabilities. *Assistive Technology Journal*, 25(2): 45–62.

Brown, P.; and Levinson, S. C. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Brown, P. F.; Della Pietra, S. A.; Della Pietra, V. J.; Lai, J. C.; and Mercer, R. L. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1): 31–40.

Cha, I.; Kim, S.-I.; Hong, H.; Yoo, H.; and Lim, Y.-k. 2021. Exploring the use of a voice-based conversational agent to empower adolescents with autism spectrum disorder. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–15.

Davies, S.; Murphy, M.; and Judd, G. 2021. The Importance of Emotional Support for Patients with Disabilities. *Disability Studies Quarterly*, 41(4): 1–16.

Giroldi, E.; Veldhuijzen, W.; Leijten, C.; Welter, D.; van der Weijden, T.; Muris, J.; van der Vleuten, C.; and Kester, A. 2019. Doctor's empathic communication and patients' somatization: A randomized controlled trial. *Journal of Family Practice*, 68(5): E1–E11.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huq, S. M.; Maskeliūnas, R.; and Damaševičius, R. 2022. Dialogue agents for artificial intelligence-based conversational systems for cognitively disabled: A systematic review. *Disability and Rehabilitation: Assistive Technology*, 1–20.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Johnson, R.; and Davis, S. 2019. Conversational Agents in Daily Living: A Survey. *Journal of Assistive Technologies*, 11(3): 145–160.

Kocaballi, A. B.; Berkovsky, S.; Quiroz, J. C.; Laranjo, L.; Tong, H. L.; Rezazadegan, D.; Briatore, A.; and Coiera, E. 2019. The personalization of conversational agents in health care: systematic review. *Journal of medical Internet research*, 21(11): e15360.

Li, Y.; Bubeck, S.; Eldan, R.; Del Giorno, A.; Gunasekar, S.; and Lee, Y. T. 2023. Textbooks Are All You Need II: **phi-1.5** technical report. *arXiv preprint arXiv:2309.05463*.

Liu, T.; Shah, P.; Chang, A.; and Lapata, M. 2020. Tag & generate: Learning to generate helpful responses with attributes. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6159–6171.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

McCrae, R. R.; and Costa, P. T. J. 1992. Personality in Adulthood: A Five-Factor Theory Perspective. *Guilford Press*.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3): 276–282.

MIRANDA MACIEL, P. 2023. *The Co-design of an Embodied-Conversational-Agent-based system to help stroke survivors to manage their recovery: the iTakeCharge study*. Ph.D. thesis, Macquarie University.

Mishra, K.; Firdaus, M.; and Ekbal, A. 2022. Please be polite: Towards building a politeness adaptive dialogue system for goal-oriented conversations. *Neurocomputing*, 494: 242–254.

Murali, P.; Arjmand, M.; Volonte, M.; Li, Z.; Griffith, J.; Paasche-Orlow, M.; and Bickmore, T. 2023. Towards Automated Pain Assessment using Embodied Conversational Agents. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, 131–140.

Newbold, J.; Doherty, G.; Rintel, S.; and Thieme, A. 2019. Politeness Strategies in the Design of Voice Agents for Mental Health.

Noll, D.; Mah, J.; and May, L. 2018. Empowering individuals with disabilities: Perspectives on assistance and support. *International Journal of Environmental Research and Public Health*, 15(4): 673.

Organization, W. H. 2021. Disability and health. *World Health Organization*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Pereira, J.; and Díaz, Ó. 2019. Using health chatbots for behavior change: a mapping study. *Journal of medical systems*, 43: 1–13.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Rashkin, H.; Smith, E. M.; Li, M.; Boureau, Y.-L.; and Liang, P. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Rogers, C. R.; and Farson, R. E. 1975. Communicating empathy in medical interviews. *The Journal of the American Medical Association*, 15: 2223–2225.

Ryan, R. M.; and Deci, E. L. 2020. The role of motivation in physical rehabilitation: understanding and applying the importance of the patient's perspective. *Rehabilitation Psychology*, 65(3): 205–214.

Samad, A. M.; Mishra, K.; Firdaus, M.; and Ekbal, A. 2022. Empathetic persuasion: reinforcing empathy and persuasiveness in dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 844–856.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sheng, Z.; Finzel, R.; Lucke, M.; Dufresne, S.; Gini, M.; and Pakhomov, S. 2023. A Dialogue System for Assessing Activities of Daily Living: Improving Consistency with Grounded Knowledge. *arXiv preprint arXiv:2307.07544*.

Smith, E.; and Johnson, R. 2020. Voice-Controlled Interfaces for Assistive Technology. In *Proceedings of the International Conference on Human-Computer Interaction*, 123–135.

Smith, R.; and Dragone, M. 2023. Generalisable Dialogue-based Approach for Active Learning of Activities of Daily Living. *ACM Transactions on Interactive Intelligent Systems*, 13(3): 1–37.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Tunstall, L.; Beeching, E.; Lambert, N.; Rajani, N.; Rasul, K.; Belkada, Y.; Huang, S.; von Werra, L.; Fourrier, C.; Habib, N.; Sarrazin, N.; Sanseviero, O.; Rush, A. M.; and Wolf, T. 2023. Zephyr: Direct Distillation of LM Alignment. arXiv:2310.16944.

United Nations. 2024a. Leave No One Behind. https://www.un.org/sustainabledevelopment/leave-no-one-behind/. Accessed: 2024-02-20.

United Nations. 2024b. Sustainable Development Goals. https://sdgs.un.org/goals. Accessed: 2024-02-20.

Vigouroux, N.; Vella, F.; Lepage, G.; and Campo, E. 2023. Design Recommendations Based on Speech Analysis for Disability-Friendly Interfaces for the Control of a Home Automation Environment. In *International Conference on Human-Computer Interaction*, 197–211. Springer.

Wang, X.; Wan, L.; Hai, J.; Xie, L.; and Zhu, J. 2019. Cross-lingual voice cloning and emotional speech synthesis for code-switching voice assistant. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1690–1699.

Wiratunga, N.; Cooper, K.; Wijekoon, A.; Palihawadana, C.; Mendham, V.; Reiter, E.; and Martin, K. 2020. FitChat: conversational artificial intelligence interventions for encouraging physical activity in older adults. *arXiv preprint arXiv:2004.14067*.

Wu, Q.; Zhang, Y.; Li, Y.; and Yu, Z. 2021. Alternating Recurrent Dialog Model with Large-scale Pre-trained Language Models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1292–1301.

Zhang, J.; Oh, Y. J.; Lange, P.; Yu, Z.; and Fukuoka, Y. 2020. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet. *Journal of medical Internet research*, 22(9): e22845.

Zhang, Q.; and Liu, W. 2021. Recent advances in conversational systems: A survey. *Journal of Intelligent Information Systems*.

Zhao, T.; and Eskenazi, M. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 654–664.