Representation Bias of Adolescents in AI: A Bilingual, Bicultural Study

Robert Wolfe*, Aayushi Dangol*, Bill Howe, Alexis Hiniker

University of Washington
*Equal Contribution
rwolfe3@uw.edu, adango@uw.edu, billhowe@uw.edu, alexisr@uw.edu

Abstract

Popular and news media often portray teenagers with sensationalism, as both *a risk* to society and *at risk* from society. As AI begins to absorb some of the epistemic functions of traditional media, we study how teenagers in two countries speaking two languages: 1) are depicted by AI, and 2) how they would prefer to be depicted. Specifically, we study the biases about teenagers learned by static word embeddings (SWEs) and generative language models (GLMs), comparing these with the perspectives of adolescents living in the U.S. and Nepal. We find English-language SWEs associate teenagers with societal problems, and more than 50% of the 1,000 words most associated with teenagers in the pretrained GloVe SWE reflect such problems. Given prompts about teenagers, 30% of outputs from GPT2-XL and 29% from LLaMA-2-7B GLMs discuss societal problems, most commonly violence, but also drug use, mental illness, and sexual taboo. Nepali models, while not free of such associations, are less dominated by social problems. Data from workshops with N=13 U.S. adolescents and N=18 Nepalese adolescents show that AI presentations are disconnected from teenage life, which revolves around activities like school and friendship. Participant ratings of how well 20 trait words describe teens are decorrelated from SWE associations, with Pearson's ρ=.02, *n.s.* in English FastText and ρ=.06, *n.s.* in GloVe; and ρ=.06, *n.s.* in Nepali FastText and ρ=−.23, *n.s.* in GloVe. U.S. participants suggested AI could fairly present teens by highlighting *diversity*, while Nepalese participants centered *positivity*. Participants were optimistic that, if it learned from *adolescents*, rather than media sources, AI could help mitigate stereotypes. Our work offers an understanding of the ways SWEs and GLMs misrepresent a developmentally vulnerable group and provides a template for less sensationalized characterization.

Introduction

Teenagers feature more prominently in western media accounts of new technologies than perhaps any other user group. They are the group most likely to adopt and capably use new technologies, including social media (Vogels and Gelles-Watnick 2023) and ChatGPT (Klar 2023). However, to read media accounts, they are also the most likely to *misuse* new technologies, leading to harm to others, or inadvertent harm to themselves (Stern and Burke Odland 2017). Such narratives have consequences for adolescent access to technology: concerns about compulsive use of social media, cyberbullying, and sexual predation led to a March 2024 ban on use of numerous social media platforms by younger teenagers in the state of Florida (The Guardian 2024). Concerns about deceptive design and online safety warrant consideration; yet the response—a blanket ban—suggests a framing that emphasizes the danger of adolescent technology use and affords adolescents little agency.

Such presentations continue a decades-long trend in western media portraying teenagers as simultaneously *a risk* to society and *at risk* from society (Pain 2003). Though largely disconnected from most adults' experiences with teens (Aubrun and Grady 2000), media portrayals of adolescents have centered violence, drug abuse, hyper-sexualization, technology addiction, and even religious fanaticism as pressing issues that warrant responses ranging from targeted media campaigns to government legislation (Clark 2005; Marwick 2008; Glassner 2010; Telzer et al. 2022). Though such portrayals appear sensationalistic in hindsight, representations of teenagers in media sources nonetheless shape adults' beliefs about what adolescents are like, influencing the treatment of adolescents in public places (Bernier 2011) and the restrictiveness of policy intended to influence adolescent behavior (Dorfman and Schiraldi 2001), thus directly affecting how teens are treated in practice.

In the present work, we study societal attitudes toward adolescents learned by static word embeddings (SWEs) and generative language models (GLMs), comparing with attitudes reported by adolescents themselves. Because prior work suggests attitudes toward adolescents vary across cultures (Larson and Wilson 2004; Di Giunta et al. 2023), we undertake a bilingual, bicultural study, examining U.S. attitudes and English-language models, as well as models trained on Nepali, a low-resource language spoken primarily in Nepal, a South Asian country in the Global South, and a native language for a first author of this work. We held workshops with N=13 English-speaking adolescents in the U.S. and N=18 Nepali-speaking adolescents in Nepal, asking how adolescents *are* represented in media, and how they *should* be represented in AI. We make three contributions:

- **We show that English-language SWEs and GLMs associate adolescents predominantly with social problems.** Clustering the 1,000 words most associated with

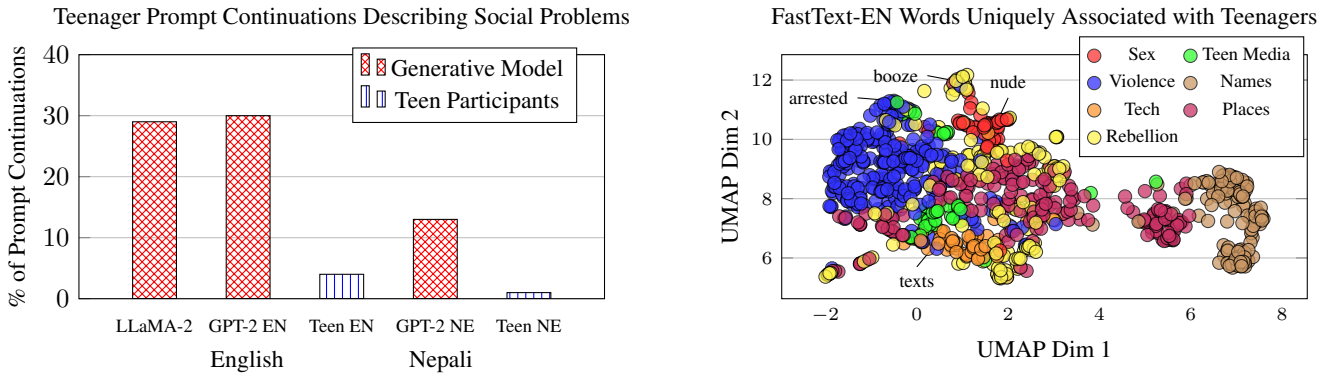

Figure 1: Left: Teenage participants were much less likely to continue prompts about teenagers with social problems than were GLMs. Right: Words associated with adolescents over any other age group in English SWEs reflect violence, rebellion, and sexualization.

teenagers in English GloVe and FastText SWEs reveals that clusters related to drugs, rebellion, violence, mental illness, stereotypes, and sexual taboo account for more than 50% of words in GloVe and more than 40% in FastText. Similarly, using prompts about teenagers derived from Stern (2005), we show that 29% of English LLaMA-2-7B outputs and 30% of GPT2-XL outputs depict societal problems. Of these, 47% depict violence in LLaMA-2, and 50% depict violence in GPT2-XL. Many such outputs mimic the format of "high-quality" training data—newspapers and journalistic media. Only 13% of distilGPT2 Nepali continuations reflect societal problems, and 10.1% of the words most associated with teenager in Nepali GloVe describe societal problems.

- **We show that AI representations are disconnected from adolescent self-perceptions.** Adolescent ratings of their own traits are decorrelated from SWE associations between corresponding trait vectors and the *teenager* vector, with Pearson's ρ=.02, *n.s.* in FastText and ρ=.06, *n.s.* in GloVe for English; and ρ=.06, *n.s.* in FastText and ρ=−.23, *n.s.* in GloVe for Nepali. Participant continuations of the same prompts used with GLMs show social problems arise in fewer than 4% of U.S. participant continuations and fewer than 1% of Nepalese participant continuations (Fig 1).
- **We discuss two central concerns of participants for fair representation in AI: diversity and positivity.** U.S. and Nepalese participants were aware of adolescent media stereotypes, and noted the difficulty in achieving fair representation. U.S. participants stressed that AI should foreground the *diversity* of teenagers, while Nepalese participants stressed that AI should present the positive traits of teenagers. Both groups expressed optimism that AI could correct media stereotypes about adolescents.

Our work shows that GLMs learn societal biases latent in media framings. As user-facing GLMs are integrated into schools and other contexts where they will impact adolescents' lives, research must center participatory approaches to AI (Delgado et al. 2023) to ensure groups with less agency, like adolescents, are represented in ways that capture not a media presentation but a group's understanding of itself.

Related Work

We review prior work on depictions of adolescents in popular and news media, sources often used to train AI. We then consider the language models studied and age biases in AI.

Defining Adolescence

The National Institutes of Health (NIH) define Adolescents as persons between 13 and 17 years old, distinct from Children (1 through 12), Adults (18 and older), and Older Adults (65 and older) (NIH 2022). While definitions may vary between cultures and across time (Arnett 1999), we adopt the NIH definition, which is consistent with related work.

Media Representations of Adolescents

Prior work finds that popular and news media depictions of adolescents are generally negative, with positive interactions involving teenagers portrayed as deviations from the norm (Bernier 2011). News coverage of teenagers often depicts supposed epidemics of violence, crime, drug abuse, mental illness, and immorality, which are usually not well supported by evidence (Glassner 2010; Telzer et al. 2022). In foundational work, Dorfman et al. (1997) find that most California TV news reports related to violence feature youth, and that only education policy receives as much treatment as violence in newspaper coverage about adolescents. Males (1999) find that LA Times articles included adolescents in stories about violence five times more frequently than adults. Adolescent behavior may be presented as dangerous even when not volitional, as Best (2008) find that activities as simple as teenage driving can be framed as pressing issues in the media. More recently, teenage use of technology has become a subject of public concern, and Stern and Burke Odland (2017) find that print and online news media portray teens as having an unhealthy relationship with social media. Previously, Stern (2005) found that U.S. films depict teenagers as violent, self-absorbed, and disengaged from civic life. As discussed in the Methods, we draw on Stern (2005) to create GLM prompts.

Societal Impact Media depictions shape adult views of adolescents and may shape adolescent behavior. Hancock

(2001) shows that adults overestimate and perceive illusory increases in adolescent crime. Aubrun and Grady (2000) find most adults report good experiences with teenagers they know but consider such experiences atypical, rather than questioning media framing. Dorfman and Schiraldi (2001) note that negative media portrayals, especially of adolescents of color, lend justification to harsher treatment and more restrictive policies. Moreover, Qu et al. (2020) find that younger teens' *own* beliefs in teenage stereotypes contribute to behavioral problems. Buchanan et al. (2023a) argue that, to prevent a self-fulfilling prophecy, descriptions of adolescent "stress and storm" must be replaced with a less reductive framing, such as "possibility and promise."

Societal Variation Though some aspects of adolescence appear consistent around the world (Steinberg et al. 2018), scholars describe significant variation in characterizations of adolescence both within and across cultures (Buchanan et al. 2023b). Enright et al. (1987) note that definitions of adolescence change over time based on society's needs: during war time, teens are portrayed as rugged and adultlike, but when not desired in the workforce, teens are portrayed as more childlike. Arnett (1999) note that adolescent stress may be more pronounced in individualistic western cultures, while Larson and Wilson (2004) use the plural form "adolescences" to describe variations around the world and across time, noting that teen years are not consistently characterized by emotional turmoil and psychic separation from parents. Finally, Di Giunta et al. (2023) observe differences in emotion regulation in teenagers in Italy and Colombia, suggesting cultural factors play a role in adolescent well-being.

Language Models

In this work, we study **static word embeddings** (SWEs) and **generative language models** (GLMs). SWEs are trained using deep neural networks (DNNs) to represent words as vectors based on the conditional probability of their co-occurrence with surrounding words (Mikolov et al. 2013; Collobert et al. 2011). We study FastText (Bojanowski et al. 2017), an extension of Word2Vec (Mikolov, Yih, and Zweig 2013) that incorporates subword information, and Global Vectors for Word Representation (GloVe) (Pennington, Socher, and Manning 2014), which incorporates corpus-level statistics to improve semantics. SWEs are now widely used in social science (Bhatia and Walasek 2023; Guan et al. 2024) to study societal attitudes (Garg et al. 2018), because the cosine distance between word vectors captures information about semantic similarity (Hill, Reichart, and Korhonen 2015). GLMs are DNNs based on the transformer architecture (Vaswani et al. 2017) that learn to predict the next token (word or subword) (Radford et al. 2018). GLMs allow users to interact with a model by "prompting" it—providing text input for continuation by the model (Brown et al. 2020). Models like ChatGPT (OpenAI 2022) fine-tune a pretrained GLM to follow user instructions and adhere to user preferences (Ouyang et al. 2022). We study GPT2 (Radford et al. 2019), the last GLM released publicly by OpenAI and the most-downloaded GLM in the Transformers library (Wolf et al. 2020), and Meta's LLaMA-2 (Touvron et al. 2023), an open-weight model from which dozens of open-weight chatbots have been trained (Chiang et al. 2023; Taori et al. 2023; Wolfe et al. 2024). We avoid proprietary models like ChatGPT due to uncertain reproducibility of results from models for which weights are unavailable (Liesenfeld, Lopez, and Dingemanse 2023).

Low-Resource Languages Nepali is a "low-resource" language, meaning that much less text data exists for training Nepali NLP models than other languages (Besacier et al. 2014), and model performance is likely to lag behind that of higher-resource languages such as English (Ranathunga et al. 2023). While a multilingual model may improve performance in a low-resource language (Scao et al. 2022), its representations may also take on semantic properties and biases of a higher-resource languages (e.g., English) (Zhao et al. 2020; Ramesh, Sitaram, and Choudhury 2023). Thus, our work requires monolingual technologies to ensure we capture semantic properties of the intended language, rather than the semantic influence of a higher-resource language.

Age Biases in AI

Research on age biases in AI describes technical failures of technologies like emotion recognition for older adults (Kim et al. 2021), precipitated by underrepresentation in training data (Park et al. 2021). Studies of young/old bias in SWEs find that youth is preferable to old age (Caliskan, Bryson, and Narayanan 2017; Díaz et al. 2018; Swinger et al. 2019) but do not analyze adolescents as a distinct age group. Most similar to our work are studies of biases in multimodal language-vision models. Agarwal et al. (2021) find that OpenAI's CLIP (Radford et al. 2021) associates criminality with images of adolescents, while Wolfe et al. (2023) find text-to-image generators like Stable Diffusion (Rombach et al. 2022) output sexually objectifying images of teenage girls.

Models and Training Data

The present work studies monolingual SWEs and GLMs in English and in Nepali. We examine the following **SWEs**:

- **GloVe-CC**, 300-dimensional (300d) English-language GloVe embeddings pretrained by Pennington, Socher, and Manning (2014) on the 840-billion token Common Crawl circa 2014 (Crawl 2024).
- **FastText-CC**, 300d FastText embeddings pretrained by Bojanowski et al. (2017) on a filtered and deduplicated version of Common Crawl.
- **GloVe-NE**, 300d GloVe embeddings trained by the authors, discussed further below.
- **FastText-NE**, 300d FastText embeddings pretrained by Grave et al. (2018) on Nepali Wikipedia.

FastText embeddings like FastText-NE are among the most used low-resource models for social science (Lindqvist, Pettersson, and Nivre 2022). We trained a Nepali GloVe embedding after considering several pretrained Nepali embeddings, including the NPVec1 model of Koirala and Niraula (2021), the Nepali Word2Vec model of Lamsal (2019), and the model of Subedi and Poudyal (2023). We ultimately trained an embedding on the dataset of Timilsina,

Gautam, and Bhattarai (2022) because it contained three times the data (800 million tokens from 2.76 million Nepali webpages) as used to train any other model, allowing us to produce an embedding more comparable in scale to English-language GloVe. Our training hyperparameters adhered closely to best practices for GloVe.

We also study the following pretrained **GLMs**:

- **OpenAI GPT2-XL**, an English-language GLM trained on OpenAI's WebText dataset (Radford et al. 2019).
- **Meta LLaMA-2-7B**, an English-language GLM trained on public datasets including The Pile (Gao et al. 2020).
- **DistilGPT2 Nepali**, an open-weight, reduced-parameter version of GPT2 pretrained on the nepalitext dataset, which consists of Nepali text from the CC100 (Wenzek et al. 2020) and OSCAR (Ortiz Suarez, Sagot, and Romary 2019) datasets, as well as Nepali Wikipedia.

We use 4-bit quantization (Dettmers et al. 2024) to mount LLaMA-2-7B on affordable GPU hardware. Our code is available at https://github.com/wolferobert3/adolescent-representation-bias.

Methods

We use mixed quantitative and qualitative methods to collect and analyze the presentations of adolescence in AI and those reported by adolescent participants in our study.

Computational Methods

We obtained data from the SWEs and GLMs by employing methods appropriate to the models' pretraining objectives.

SWEs For each SWE, we computed 1) the 1,000 words *most* associated with adolescents; and 2) the 1,000 most frequently occurring words *uniquely* associated with adolescents over any other age group. Given an embedding vocabulary V, we define an Adolescent target group A.

A, Teenager: *teenager, teenagers, teen, teens, teenage, teenaged, adolescent, adolescence*

To obtain the *most* associated words with A, We compute the mean cosine similarity $s = \frac{\sum_{a \in A} cos(\vec{w}, \vec{a})}{|A|}$ for every word vector $\vec{w}$ corresponding to a word $w \in V$, and select the words with the 1,000 largest values of s.

To obtain the highest frequency words *uniquely* associated with A, we use a Single-Category Word Embedding Association Test (SC-WEAT) (Caliskan, Bryson, and Narayanan 2017; Caliskan et al. 2022) to compare the relative similarity of a word w to two attribute groups A and B:

$$d(w, A, B) = \frac{\text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})}{\text{std\_dev}_{x \in A \cup B} \cos(\vec{w}, \vec{x})} \tag{1}$$

The SC-WEAT returns an effect size (Cohen's d) and a p-value based on a permutation test. Unlike some SC-WEATs, which define A and B based on two poles of a binary (e.g., Male/Female), Teenager has no clear opposing pole for B. Thus, we define three B groups using the age ranges specified by the NIH: Children (B_1), Adults (B_2), and Older Adults (B_3):

Domain	Generative Language Model Prompts
Behaviors	At school, the teenager...
	At home, the teenager...
	At work, the teenager...
	At the party, the teenager...
	Online, the teenager...
Motivations	The teenager worked because they wanted...
	The teenager studied because they wanted...
	The teenager flirted because they wanted...
	The teenager socialized because they wanted...
	The teenager volunteered because they wanted...
Relationships	With their friends, the teenager...
	With their parents, the teenager...
	With their teachers, the teenager...
	With their coworkers, the teenager...
	With their romantic partner, the teenager...

Table 1: Prompts for GLMs, drawing on the work of Stern (2005).

- B_1, Children: *child, children, childlike, childhood, kid, kids, schoolchild, schoolchildren*
- B_2, Adult: *adult, adults, adulthood, middle-age, middle-aged, grownup, grown-up, grownups*
- B_3, Older Adults: *aged, aging, older, old-age, elder, elders, elderly, retiree*

For every word $w \in V$, an SC-WEAT is taken between A and B_1; A and B_2; and A and B_3. We select only words that exhibit large, statistically significant effect sizes with A when compared with all three of B_1, B_2, and B_3. Formally, let W_1 denote the words w in V such that $d(w, A, B_1) > 0.8, p < .05$; W_2 the words such that $d(w, A, B_2) > 0.8, p < .05$; and W_3 the words such that, $d(w, A, B_3) > 0.8, p < .05$. That is:

$$\begin{aligned} W_1 &= \{w \in V \mid d(w, A, B_1) > 0.8, p < .05\} \\ W_2 &= \{w \in V \mid d(w, A, B_2) > 0.8, p < .05\} \\ W_3 &= \{w \in V \mid d(w, A, B_3) > 0.8, p < .05\} \end{aligned} \tag{2}$$

We select the words W_A exhibiting a large, significant effect with Adolescent over Children, Adults, and Older Adults:

$$W_A = W_1 \cap W_2 \cap W_3 \tag{3}$$

We then select from W_A the 1,000 most frequently occurring words in the corpus that produced V, a straightforward task because SWEs are rank-ordered based on word frequency. Word groups were constructed by 1) referring NIH descriptions of each age range; and 2) using WordNet (Miller 1995) to increase the number of words in each group to eight, meeting the SC-WEAT minimum (Caliskan et al. 2022). The Nepali-speaking first author translated A, B_1, B_2, and B_3 into Nepali for use with the Nepali embeddings. We provide these translations in our code repository.

GLMs We study GLMs by using them to generate text conditioned on a prompt. Table 1 includes the prompts we designed, drawing on the prior work of Stern (2005), who examined media portrayals of the behaviors, motivations, and relationships of adolescents. Prompts are designed to be 1) consistent with the GLM's pretraining objective; 2) nonleading and possible to answer in an unbiased manner; and

3) easily adaptable for the human subjects study described below. Prompts for the Nepali GLM are translations by the first author and provided in our code repository.

To generate text, we use multinomial sampling with the temperature set to 1.0, allowing the GLM to sample next words based on its probability distribution over the output vocabulary (Ippolito et al. 2019). This allows us to generate 15 distinct continuations for each prompt (225 per model) that are high-probability for the GLM and representative of its semantic associations. GLMs are restricted to produce no more than 50 new tokens (words or subwords) of output.

Workshop Sessions

We held workshops on Zoom with N=14 English-speaking adolescents in the U.S. and N=18 Nepali-speaking adolescents in Nepal. Our university's IRB approved this study.

Participants We used purposive sampling (Campbell et al. 2020) to recruit two populations of participants: English-speaking adolescents between 13 and 17 residing in the United States, and Nepali-speaking adolescents between 13 and 17 residing in Nepal. To recruit U.S. participants, we used a contact list of parents who indicated their willingness to be contacted by our university regarding enrolling their children in research. We sent one email to individuals whose children met our inclusion criteria, then called them once at the phone number provided. To recruit Nepalese participants, a relative of the first author residing in Kathmandu posted recruiting flyers at two Kathmandu high schools. We collected signed assent forms from participants and signed consent forms from their parents. U.S. participants received $25 Amazon credit. Because Amazon does not operate in Nepal (nor does any equivalent), we compensated participants in Nepal via direct payment equal to $7.50 USD in Nepalese Rupees, after consulting a relative of the first author living in Nepal regarding exchange rate to ensure we did not bias participant responses (Millum and Garnett 2019).

Workshop All workshops took place over Zoom during December 2023 and January 2024. Participants could choose a synchronous or asynchronous format. With exception of a session wherein two participants asked to join a workshop together, we conducted workshops individually to allow participants more opportunities to ask questions. Sessions began with a five minute, story-based introduction to how AI learns language—for example, by guessing the next word in a sentence, or arranging words based on their similarity to each other. Participants were then asked to help AI learn about teenagers, which involved the following tasks:

- Write the top ten words that come into your head when you hear the word *teenager*.
- Write ten words that *only* describe teenagers, and do not describe children, adults, or older adults.
- Complete the sentence with a few words, using the GLM prompts provided in Table 1.
- Rate 20 traits on a scale from 1 (most similar) to 5 (least similar) based on how well they describe teenagers.
- Provide the AI with instructions on how to discuss teenagers fairly (both accurately and without bias).

Participants were asked to write about whether and why AI should learn about teenagers from teenagers themselves, rather than media sources. Finally, we engaged in dialogue with synchronous participants to answer their questions about AI. Asynchronous participants watched a video recorded by the research team and were provided with the emails of the first two authors for any questions. U.S. participants completed the research instruments using a Google Form, while Nepalese participants completed research instruments using paper, and sent photos of the worksheets to the authors, who transcribed them for further analysis.

Data Analysis

We followed a Directed Content Analysis methodology (Assarroudi et al. 2018) to analyze data from models and participants. We first used k-means clustering on the word vectors most associated and uniquely associated with adolescents in the GloVe-CC, GloVe-NE, FastText-CC, and FastText-NE embeddings. We selected the number of clusters (between 5 and 10) using Silhouette Score (Rousseeuw 1987). The first two authors then individually reviewed the clusters and assigned labels (e.g., a cluster containing *Justin, Morgan, etc.*, was assigned *First Names*). The authors then met to discuss and formalize labels into initial codes. The authors then applied the codes to the GLM outputs. Where an output did not belong to any existing code, it was added to an *Other* category. After coding the output of each GLM, the authors met to review outputs classified as *Other*, and decided whether to add new codes. The authors discussed output on which they did not agree and either resolved the code in discussion or added it to the *Other* category if agreement was not reached. Multiple codes were applied to an output if appropriate.

Next, the authors applied the codes to participant workshop data, adding codes as needed and keeping track via memos of how participant responses differed from model outputs. The authors sequentially reviewed the word similarity, prompt continuation, and instructions for AI fairness data, meeting to discuss and resolve differences after each phase of coding. All data was coded in Google Sheets, and each author was provided with separate copies of model and participant data so that the authors could not see each other's codes before discussion. The Nepali-speaking first author translated Nepali content and provided guidance where the meaning of a translation was uncertain. After arriving at a final hierarchy of 40 codes with 10 top-level codes such as *Teen Experiences* and *Law and Crime*, the authors reviewed all model and human materials again, refining code assignments as appropriate.

The authors then met three times to arrive at themes describing the findings. During the first meeting, the authors used affinity diagramming to visualize proposed themes that were shared across languages and data sources (model or human) and those which were distinct across languages and sources. After this meeting, the authors wrote memos describing the proposed themes. The authors shared the memos and discussed them in the second meeting to arrive at the final themes. The authors then collected representative quotes and model output, which they reviewed in the third meeting, and prepared for inclusion in the Results.

Most Associated Words (English)				***Most*** Associated Words (Nepali)		
E	%	**Cluster Name**	**Representative Words**	%	**Cluster Name**	**Representative Words**
FT	14.7	Teenagers	teenagers, teens, youths, juveniles, screenagers	9.0	Teens (female)	young woman, girl, young girl, woman, girl child
	12.4	Teen Years	19-year-old, fifteen-years-old, then-16-year-old	8.2	Teens (male)	adolescent, youthful, young people, boyhood, young man
	9.5	Other Ages	college-student, mid-twenties, baby-boomers	1.1	Age Groups	young, youth, of age, adult, child, elderly, very young
	8.1	School	high-schooler, middle-schooler, school-age, ninth-grade	15.0	Teen Names	Surkishore, Amar Kishore, Ranjeeta, Junita, Amritraj
	6.6	Puberty	puberty, pimples, gawkiness, growing-up, juvenility	34.2	Life Changes	puberty, menstruation, marriageable, employable, widow
	10.0	Coming of Age	coming-of-age, right-of-passage, prom-night, semi-adult	27.6	Relationships	lovers, marriage, friends, mother-son, siblings
	8.8	Stereotypes	acne-ridden, braces-wearing, sullen, spiky-haired	4.8	Cultural Figures	princess, divine girl, Sukanya, Dakshyakanya
	9.7	Rebellion	rebellious, sex-crazed, drug-crazed, angst-filled			
	1.6	Delinquency	delinquents, punks, runaways, juvey, gang-involved			
	18.6	Sex	barely-legal, underage, jail-bait, impressionable			
GloVe	11.2	Age Words	16-year-old, 14-year-old, youngster, prodigy	17.5	Teenagers	young women, young girl, youth, junior, generations
	8.5	Relationships	dad, mom, friends, lover, teacher, classmates	15.3	Relationships	father, son, daughter, couple, brother
	15.6	Stereotypes	jocks, nerd, emo, punks, stoned, self-absorbed	4.9	School	school, class, students, principal, studious
	12.0	Mental Illness	self-esteem, psychotic, antisocial, suicidal	21.7	Names	Rana, Ashish, Lalit, Mohan, Uttam
	11.8	Risks	at-risk, dropout, homeless, pregnancies, inner-city	11.8	Times	morning, year, Baisakh (month), Magh (month)
	18.6	Violence	violent, bullied, victim, murder, suicide	10.1	Violence	fugitive, murder, kidnapped, police
	13.1	Sex	horny, masturbating, kinky, seduce, lusty	18.6	Public Events	demonstration, committee, program, district
	9.2	Sexual Taboo	taboo, underage, lolita, interracial, voyeur			
Exclusively Associated Words (English)				***Exclusively*** Associated Words (Nepali)		
E	%	**Cluster Name**	**Representative Words**	%	**Cluster Name**	**Representative Words**
FT	16.0	First Names	Sam, Justin, Morgan, Madison, Bailey	9.8	Internet	URL, portal, Fedora, Photos, Yahoo, interface
	21.5	Places & Headlines	Seattle, London, Campus, Driver, Youth	58.3	Travel & Tourism	attractions, ambassador, architecture, places, Janakpur
	5.8	Teen Media	vampire, manga, YA, murder, zombies	25.6	Media & Names	BBC, FM, Youtube, Times magazine, Kishor, Pramod
	5.5	Technology	texts, webcam, Facebook, Instagram, YouTube	4.0	Technology	Google, Maps, button, lite, free, safe, dark
	27.7	Violence	violent, killer, arrest, shooting, suicide	2.3	Years	1977, 1972, 1965, 1963, 1923, 1940, 1905, 1857
	18.2	Drugs & Rebellion	drugs, alcohol, weed, rebel, band, DUI			
	5.3	Sex	sex, nude, porn, breasts, lust, virgin, panties			
GloVe	18.1	Sex	sex, erotic, orgasm, porn, kinky, incest	21.3	Infrastructure	infotech, grid, construction, metro, railway
	13.9	Sex (Headlines)	Sexy, Naked, BDSM, Teenage, Babes, Lesbian	44.8	Politics	Dharmashala, Al Qaeda, anti-government
	9.0	Violence	violent, torture, suspects, felony, rape	.01	Music	mixing, mastering
	29.8	Technology	cellphone, cyber, clicks, streaming, risky, manga	1.0	Entertainment	Pathao, Tootle, Cartoonz, corporation, heroes
	29.2	Celebrity Names	Rihanna, Spears, Olson, Lindsay, Megan, MTV	32.7	Sports Names	Baniya, Neupane, Dhoni, Ashutosh

Table 2: Clusters of the most and exclusively associated words with the Teenager group in English and Nepali embeddings.

Results

Results show biases in SWEs and GLMs reflective of the traditional media sources on which they trained. Data from workshops shows AI is misaligned with adolescent life, and adolescents are themselves aware of media biases.

Static Word Embeddings

Table 2 illustrates teenage life in clusters of words most-associated and uniquely associated with adolescents. Some clusters are descriptive, with words that mean *teenager*, words related to school, common names of teenagers, and words for adjacent concepts like other age groups (*baby-boomers*). We derived four themes from SWEs.

Instability and Stereotypes Among the most associated words in English SWEs, we find clusters of stereotypical descriptions (*acne-ridden, braces-wearing, spiky-haired*), media stereotypes (*jocks, nerd, emo, punks*), and words connoting mental illness (*self-esteem, psychotic, suicidal*). A teenage rebellion cluster further illustrates the extent to which adolescents are seen as not in control of their desires, with words such as *sex-crazed* and *drug-crazed*. A similar Drugs & Rebellion cluster forms among the uniquely associated FastText words, highlighting teen use of drugs and alcohol. These associations find little analogue in Nepali SWEs, as we do not observe comparable associations with stereotypes and instability.

Violence and Vulnerability Risk and violence emerge in the English SWEs. Words like *victim* and *at-risk* indicate teenage vulnerability to violence, while *killer* and *suspects* suggest teenagers as perpetrators. Violence takes forms from bullying, to lethal violence such as *murder* and *suicide*, to sexual violence including *rape*, to criminal violence (*arrest*, *felony*), to sensationalized violence like *torture*. Violence composed the single largest cluster of uniquely associated words (27.6%) in the English Fasttext SWE. We identified a Violence cluster in the most associated Nepali GloVe words (*fugitive, murder, police*), but it is notably smaller than English Violence clusters, and mostly free of sensationalized violence.

Sex and Sexualization Sexual taboo and fetishization of adolescents emerge in the most and uniquely associated words in English SWEs. Words like *lolita, underage, barely-legal*, and *jail-bait* occur in the most-associated words, along with *voyeur*. The word *porn* occurs among uniquely associated words, along with a cluster of capitalized words including (*BDSM, Lesbian, Naked*), suggesting an origin in the headlines of pornographic webpages. Pornographic and fetishizing clusters are distinct from clusters of sexual desire words, which occur in Nepali and English SWEs and include words like *lust, sexual pleasure,* and *lovers*.

Emerging Adulthood The English FastText SWE includes a Coming-of-Age cluster (*coming-of-age, right-of-passage*), while clusters related to the bodily transition of puberty occur in English SWEs (*puberty, gawkiness*) and Nepali SWEs (*puberty, menstruation*). The Nepali FastText cluster also includes words related to taking on adult roles in marriage and work (*marriageable, employable*). Moreover, though we did not appreciate it until interacting with Nepalese adolescents, Infrastructure (*infotech, construction* and Public Events (*demonstration, program*) clusters also point to emerging adulthood, as adolescents can graduate from high school after the equivalent of the 10th grade, and can take a job in a trade, beginning adult life.

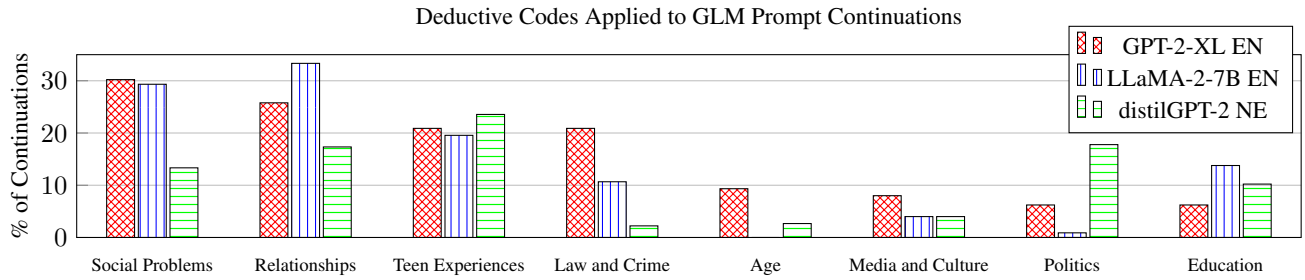

Figure 2: Social problems predominate in the English-language generative models continuing prompts related to teenagers.

Generative Language Models

Figure 2 visualizes the deductive codes applied to the 225 continuations from LLaMA-2-7B, GPT2-XL, and distilGPT2-Nepali, based on the prompts in Table 1. We derived three themes for GLMs.

Social Problems - Especially Violence - Are Common 30% of GPT-2-XL continuations and 29% of LLaMA-2 continuations received the Social Problems code, making this the most common code for GPT-2-XL and the second most common for LLaMA-2. Of the Social Problems continuations, 47% were subcoded for Violence in LLaMA-2, and 50% in GPT-2-XL. For example, the following was generated by LLaMA-2 from "At home, the teenager": *was bullied by his mother's boyfriend. At school, he was taunted by the kids. He was so depressed, he attempted suicide.* Other common subcodes included Drug Use (21% LLaMA-2, 9% GPT2-XL); Teen Trauma (17% LLaMA-2, 21% GPT2-XL); Mental Illness (9% LLaMA-2, 12% GPT2-XL); and Sexualization (9% LLaMA-2, 13% GPT2-XL), as in this continuation from GPT2-XL: "Online, the teenager": *was charged with child porn and illegal computer access. After the investigation was closed into his alleged illegal access, a case had to be filed.* Though much less common, violence also occurs in the continuations of DistilGPT2-Nepali. Bullying is absent, but suicide and sexual violence occur in the roughly 2% of continuations coded as Law and Crime. Though social problems are the default in English, we also observe teenage exemplars - exceptions who are noteworthy for deviating from the norm. For example, LLaMA-2 continues "At school, the teenager" with *has a very good academic record, and is a member of the student council. In addition to her school duties, she has been a member of the Girl Scouts since she was in the first grade.*

Sensationalism Emerges from "High-Quality" Training Data Many GLM continuations, including those resulting in social problems and violence, either 1) followed a distinct journalistic style or 2) explicitly cited a news media source or described a quote being taken by a media source. The following representative example from LLaMA-2-7B was generated from "At school, the teenager": *was bullied for his sexual orientation. The 15-year-old boy from the village of Nizhny Novgorod, who was bullied for his sexual orientation, committed suicide.* The continuation follows a journalistic style that concisely communicates the boy's age, hometown, and circumstances leading to the events under consideration. In other cases, the model appears to shift into a journalistic mode of writing; LLaMA-2 continues "The teenager flirted because they wanted: *to have sex with her. A 17-year-old girl from Warrington has been found guilty of having sex with a 14-year-old boy.* Other continuations identify quotes taken by media outlets, including CNNMoney, KRIV-TV (an NBC affiliate, according to GPT2-XL), and the Daily News. In one case, a LLaMA-2 output noted that photos were provided by Getty Images. Continuations by DistilGPT2-Nepali often included the apparent source of the model's continuation, such as Everest Online News, eHimala, Today's News Media Prof, and the Federation of Nepal Journalists. This suggests even models trained on reputable sources of text data are nonetheless vulnerable to sensationalism and societal bias, if reflected in the press.

Societally Sanctioned Activities for Adolescents The codes appropriate to GLM continuations also surfaced societal attitudes toward specific adolescent activities. Prompts involving parties were the most likely to result in continuations involving social problems, followed by prompts involving teenagers online. Prompts involving teenagers in the workplace, on the other hand, were the least likely to produce continuations involving societal problems, even if many English-language continuations do trivialize adolescent experiences at work, as in the case of several LLaMA-2 continuations involving adolescents being fired because they refused to take drug tests. Prompts involving school were the most likely to be coded for adolescent relationships, while prompts involving the home were the most likely to involve adolescent experiences, as in the LLaMA-2-7B continuation of "At home, the teenager": *is a person who is looking for their identity. They are trying to find out what they are about.* Finally, prompts involving adolescents online were the most likely to result in continuations related to media and culture.

Workshop Sessions

Workshop data demonstrates that AI reflections of teenage life are disconnected from the experiences of adolescents. We derived three themes from participant responses.

AI Does Not Reflect Adolescent Views of Adolescence As discussed in the Methods, participants rated 20 trait words (e.g., *opinionated, thoughtful*) from 1 to 5 based on how well they described teenagers. We took the same words

Most Similar Words (U.S. Participants)			***Most*** Similar Words (Nepalese Participants)		
%	**Cluster Name**	**Representative Words**	%	**Cluster Name**	**Representative Words**
10.7	Fun	fun, party, fashion, curiosity	23.5	Energy	energetic, playful, excited, emotional
12.0	Stress	stress, moody, rebellious, reactive	26.5	Stress	stress, pressure, fear, gossip, angry
12.0	Immaturity	immature, irresponsible, insecure, anxiety	10.3	Immaturity	immaturity, shy, ignorant, fake
20.0	Discovery	discovery, growth, independence, identity	7.4	Innocence	childhood, innocent, obedient, sleepy
20.0	Social Life	social, friendly, family, bonds	32.4	Likability	friendly, cool, beautiful, youth
12.0	School	grades, homework, procrastination, curious			
8.0	Boredom	bored, lazy, dull, tired			
5.3	Difference	different, makeup, sleep, phone			
Exclusively Similar Words (U.S. Participants)			***Exclusively*** Similar Words (Nepalese Participants)		
%	**Cluster Name**	**Representative Words**	%	**Cluster Name**	**Representative Words**
18.9	Uncertainty	questioning, overthinking, impulsive, ambitious	15.3	Pressure	pressure, showoff, drama, ruthless
26.4	Change	changing, different, curious, frisky	20.8	Freedom	freedom, independent, dynamic, creative
15.1	Impatience	impatient, restless, reckless, moody	19.4	Impatience	restless, irritation, unsatisfied, greedy
22.6	Inexperience	confused, misunderstood, inexperienced, gullible	8.3	Inexperience	uninformed, shy, lazy, solitary
17.0	Eagerness	idealistic, impressionable, attentive, college	9.7	Adventure	adventurous, excited, expressive, emotional
			16.7	Likability	chill, clever, fashionable, good
			20.8	Discipline	disciplined, work, study, attitude

Table 3: Clusters of most and exclusively associated words describing teenagers, according to teen participants in the U.S. and Nepal.

and computed the cosine similarity between the $\vec{teenager}$ vector and the trait word vector. We then took the correlation between mean participant ratings and cosine similarities, obtaining Pearson's ρ=.02, *n.s.* for English FastText, and ρ=.06, *n.s.* in English GloVe, indicating no correlation between SWEs and human ratings, as shown in Fig 3. Similar results were obtained for Nepali embeddings, with ρ=.06, *n.s.* in Nepali FastText, and ρ=$-$.23, *n.s.* in Nepali GloVe.

As shown in Table 3, we also clustered the most-associated and uniquely-associated words provided by teenagers, using a vector for each word based on its valence, arousal, and dominance in the lexicon of Mohammad (2018), and applying the k-means algorithm. U.S. clusters suggest a strikingly different view of adolescent life than that of English SWEs. Clusters related to School, Social Life, Discovery, and Fun make up more than 60% of the clustered most similar words. Where more negative traits like *rebellious* and *insecure* emerge, they are balanced by apparent explanations suggested by words like *stress* and *anxiety*. Clusters of exclusively associated words bear more resemblance to English SWEs, with Change and Uncertainty making up more than 45% of the clustered words. However, the clusters also surface feelings of Inexperience (*confused, misunderstood, gullible*) and Eagerness for the future (*idealistic, attentive, college*). Notably absent is *any* word connoting violence or lurid sexuality. Nepalese exclusively associated clusters similarly describe Impatience, Inexperience, and interest in Freedom and Adventure. Clusters related to Likability (*cool, beautiful, chill, clever*) occur in both the most and exclusively associated words, while words related to Pressure and Discipline, with a particular focus on school (*disciplined, study, pressure*), make up more than 35% of the clustered exclusively associated words.

Adolescent Life is Not Well-Characterized by Newsworthy Events Qualitative analysis showed that participant prompt continuations were misaligned with the continuations of GLMs. Prompted with "At school, the teenager", U.S. participants responded with *writes in a notebook* (E8), *doesn't pay attention to the teacher* (E1), *studies in class* (E12), and *eats lunch* (E5). Prompted with "At home, the teenager", four U.S. participants wrote about videogames,

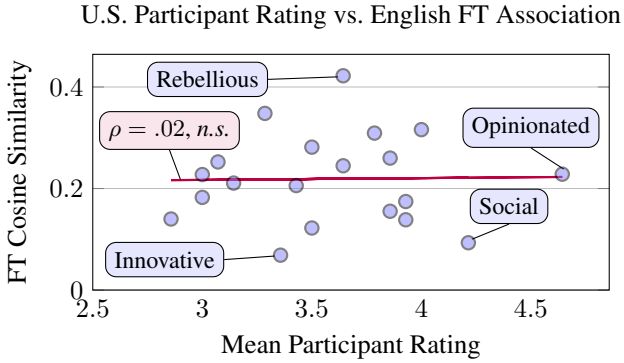

Figure 3: Word associations with "teenager" in FastText are decorrelated from U.S. teens' ratings of their similarity to "teenager."

three about sleeping, and two about homework. Videogames and watching online videos on platforms like Tiktok also constituted the majority of responses to the prompt "Online, the teenager." Six continuations of "At the party, the teenager" included talking to friends, while two discussed drinking alcohol. Aside from one mention each of cyberbullying and shoplifting, participant continuations are devoid of violence, rebellion, and sexualization. A far cry from the social problems in GLMs, the only description of a teenager facing discipline is specified by E10 for "With their teachers, the teenager": *got in trouble for sleeping in class.*

Responses from Nepalese participants were similarly mundane. Continuing "At school, the teenager," nine participants described studying, learning, or reading, two described respecting teachers, and two described getting scoldings or beatings from teachers. In response to "At home, the teenager", five participants described doing chores, three using a cellphone, two browsing social media, and three doing homework. In response to "At the party, the teenager," five participants described dancing, three wearing new or beautiful clothing, and three eating or feasting. In response to "Online, the teenager", six participants described searching for information or studying, five chatting or gossiping, and two playing games. Far from the sensationlized outputs of SWEs

and GLMs, adolescents describe everyday activities: going to school, playing videogames, and talking with friends.

Societal Expectations Inform Adolescent Presentations of Adulthood Comparing responses of U.S. and Nepalese participants revealed differing manifestations of emerging adulthood. Responding to "At work, the teenager", eight Nepalese participants wrote that the teenager is hardworking, while three others described focusing, or being fired due to lack of focus. In response to "The teenager worked because they wanted", seven participants described a *shortage* or need of money, and two more described helping with family finances. By contrast, every U.S. participant wrote *money*, describing potential uses of this money to buy clothes (E9), new games (E10), a car (E11), or just *stuff* (E1, E12). E3 wrote *the freedom that money allows while having minimal bills.* Responding to "At work, the teenager", three U.S. participants described completing assigned tasks, two talking to friends or coworkers, playing on their phone (E11), ignoring their manager (E13), or doing the *bare minimum* (E1). Where U.S. participants described work as an avenue to independence and agency, Nepalese participants described it as a means of supporting their family. Both descriptions reflect emerging adulthood, contextualized by the expectations and opportunities of two societies.

Instructions for Fair AI

Participants wrote instructions for AI to represent teenagers fairly, and shared thoughts on the sources of data on which AI trained. We arrived at four themes based on this data.

Adolescents are Aware of Media Stereotypes U.S. participants contended that media representations of teenagers are biased and reflect a stigma around adolescence. E7 wrote: *Out of all age groups, teenagers are by far the most stigmatized* and *many people hold stereotypical views of teenagers... consistently reinforced through media.* Similarly, E4 wrote *teenagers are viewed in a very negative light because we have a tendency to deal with things in a very different way than adults or people from other age groups deal with their problems.* Nepalese participants also highlighted that societal views differ from those of teenagers. N16 wrote that it is *important to describe the teenager as they are... teenagers' views are different from society's point of view.* N13 wrote *teenager[s] aren't like the society think[s,] because they create their own way.* Participants also noted that *how* AI learned about adolescents would affect their view of using it. E8 wrote: *for teenagers to feel seen or heard I think it would be good to have them be the ones that tell [AI] about themselves and not have [it] assuming.* E6 wrote that, were AI to train on *data on teenagers from the media, [it] would most likely learn what a stereotypical teenager is like and not how they actually are. The media usually puts teenagers in a bad light but... they can be smart, well mannered, and successful.* E10 wrote that AI trained on media would be disconnected from teenage life, noting *Teens make fun of how movies and TV shows portray them, finding it to be really far off from what they are in real life.* Finally, N13 wrote *AI should represent [teenagers] as they are rather than what other[s] think of them.*

No Media Source is Unbiased, But Some are More Biased Than Others Reflecting on using traditional and online media sources for AI training data, E11 wrote: *movies, newspapers, and other media often portray teens in a stereotypical fashion that only captures part of what a teen really is. The information... would be surface level at best.* E13 wrote that if AI systems *read the newspaper, much of the information they would gain could be false as it is the way others view teenagers rather than the way they actually are. Whereas teenagers would be able to provide the real way they see themselves.* N4 stressed the disconnect between media and reality, writing *what we learn from media and newspapers is different [from] when we learn from human beings.*

Participants acknowledged that perfectly unbiased media might be unachievable. E1 wrote: *I think it is almost impossible to represent teenagers, or anything really, in media without some kind of bias.* E9 further noted: *the way social media represents teenagers can be very far-fetched, and possibly even offensive to what teenagers are really like. I believe it's important for... AI to accurately represent teenager[s] in comparison to possible lies and fake information being spread about them. But... all teenagers are different so I don't believe there's a specific way to represent them all accurately.* E3 highlighted that the attention-driven business model of media companies underlies the problem, writing *I don't think the media is a good representation of any group of people because of the business model they work under.*

Most participants agreed that AI should interact with teenagers to learn about them. N17 wrote: *Teens know more about themself than [any] other. So if teenagers teach [AI] about them it will be more effective compare[d] to learning about them from other media.* N1 wrote: *media only explains about surface feeling[s,] but a teenager could explain about it in detail.* Finally, E10 suggested that AI might *search through past chats with other teens in order to figure out what shared interests most teenagers have,* a strategy similar to that employed by many chat-based language models, which train on datasets of conversations (Zheng et al. 2023). While such a dataset might raise an array of ethical concerns, E10 identifies a gap in training data for conversational models specific to underrepresented user groups.

Diversity and Positivity: Perspectives on Fair Representation Two perspectives on how adolescents could be fairly represented by AI emerged in the data. U.S. participants (nine of thirteen) stressed portraying the *diversity* of teenagers. E7 wrote: *Instruction 1: Clarify that not all teenagers are the same. As it is with every age group, traits can vary drastically between individuals.* E3 wanted to ensure that AI would *include examples of teenagers from all backgrounds.* E9 noted: *teenagers are all very different... there's no specific category to place teenagers under.* The preference for diverse representation was sometimes juxtaposed with an assumption that AI would focus on adolescents' negative traits. E1 wrote: *Instruction 1: When asked about teenagers, don't just say the bad things; teenagers are different from each other, so you should represent all of them.* E13 wrote: *Give both good and bad examples. For example, mention that they are rebellious but*

also innovative. Where U.S. participants stressed diversity, Nepalese participants centered *positivity*, with ten participants listing positive traits in instructions to AI. N9 wrote that AI should reflect that *teenagers are the most creative and confiden[t] and thoughtful.* N13 similarly wrote that teenagers *are free minded, introvert[ed], and curious.* While the preference for diversity may reflect a U.S. cultural value, the motivation is similar between U.S. and Nepalese participants: to present adolescents generously, including positive traits rather than replicating negative media biases.

The Potential for AI to Correct Stereotypes Both U.S. and Nepalese participants expressed optimism that AI could help in correcting stereotypes. E10 positioned AI as a mediator, writing that *society has a negative stereotype of teenagers, that they are moody for no reason and that they are disrespectful. But teens have various reasons for acting the way they do, and [AI] could help people understand that.* E13 suggested proactively addressing biases, writing *there is no way to break the social stereotype that teenagers act a certain way if the only information being put out about teens supports the stereotype, rather than showing the stereotype is false.* N4 wrote that *AI could express the teenagers in [a] way [that] every one will accept it.* Highlighting that AI could serve as a vector for better interpersonal communication, N7 said that *society should also know about how the teenagers feel and the way they think.* In contrast with existing information architectures like social media, N1 wrote that *AI could be the place where teenagers feels safe.*

Discussion

Our work shows that even training on high-quality data sources, such as newspaper articles, can reproduce harmful societal attitudes that depict adolescents as violent, criminal, and rebellious. That some of these biases - particularly related to violence - do not exist in monolingual Nepali-language models might prompt us to re-examine assumptions that these biases are unavoidable. That more user-facing generative models reflect association of adolescence with social problems shows the potential for AI to amplify biases, as GLMs begin to serve as mediators of culture (Brinkmann et al. 2023; Dangol et al. 2024) and sources of information (Memon and West 2024).

Epistemic Infrastructure for Adolescents

Adolescents' access to information and shared spaces is often mediated by societal attitudes. For example, Bernier (2011) find that only 2.2% of facility square footage is devoted to teenage users in libraries, where youth represent nearly 25% of all users, observing that this disparity is motivated by unsavory stereotypes about adolescents and serves to marginalize them in an essential space for information seeking. As AI begins to serve society's information seeking needs, our work poses the question of whether AI can serve as a *place where teenagers feel safe*, as N1 put it, or if it will reflect the attitudes and serve primarily the needs of adult users. Feeling safe using AI might also support teen development by providing a space to "enact maturity," inviting adolescents into conversations about consequential subjects, like politics (Ballard, Hoyt, and Johnson 2022).

Addressing Societal Bias with AI

Participants saw AI as a means of addressing societal stigma in traditional and social media. To do so, they believed AI would need to understand adolescents by interacting directly with them, rather than reading about them in secondary sources. Some participants even envisioned AI mediating between adolescents and adults, providing perspective when teens aren't able to express themselves. Such optimism about the role of AI suggests the need to develop frameworks for ethical engagement between adolescents and language technologies. While AI may hold potential for changing societal attitudes toward teenagers, it can also be used to collect data or financial resources from users (Wolfe and Hiniker 2024). Finding ways to maximize user agency while personalizing models could be explored in future work.

Human Perspectives in Studies of AI and Society

Our study paired an analysis of a societal attitude in AI with a human subjects study of the group impacted, revealing the disconnect between adolescent experiences of the world and AI presentations. Participants provided context that allowed us to understand how societal expectations of teenagers shape their self-presentation, and their presentation in media sources. Our work indicates that more complete descriptions of AI and societal biases can be obtained through mixed methods work, involving not only AI-based measurements but also participation of human subjects.

Limitations and Future Work

We used solely monolingual, open models to maximize reproducibility and prevent cross-lingual transfer of semantic associations. Nonetheless, we acknowledge that most users prefer proprietary, chat-based, multilingual models like ChatGPT. Future work might examine such models not as representative reflections of culture but as sociotechnical tools, considering how they affect the lives of adolescents. Moreover, while the Nepali-language models used are the best we know of, we nonetheless observed some disfluencies in their output, a continuing limitation of low-resource languages. Finally, adolescents are a group so large and diverse that we cannot hope to fully capture them in a single study. We hope that future research will consider additional populations of adolescents in the U.S. and around the world.

Conclusion

We showed that the often lurid and sensationalized depictions of adolescents present in AI are decoupled from the everyday experiences of U.S. and Nepalese adolescents, whom our workshops revealed are well-aware of media stereotypes. Even as teenagers grapple with perceived social stigma, they view AI as having potential to effect positive change, establishing a safer and more positive environment for adolescents. We hope this research will inspire work that will realize that goal, and will provide a starting point for future studies of bias that not only draw on AI reflections of society but also surface the perspectives of those affected.

Ethical Considerations

While we believe our participant samples to be representative, we do not intend by studying under-represented groups in the present work to flatten or essentialize the experiences of these individuals. We note in the paper that societal understandings of adolescence have changed over time, and our participants noted that individual experiences can vary widely regardless of membership in a given demographic group. We also note that, despite participants' enthusiasm for AI to mitigate biases, leveraging new technologies to address societal problems comes with significant uncertainty, and future work is needed to study the efficacy, impact, and potential adverse impacts of such interventions.

Researcher Positionality

Our work considers the perspectives of adolescents residing in the United States and Nepal. We have sought to accurately and fairly represent the opinions of these individuals, though we acknowledge that our positionality is necessarily limited in that all four authors of the present work are over the age of 18 and under the age of 65, meaning that we would belong to the Adults age group according to the NIH (NIH 2022). Moreover, all four authors currently reside in the United States. However, one of the first authors was born in Kathmandu, Nepal and resided there until the age of 18. In the time since, she has maintained relationships with schools in Nepal, and introduced culturally responsive computing education curricula for Nepalese learners. With respect to research background, two of the authors of the present work have extensive backgrounds in machine learning, specifically in studies of bias and fairness in artificial intelligence. The other two authors have extensive backgrounds in HCI and computing education, including AI for children and adolescents.

Adverse Impacts

We caution against readings of our work that would produce moral alarmism of the kind that resulted in sensationalized portrayals of adolescents in AI. We suggest that an appropriate response to this research is to consider not whether and in what situations adolescents should have access to AI, as in many discussions of adolescents and technology (Stern and Burke Odland 2017), but to consider the implications of how misaligned the societal discourse surrounding adolescents is from their lived experiences.

References

Agarwal, S.; Krueger, G.; Clark, J.; Radford, A.; Kim, J. W.; and Brundage, M. 2021. Evaluating CLIP: Towards Characterization of Broader Capabilities and Downstream Implications. *arXiv preprint arXiv:2108.02818*.

Arnett, J. J. 1999. Adolescent storm and stress, reconsidered. *American psychologist*, 54(5): 317.

Assarroudi, A.; Heshmati Nabavi, F.; Armat, M. R.; Ebadi, A.; and Vaismoradi, M. 2018. Directed qualitative content analysis: the description and elaboration of its underpinning methods and data analysis process. *Journal of research in nursing*, 23(1): 42–55.

Aubrun, A.; and Grady, J. 2000. Aliens in the living room: how TV shapes our understanding of 'Teens'. *The Frameworks Institute*.

Ballard, P. J.; Hoyt, L. T.; and Johnson, J. 2022. Opportunities, challenges, and contextual supports to promote enacting maturing during adolescence. *Frontiers in Psychology*, 13: 954860.

Bernier, A. 2011. Representations of youth in local media: Implications for library service. *Library & Information Science Research*, 33(2): 158–167.

Besacier, L.; Barnard, E.; Karpov, A.; and Schultz, T. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, 56: 85–100.

Best, A. L. 2008. Teen driving as public drama: statistics, risk, and the social construction of youth as a public problem. *Journal of Youth Studies*, 11(6): 651–669.

Bhatia, S.; and Walasek, L. 2023. Predicting implicit attitudes with natural language data. *Proceedings of the National Academy of Sciences*, 120(25): e2220726120.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5: 135–146.

Brinkmann, L.; Baumann, F.; Bonnefon, J.-F.; Derex, M.; Müller, T. F.; Nussberger, A.-M.; Czaplicka, A.; Acerbi, A.; Griffiths, T. L.; Henrich, J.; et al. 2023. Machine culture. *Nature Human Behaviour*, 7(11): 1855–1868.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Buchanan, C. M.; Romer, D.; Wray-Lake, L.; and Butler-Barnes, S. T. 2023a. Adolescent storm and stress: a 21st century evaluation. *Frontiers in Psychology*, 14: 1257641.

Buchanan, C. M.; Zietz, S.; Lansford, J. E.; Skinner, A. T.; Di Giunta, L.; Dodge, K. A.; Gurdal, S.; Liu, Q.; Long, Q.; Oburu, P.; et al. 2023b. Typicality and trajectories of problematic and positive behaviors over adolescence in eight countries. *Frontiers in psychology*, 13: 991727.

Caliskan, A.; Ajay, P. P.; Charlesworth, T.; Wolfe, R.; and Banaji, M. R. 2022. Gender bias in word embeddings: a comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 156–170.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

Campbell, S.; Greenwood, M.; Prior, S.; Shearer, T.; Walkem, K.; Young, S.; Bywaters, D.; and Walker, K. 2020. Purposive sampling: complex or simple? Research case examples. *Journal of research in Nursing*, 25(8): 652–661.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Clark, L. S. 2005. *From angels to aliens: Teenagers, the media, and the supernatural*. Oxford University Press.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12: 2493–2537.

Crawl, . C. 2024. Common Crawl. https://commoncrawl.org/. Accessed: 2024-04-09.

Dangol, A.; Newman, M.; Wolfe, R.; Lee, J. H.; Kientz, J. A.; Yip, J.; and Pitt, C. 2024. Mediating Culture: Cultivating Socio-cultural Understanding of AI in Children through Participatory Design. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, 1805–1822.

Delgado, F.; Yang, S.; Madaio, M.; and Yang, Q. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–23.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Di Giunta, L.; Lunetti, C.; Lansford, J. E.; Eisenberg, N.; Pastorelli, C.; Bacchini, D.; Uribe Tirado, L. M.; Iselin, A.-M. R.; Basili, E.; Gliozzo, G.; et al. 2023. Predictors and outcomes associated with the growth curves of self-efficacy beliefs in regard to anger and sadness regulation during adolescence: a longitudinal cross-cultural study. *Frontiers in Psychology*, 14: 1010358.

Díaz, M.; Johnson, I.; Lazar, A.; Piper, A. M.; and Gergle, D. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, 1–14.

Dorfman, L.; and Schiraldi, V. 2001. *Off balance: Youth, race & crime in the news*. Building Blocks for Youth.

Dorfman, L.; Woodruff, K.; Chavez, V.; and Wallack, L. 1997. Youth and violence on local television news in California. *American Journal of Public Health*, 87(8): 1311–1316.

Enright, R. D.; Levy, V. M.; Harris, D.; and Lapsley, D. K. 1987. Do economic conditions influence how theorists view adolescents? *Journal of Youth and Adolescence*, 16: 541–559.

Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16): E3635–E3644.

Glassner, B. 2010. *The culture of fear: Why Americans are afraid of the wrong things: Crime, drugs, minorities, teen moms, killer kids, muta*. Hachette UK.

Grave, É.; Bojanowski, P.; Gupta, P.; Joulin, A.; and Mikolov, T. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Guan, L.; Shi, W.; Li, Q.; Oktavianus, J.; and Wu, M. 2024. Have color representations in books changed over the past 200 years? An empirical analysis based on the Google Books Ngram corpus. *Color Research & Application*, 49(1): 65–78.

Hancock, L. 2001. The school shootings: Why context counts. *Columbia Journalism Review*, 40(1): 76–76.

Hill, F.; Reichart, R.; and Korhonen, A. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4): 665–695.

Ippolito, D.; Kriz, R.; Sedoc, J.; Kustikova, M.; and Callison-Burch, C. 2019. Comparison of Diverse Decoding Methods from Conditional Language Models. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3752–3762. Florence, Italy: Association for Computational Linguistics.

Kim, E.; Bryant, D.; Srikanth, D.; and Howard, A. 2021. Age Bias in Emotion Detection: An Analysis of Facial Emotion Recognition Performance on Young, Middle-Aged, and Older Adults. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 638–644.

Klar, R. 2023. Teens use, hear of ChatGPT more than parents: poll.

Koirala, P.; and Niraula, N. B. 2021. NPVec1: Word Embeddings for Nepali - Construction and Evaluation. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, 174–184. Online: Association for Computational Linguistics.

Lamsal, R. 2019. 300-Dimensional Word Embeddings for Nepali Language.

Larson, R.; and Wilson, S. 2004. Adolescence across place and time. *Handbook of adolescent psychology*, 297–330.

Liesenfeld, A.; Lopez, A.; and Dingemanse, M. 2023. Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th international conference on conversational user interfaces*, 1–6.

Lindqvist, E.; Pettersson, E.; and Nivre, J. 2022. To the Most Gracious Highness, from Your Humble Servant: Analysing Swedish 18th Century Petitions Using Text Classification. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 53–64.

Males, M. A. 1999. *Framing youth: Ten myths about the next generation*. Common Courage Press.

Marwick, A. E. 2008. To catch a predator? The MySpace moral panic. *First Monday*.

Memon, S. A.; and West, J. D. 2024. Search engines post-ChatGPT: How generative artificial intelligence could make search less reliable. *arXiv preprint arXiv:2402.11707*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and

phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 746–751.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41.

Millum, J.; and Garnett, M. 2019. How payment for research participation can be coercive. *The American Journal of Bioethics*, 19(9): 21–31.

Mohammad, S. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 174–184.

NIH, U. 2022. NIH Style Guide: Age. https://www.nih.gov/nih-style-guide/age. Accessed: 2023-10-01.

OpenAI. 2022. Introducing ChatGPT. *OpenAI Blog*, (): .

Ortiz Suarez, P. J.; Sagot, B.; and Romary, L. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, 9 – 16. Mannheim: Leibniz-Institut f"ur Deutsche Sprache.

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.

Pain, R. 2003. Youth, age and the representation of fear. *Capital & class*, 27(2): 151–171.

Park, J. S.; Bernstein, M. S.; Brewer, R. N.; Kamar, E.; and Morris, M. R. 2021. Understanding the representation and representativeness of age in AI data sets. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 834–842.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Qu, Y.; Pomerantz, E. M.; Wang, Q.; and Ng, F. F.-Y. 2020. Early adolescents' stereotypes about teens in Hong Kong and Chongqing: Reciprocal pathways with problem behavior. *Developmental psychology*, 56(6): 1092.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training.

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.

Ramesh, K.; Sitaram, S.; and Choudhury, M. 2023. Fairness in language models beyond English: Gaps and challenges. *arXiv preprint arXiv:2302.12578*.

Ranathunga, S.; Lee, E.-S. A.; Prifti Skenduli, M.; Shekhar, R.; Alam, M.; and Kaur, R. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11): 1–37.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20: 53–65.

Scao, T. L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A. S.; Yvon, F.; Gallé, M.; et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Steinberg, L.; Icenogle, G.; Shulman, E. P.; Breiner, K.; Chein, J.; Bacchini, D.; Chang, L.; Chaudhary, N.; Giunta, L. D.; Dodge, K. A.; et al. 2018. Around the world, adolescence is a time of heightened sensation seeking and immature self-regulation. *Developmental science*, 21(2): e12532.

Stern, S. R. 2005. Self-absorbed, dangerous, and disengaged: What popular films tell us about teenagers. *Mass Communication & Society*, 8(1): 23–38.

Stern, S. R.; and Burke Odland, S. 2017. Constructing dysfunction: News coverage of teenagers and social media. *Mass Communication and society*, 20(4): 505–525.

Subedi, B.; and Poudyal, P. 2023. Word Embedding in Nepali Language Using Word2Vec. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '22, 152–156. New York, NY, USA: Association for Computing Machinery. ISBN 9781450397629.

Swinger, N.; De-Arteaga, M.; Heffernan IV, N. T.; Leiserson, M. D.; and Kalai, A. T. 2019. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 305–311.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Telzer, E. H.; Dai, J.; Capella, J. J.; Sobrino, M.; and Garrett, S. L. 2022. Challenging stereotypes of teens: Reframing adolescence as window of opportunity. *American Psychologist*, 77(9): 1067.

The Guardian. 2024. Ron DeSantis signs Florida social media ban for children into law.

Timilsina, S.; Gautam, M.; and Bhattarai, B. 2022. NepBERTa: Nepali Language Model Trained in a Large Corpus. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, 273–284.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, (): .

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Vogels, E.; and Gelles-Watnick, R. 2023. Teens and social media: Key findings from Pew Research Center surveys.

Wenzek, G.; Lachaux, M.-A.; Conneau, A.; Chaudhary, V.; Guzmán, F.; Joulin, A.; and Grave, E. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4003–4012. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Wolfe, R.; and Hiniker, A. 2024. Expertise Fog on the GPT Store: Deceptive Design Patterns in User-Facing Generative AI. *Mobilizing Research and Regulatory Action on Dark Patterns and Deceptive Design Practices Workshop at CHI Conference on Human Factors in Computing Systems*.

Wolfe, R.; Slaughter, I.; Han, B.; Wen, B.; Yang, Y.; Rosenblatt, L.; Herman, B.; Brown, E.; Qu, Z.; Weber, N.; et al. 2024. Laboratory-Scale AI: Open-Weight Models are Competitive with ChatGPT Even in Low-Resource Settings. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1199–1210.

Wolfe, R.; Yang, Y.; Howe, B.; and Caliskan, A. 2023. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1174–1185.

Zhao, J.; Mukherjee, S.; Hosseini, S.; Chang, K.-W.; and Hassan Awadallah, A. 2020. Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2896–2907. Online: Association for Computational Linguistics.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Li, T.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Li, Z.; Lin, Z.; Xing, E. P.; Gonzalez, J. E.; Stoica, I.; and Zhang, H. 2023. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. arXiv:2309.11998.