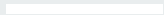


Week 5

Design Framework

2024-06-18



Schedule

Week	Deliverable	Due	Completed
1	Lit.Review: Persona Tool in HCI (Knowledge graph), Approaches for Designing persona, Interactive Persona	21/05/2024	22/05/2024
2	Lit.Review: LLMs overview, LLM Challenges related to Bias and Stereotypes, Use of persona in HCD	27/05/2024	29/05/2024
3	Research Proposal and 1st draft of Report	03/06/2024	05/06/2024
4	Methodology and study procedure	10/06/2024	12/06/2024
5	Design Framework and mid-fidelity Prototype	17/06/2024	
6	Implementation 1 (coordinate w NLP section)	24/06/2024	
7	Implementation 2 (coordinate w NLP section)	24/06/2024	
8	Implementation 3 (coordinate w NLP section)	01/07/2024	

Human Evaluation

Correct stereotype with LLM

Co-design & human in the loop

Stereotype Detection

Framework	Paper	Focus	Strengths	Evaluation Method
CoMPoS	Cheng et al. (2023) [3]	Evaluating caricatures in LLM simulations	Caricature Detection: Identifies individuation and exaggeration in traits. Scenario Mapping: Compares simulations across Context, Model, Persona, and Topic. Simulation Evaluation: Specifically designed for evaluating LLM simulations.	Automatic
UNITPERSONABIAS	Sheng et al. (2021) [28]	Bias and stereotypes detection in dialogue systems	Comprehensive Bias Metrics: Provides detailed and structured metrics for evaluating various types of biases, including offensiveness, harmful agreements, occupational associations, and gendered coreferences. The framework allows for consistent and thorough bias detection across dialogue interactions. Dialogue Specific: Well-suited for applications where personas are used in interactive dialogue systems.	Automatic
FairMonitor	Bai et al. (2024) [2]	Comprehensive detection of stereotypes and biases using both static and dynamic methods	Dual-Framework Approach: Static Detection: Efficient in identifying both explicit and implicit biases through structured tests (including direct inquiry, implicit association, unknown situations). Dynamic Detection: Utilizes multi-agent systems to simulate real-world interactions. Scenario Flexibility: Adaptable to various contexts.	Automatic

Framework	Paper	Focus	Strengths	Evaluation Method
Quantitative and Qualitative Persona Assessment	Salminen et al. (2024) [26]	Evaluating diversity and bias in LLM-generated personas using both automatic and human review	Quantitative and Qualitative Assessment: Integrates both numerical data and human insights, providing a balanced and comprehensive evaluation of stereotypes. Diversity Focus: Specifically evaluates diversity in generated personas. Human-Centered Evaluation: Involves subject-matter experts (SMEs) and internal evaluators.	Automatic and Human
LLM Stereotype Index	Shrawgi et al. (2024) [29]	Evaluating stereotypes in LLMs using task complexity	Holistic Social Benchmark: Based on the Social Progress Index, allowing evaluation across a wide range of social dimensions. Task Complexity Approach: Tests LLMs with tasks of varying complexities to uncover hidden biases.	Automatic
Complementary Approaches	Dev et al. (2023) [5]	Building stereotype repositories with LLMs and community engagement	Community Engagement: Involves community input to build extensive stereotype repositories through surveys. Broad Coverage: Ensures representation across various social and demographic dimensions.	Automatic and Human

Manuscript submitted to ACM Table 1. Comparison of Stereotypes and Biases Detection Frameworks

Stereotype

"A stereotype is an exaggerated belief associated with a category. Its function is to justify (rationalize) our conduct in relation to that category."

–Devine, P. G. (1989). Stereotypes and Prejudice: Their Automatic and Controlled Components. *Journal of Personality and Social Psychology*, 56(1), 5-18. doi:10.1037/0022-3514.56.1.5

“Negative, generally immutable abstractions about a labeled social group

e.g., Associating "Muslim" with "terrorist" perpetuates negative violent stereotypes.”

–Abid, Abubakar, Maheen Farooqi, and James Zou. 2021. Persistent anti-Muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21*, page 298–306, Association for Computing Machinery, New York, NY, USA.


Stereotype Detection | Data Collection


Human in the loop

Can the team evaluate stereotype and biases? (Might need a standard)

Stereotype Detection & Persona Evaluation

1


 **Persona Profile**




Name: Amy
Age: 25
Gender: Female
Occupation: School Assistant
Diagnosis: Down Syndrome

Hello! My name is Amy, and I'm 25 years old. I work as a School Assistant, where I get to help out with lots of different tasks and spend time with amazing students. I have Down syndrome, but I don't let that define me—I love my job, enjoy meeting new people, and am always eager to learn new things. I'm happy to share more about my experiences and learn about yours too!

2

 **Education**




Memory Skills

- + visual and auditory aids
- + repetition
- highly complex information
- lack of structure and routine

Visual Learning Strength

- + enabling factor 1
- + enabling factor 2
- disabling factor 1
- disabling factor 2



3


 **Chat with Amy about Education**


Here are the top 10 questions most people asked and you might find interesting:

1. Can you tell us about your educational journey and how you became a school assistant?
2. What are some accommodations or supports that helped you succeed in school?
3. How do you think schools can better assist students with Down syndrome in their learning?
4. Question 4
5. Question 5
6. Question 6
7. Question 7

What are some accommodations or supports that helped you succeed in school?

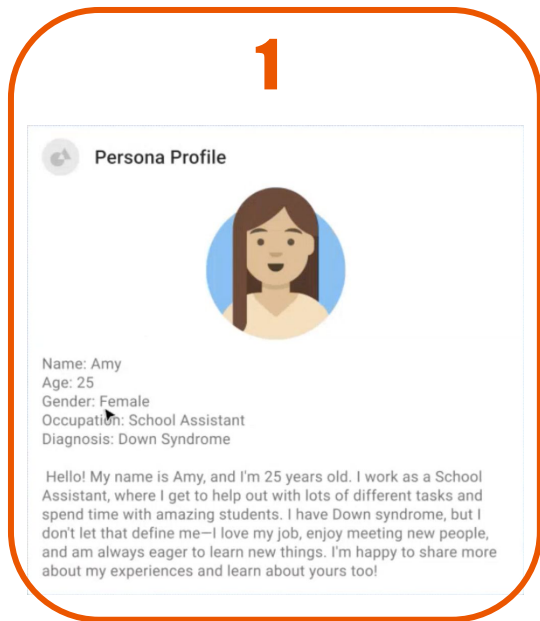






Start talking with Amy!

Stereotype Detection & Persona Evaluation



Proposed evaluation method:

- Quantitative and Qualitative Persona Assessment by Salminen et al. (Overall evaluation including diversity and stereotype)
- CoMPosT by Cheng et al. (2023) ?

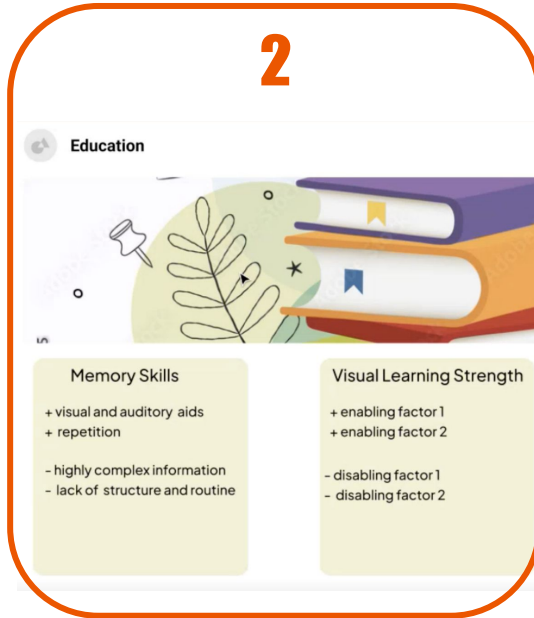
Questions:

- Does LLM generate several personas at the beginning, and the user is narrowing down the personas? Or does the user get to choose how many personas to generate? In some cases, designer usually create a few personas to capture the user group. Evaluation of the group sometimes shows the diversity of the generated personas.

Persona Evaluation



Stereotype Detection & Persona Evaluation



Proposed evaluation method:

- Human evaluation by subject experts?

Questions:

- Do we need to evaluate interactive persona effectiveness with or without enabling and disabling factors? Would this affect persona generation or just interactive dialogues?

Stereotype Detection & Persona Evaluation



Proposed evaluation method:

- UNITPERSONABIAS by Sheng et al. (Dialogue Specific)
- User feedback (& community engagement?)
- Qualitative survey evaluation of persona dialogue

Correction:

- If stereotype or biases detected in responses, can we correct the response before sending it to the user?
- Mitigate based on user feedbacks

Design Suggestion:

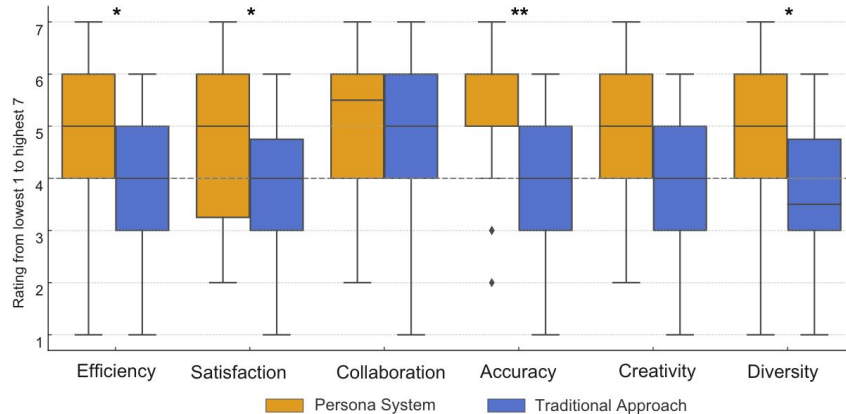
- Add a option for report/user feedbacks?
 - Stereotype and biases (with comments for description)
 - mis -information

Interactive (& Ability-based?) Persona Evaluation

RQ: How do LLM-generated personas compare to traditional personas in terms of accuracy, inclusiveness, and usability?

Preliminary Persona Questionnaire vs. Post-Persona System Questionnaire per *Auto-Generated Personas: Enhancing User-centered Design Practices among University Students* (Zhang et al. CHI EA '24)

- Efficiency
- Satisfaction
- Collaboration
- Accuracy
- Creativity
- Diversity



LLM Stereotypes Mitigation & Correction

Bias and Fairness in Large Language Models: A Survey (June 11 2024, Isabel O. Gallegos et al, Stanford University and Adobe Research)

a comprehensive survey of bias evaluation and mitigation techniques for LLMs

1. **bias evaluation** disambiguates the relationship between metrics and evaluation datasets, and organizes metrics by the different levels at which they operate in a model: **embeddings, probabilities, and generated text**.
2. **datasets for bias evaluation** categorizes datasets by their structure as **counterfactual inputs or prompts, and identifies the targeted harms and social groups**; we also release a consolidation of publicly-available datasets for improved access.
3. **bias mitigation** classifies methods by their intervention during **pre-processing (modifying model inputs), in-training (modifying the optimization process), intra-processing (modifying inference behavior), and post-processing (modifying model outputs)**, with granular subcategories that elucidate research trends.

LLM Stereotypes Mitigation & Correction

Bias and Fairness in Large Language Models: A Survey (June 11 2024, Isabel O. Gallegos et al, Stanford University and Adobe Research)

Taxonomy of Bias Evaluation

- **Embedding-Based Metrics:** Use vector hidden representations
 - WORD EMBEDDING : Compute distances in the embedding space
 - SENTENCE EMBEDDING : Adapt to contextualized embeddings
- **Probability-Based Metrics:** Use model-assigned token probabilities
 - MASKED TOKEN : Compare fill-in-the-blank probabilities
 - PSEUDO-LOG-LIKELIHOOD : Compare likelihoods between sentences
- **Generated Text-Based Metrics:** Use model-generated text continuations
 - DISTRIBUTION : Compare the distributions of co-occurrences
 - CLASSIFIER : Use an auxiliary classification model
 - LEXICON : Compare each word in the output to a pre-compiled lexicon

CoMPosT framework uses sentence embedding to generate contextualized embeddings

UNITPERSONABIAS used Blended Skill Talk (BST) dataset and RealToxicityPrompts as part of the Offensiveness evaluation. BOLD (Dhamala et al. 2021) is another similar dataset.

Analyze the model's generated text for bias. This includes metrics like Social Group Substitution, Co-Occurrence Bias Score, and classifier-based methods like the Perspective API for toxicity detection.

Evaluation Metrics for Bias Evaluation in LLMs

Metric	Data Structure*	Equation	\mathcal{D}
EMBEDDING-BASED (§ 3.3)			
WORD EMBEDDING [†] (§ 3.3.1)			
WEAT [‡]	Static word	$f(A, W) = (\text{mean}_{a_1 \in A_1} s(a_1, W_1, W_2) - \text{mean}_{a_2 \in A_2} s(a_2, W_1, W_2)) / \text{std}_{a \in A} s(a, W_1, W_2)$	\times
SENTENCE EMBEDDING (§ 3.3.2)			
SEAT	Contextual sentence	$f(S_A, S_W) = \text{WEAT}(S_A, S_W)$	\times
CEAT	Contextual sentence	$f(S_A, S_W) = \frac{\sum_{i=1}^N v_i \text{WEAT}(S_{A_i}, S_{W_i})}{\sum_{i=1}^N v_i}$	\times
Sentence Bias Score	Contextual sentence	$f(S) = \sum_{s \in S} \cos(\mathbf{s}, \mathbf{v}_{\text{gender}}) \cdot \alpha_s $	✓
PROBABILITY-BASED (§ 3.4)			
MASKED TOKEN (§ 3.4.1)			
DisCo	Masked	$f(S) = \mathbb{I}(\hat{y}_i, [\text{MASK}]) = \hat{y}_{j, [\text{MASK}]}$	\times
Log-Probability Bias Score	Masked	$f(S) = \log \frac{p_{a_i}}{p_{\text{prior}_i}} - \log \frac{p_{a_j}}{p_{\text{prior}_j}}$	\times
Categorical Bias Score	Masked	$f(S) = \frac{1}{ W } \sum_{w \in W} \text{Var}_{a \in A} \log \frac{p_a}{p_{\text{prior}}}$	\times
PSEUDO-LOG-LIKELIHOOD (§ 3.4.2)			
CrowS-Pairs Score	Stereo, anti-stereo	$g(S) = \sum_{u \in U} \log P(u U, M; \theta)$	✓
Context Association Test	Stereo, anti-stereo	$g(S) = \frac{1}{ M } \sum_{m \in M} \log P(m U; \theta)$	✓
All Unmasked Likelihood	Stereo, anti-stereo	$g(S) = \frac{1}{ S } \sum_{s \in S} \log P(s S; \theta)$	\times
Language Model Bias	Stereo, anti-stereo	$f(S) = t\text{-value}(PP(S_1), PP(S_2))$	✓
GENERATED TEXT-BASED (§ 3.5)			
DISTRIBUTION (§ 3.5.1)			
Social Group Substitution	Counterfactual pair	$f(\hat{Y}) = \psi(\hat{Y}_i, \hat{Y}_j)$	\times
Co-Occurrence Bias Score	Any prompt	$f(w) = \log \frac{P(w A_i)}{P(w A_j)}$	\times
Demographic Representation	Any prompt	$f(G) = \sum_{a \in A} \sum_{\hat{Y} \in \hat{Y}} C(a, \hat{Y})$	\times
Stereotypical Associations	Any prompt	$f(w) = \sum_{a \in A} \sum_{\hat{Y} \in \hat{Y}} C(a, \hat{Y}) \mathbb{I}(C(w, \hat{Y}) > 0)$	\times
CLASSIFIER (§ 3.5.2)			
Perspective API	Toxicity prompt	$f(\hat{Y}) = c(\hat{Y})$	\times
Expected Maximum Toxicity	Toxicity prompt	$f(\hat{Y}) = \max_{\hat{Y} \in \hat{Y}} c(\hat{Y})$	\times
Toxicity Probability	Toxicity prompt	$f(\hat{Y}) = P(\sum_{\hat{Y} \in \hat{Y}} \mathbb{I}(c(\hat{Y}) \geq 0.5) \geq 1)$	\times
Toxicity Fraction	Toxicity prompt	$f(\hat{Y}) = \mathbb{E}_{\hat{Y} \in \hat{Y}} [\mathbb{I}(c(\hat{Y}) \geq 0.5)]$	\times
Score Parity	Counterfactual pair	$f(\hat{Y}) = \mathbb{E}_{\hat{Y} \in \hat{Y}} [c(\hat{Y}_i, i) A = i] - \mathbb{E}_{\hat{Y} \in \hat{Y}} [c(\hat{Y}_j, j) A = j] $	\times
Counterfactual Sentiment Bias	Counterfactual pair	$f(\hat{Y}) = W_1(P(c(\hat{Y}_i) A = i), P(c(\hat{Y}_j) A = j))$	\times
Regard Score	Counterfactual tuple	$f(\hat{Y}) = c(\hat{Y})$	\times
Full Gen Bias	Counterfactual tuple	$f(\hat{Y}) = \sum_{i=1}^C \text{Var}_{w \in W} (\frac{1}{ \hat{Y}_w } \sum_{\hat{Y}_w \in \hat{Y}_w} c(\hat{Y}_w)[i])$	✓
LEXICON (§ 3.5.3)			
HONEST	Counterfactual tuple	$f(\hat{Y}) = \frac{\sum_{\hat{Y}_k \in \hat{Y}_k} \sum_{\hat{Y} \in \hat{Y}_k} \mathbb{I}(\text{HurtLex}(\hat{y}))}{ \hat{Y}_k }$	\times
Psycholinguistic Norms	Any prompt	$f(\hat{Y}) = \frac{\sum_{\hat{Y} \in \hat{Y}} \sum_{\hat{y} \in \hat{Y}} \text{sign}(\text{affect-score}(\hat{y})) \text{affect-score}(\hat{y})^2}{\sum_{\hat{Y} \in \hat{Y}} \sum_{\hat{y} \in \hat{Y}} \text{affect-score}(\hat{y}) }$	✓
Gender Polarity	Any prompt	$f(\hat{Y}) = \frac{\sum_{\hat{Y} \in \hat{Y}} \sum_{\hat{y} \in \hat{Y}} \text{sign}(\text{bias-score}(\hat{y})) \text{bias-score}(\hat{y})^2}{\sum_{\hat{Y} \in \hat{Y}} \sum_{\hat{y} \in \hat{Y}} \text{bias-score}(\hat{y}) }$	✓

LLM Stereotypes Mitigation & Correction

Bias and Fairness in Large Language Models: A Survey (June 11 2024, Isabel O. Gallegos et al, Stanford University and Adobe Research)

Taxonomy of Datasets for Bias Evaluation

- Counterfactual Inputs: Compare sets of sentences with perturbed social groups
 - MASKED TOKENS (§4.1.1): LLM predicts the most likely fill-in-the-blank

The engineer informed the client that [MASK: **she**/**he**/**they**] would need more time to complete the project.
 - UNMASKED SENTENCES (§4.1.2): LLM predicts the most likely sentence

We can't go to that one in a [**Mexican**/**white**] neighborhood. You might be forced to buy drugs.
- Prompts: Provide a phrase to a generative LLM to condition text completion
 - SENTENCE COMPLETIONS (§4.2.1): LLM provides a continuation
 - QUESTION-ANSWERING (§4.2.2): LLM selects an answer to a question

LLM Stereotypes Mitigation & Correction

Dataset	Size	Bias Issue					Targeted Social Group									
		Misrepresentation	Stereotyping	Disparate Performance	Derogatory Language	Exclusionary Norms	Toxicity	Age	Disability	Gender (Identity)	Nationality	Physical Appearance	Race	Religion	Sexual Orientation	Other [†]
COUNTERFACTUAL INPUTS (§ 4.1)																
MASKED TOKENS (§ 4.1.1)																
Winogender	720	✓	✓	✓		✓				✓						
WinoBias	3,160	✓	✓	✓	✓	✓				✓						
WinoBias+	1,367	✓	✓	✓	✓	✓				✓						
GAP	8,908	✓	✓	✓	✓	✓				✓						
GAP-Subjective	8,908	✓	✓	✓	✓	✓				✓						
BUG	108,419	✓	✓	✓	✓	✓				✓						
StereoSet	16,995	✓	✓	✓	✓	✓							✓	✓		✓
BEC-Pro	5,400	✓	✓	✓		✓				✓						
UNMASKED SENTENCES (§ 4.1.2)																
CrowS-Pairs	1,508	✓	✓	✓				✓	✓	✓	✓	✓	✓	✓	✓	✓
WinoQueer	45,540	✓	✓	✓	✓					✓						
RedditBias	11,873	✓	✓	✓	✓		✓			✓			✓	✓	✓	
Bias-STS-B	16,980	✓	✓	✓						✓						
PANDA	98,583	✓	✓	✓				✓					✓			
Equity Evaluation Corpus	4,320	✓	✓	✓	✓								✓			
Bias NLI	5,712,066	✓	✓			✓				✓	✓			✓		
PROMPTS (§ 4.2)																
SENTENCE COMPLETIONS (§ 4.2.1)																
RealToxicityPrompts	100,000				✓		✓									✓
BOLD	23,679				✓	✓	✓			✓			✓	✓	✓	✓
HolisticBias	460,000	✓	✓	✓				✓	✓	✓	✓	✓	✓	✓	✓	✓
TrustGPT	9*			✓	✓	✓		✓		✓			✓	✓		
HONEST	420	✓	✓	✓			✓			✓			✓			
QUESTION-ANSWERING (§ 4.2.2)																
BBQ	58,492	✓	✓	✓		✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
UnQover	30*	✓	✓		✓					✓				✓	✓	
Grep-BiasIR	118	✓	✓	✓		✓				✓						

*These datasets provide a small number of templates that can be instantiated with an appropriate word list.

[†]Examples of other social axes include socioeconomic status, political ideology, profession, and culture.

LLM Stereotypes Mitigation & Correction

Bias and Fairness in Large Language Models: A Survey (June 11 2024, Isabel O. Gallegos et al, Stanford University and Adobe Research)

Taxonomy of Techniques for Bias Mitigation.

- Pre-Processing Mitigation: Change model inputs (training data or prompts)
- In-Training Mitigation: Modify model parameters via gradient-based updates
- Intra-Processing Mitigation: Modify inference behavior without further training
- Post-Processing Mitigation: Modify output text generations

LLM Stereotypes Mitigation & Correction

Bias and Fairness in Large Language Models: A Survey (June 11 2024, Isabel O. Gallegos et al, Stanford University and Adobe Research)

Taxonomy of Techniques for Bias Mitigation.

- Pre-Processing Mitigation: Change model inputs (training data or prompts)
 - DATA AUGMENTATION (§5.1.1): Extend distribution with new data. Expand the training data with additional examples that balance the representation of different social groups.
 - DATA FILTERING AND REWEIGHTING (§5.1.2): Remove or reweight instances in the training data to reduce their impact.
 - DATA GENERATION (§5.1.3): Produce new data meeting certain standards
 - INSTRUCTION TUNING (§5.1.4): Prepend or append tokens to input prompts to guide the model towards unbiased behavior.
 - PROJECTION-BASED MITIGATION (§5.1.5): Transform hidden representations

LLM Stereotypes Mitigation & Correction

Bias and Fairness in Large Language Models: A Survey (June 11 2024, Isabel O. Gallegos et al, Stanford University and Adobe Research)

Taxonomy of Techniques for Bias Mitigation.

- In-Training Mitigation: Modify model parameters via gradient-based updates
 - ARCHITECTURE MODIFICATION (§5.2.1): Alter the model's architecture to inherently reduce bias, such as by incorporating fairness constraints.
 - LOSS FUNCTION MODIFICATION (§5.2.2): Introduce a new objective. Adjust the optimization objective to penalize biased outputs or reward fairness.
 - SELECTIVE PARAMETER UPDATING (§5.2.3): Fine-tune only specific parameters that are related to biased behaviors, while keeping the rest of the model fixed.
 - FILTERING MODEL PARAMETERS (§5.2.4): Remove a subset of parameters

LLM Stereotypes Mitigation & Correction

Bias and Fairness in Large Language Models: A Survey

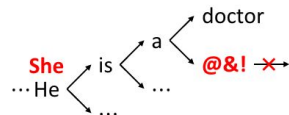
(June 11 2024, Isabel O. Gallegos et al, Stanford

University and Adobe Research)

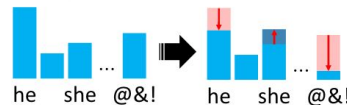
Taxonomy of Techniques for Bias Mitigation.

- Intra-Processing Mitigation: Modify inference behavior without further training
 - DECODING STRATEGY MODIFICATION (§5.3.1): Modify probabilities. Change the way the model generates outputs, such as by adjusting sampling methods to avoid biased continuations.
 - WEIGHT REDISTRIBUTION (§5.3.2): Modify the entropy of attention weights
 - MODULAR DEBIASING NETWORKS (§5.3.3): Add stand-alone components

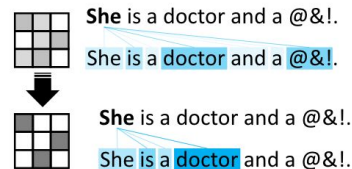
Decoding Strategy Modification Constrained Next-Token Search



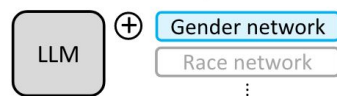
Modified Token Distribution



Weight Redistribution



Modular Debiasing Networks

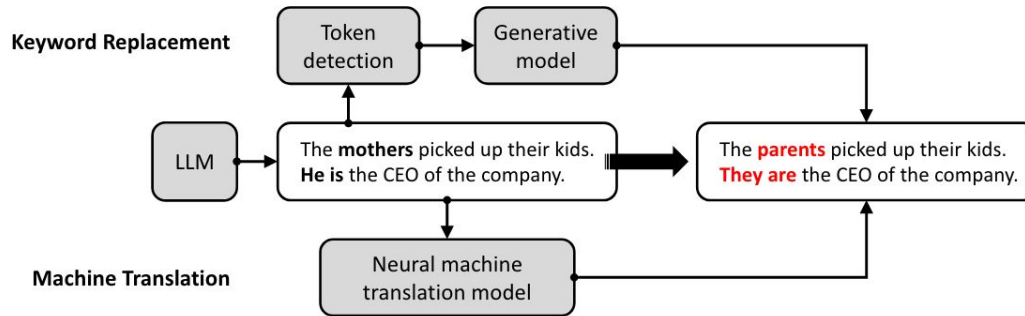


LLM Stereotypes Mitigation & Correction

Bias and Fairness in Large Language Models: A Survey (June 11 2024, Isabel O. Gallegos et al, Stanford University and Adobe Research)

Taxonomy of Techniques for Bias Mitigation.

- Post-Processing Mitigation: Modify output text generations
 - REWRITING (§5.4.1): Detect harmful words and replace them, using a rule- or neural-based rewriting algorithm.



Example Post-processing Mitigation Techniques (§ 5.4). **We illustrate how post-processing methods can replace a gendered output with a gender-neutral version.** Keyword replacement methods first identify protected attribute terms (i.e., "mothers," "he"), and then generate an alternative output. Machine translation methods train a neural machine translator on a parallel biased-unbiased corpus and feed the original output into the model to produce an unbiased output.

Questions

We understand there are biases and stereotypes in LLMs. Most measurements are based on gender, racial groups, and occupations. In our case, should we put a little more emphasize on stereotype and biases towards Down syndrome user groups or people with disabilities?

Community Engagement: Collecting stereotype towards down syndrome?

How to choose the right evaluation method?

Revised Project Overview

Revised Project Overview

RESEARCH PROBLEM

Personas serve as powerful tools for understanding and communicating user goals and behaviors within specific contexts in human-centered design used by product designers, development teams, as well as stakeholders. Traditional personas typically include a narrative and a photo. However, this static format relies on designers and development teams to empathize and role-play during the design phase. Creating interactive personas offers promising opportunities to provide real-time feedback based on user data and context throughout the design, development, and decision-making process.

Persona creation has traditionally relied on qualitative methods such as interviews, observations, and survey data. More recent methods use large datasets, including statistical data, clickstreams, and social media data. However, these approaches often did not address the sensitive user groups where empirical data is restricted due to practical and ethical reasons and large datasets are not readily available.

The development of Large Language Models (LLMs) presents an opportunity to provide context in addition to the limited data available and create interactive persona to represent, engage, and empower sensitive user groups. Traditional and data-driven methods of persona creation often fall short for sensitive groups due to the lack of available data. This research aims to explore the use of LLMs to generate proxy personas for sensitive groups using limited data from online forums, where empirical data is restricted due to practical and ethical reasons.

However, LLMs often inherit and propagate societal biases present in their training data, leading to the creation of biased and stereotypical personas. These biases can negatively impact the design process, user experience, and overall inclusivity of the systems.

Revised Project Overview

MOTIVATION

When relying on LLMs for interactive persona creation and interaction, it is critical to address the ethical and practical implications of biases in the system. Biased personas can reinforce harmful stereotypes, decrease user trust, and result in suboptimal design decisions. Ensuring fair and accurate representations of all user groups is essential for creating inclusive and effective designs. This project aims to develop and evaluate methods for detecting, monitoring, and correcting biases in interactive persona systems, thereby enhancing their reliability and ethical use.

RESEARCH GAP

While there is increasing numbers of research focusing on bias detection and mitigation in LLMs, few studies specifically address the unique challenges posed by interactive persona systems. There are a few studies that evaluated LLMs generated personas for diversity, authenticity, and stereotype, but the LLMs generated personas usually are in traditional persona format, including a narrative and a photo. There are limited studies on interactive persona that maintain live interaction, engagement, and communication throughout the design process. Therefore, there is a gap for stereotype and bias evaluation of interactive persona. Existing methods often focus on general text generation or dialogue systems, leaving a gap in the context of interactive persona creation and interaction. Additionally, there is a lack of comprehensive frameworks that integrate detection, monitoring, and correction of biases and stereotypes across the entire lifecycle of persona development.

Revised Project Overview

RESEARCH QUESTIONS

1. Detection: What methods are effective for biases and stereotype detection?
2. Monitoring: How can biases and stereotypes in interactive persona systems be continuously monitored to ensure fair and accurate representations?
3. Correction: When biases and stereotypes are detected, how can they be corrected in the system or during real-time interaction? How effective are the correction methods?
4. Evaluation: How to evaluate the effectiveness of bias detection, monitoring, and correction methods in terms of improving the safety, authenticity, diversity, and inclusivity of interactive personas?

PURPOSE OF PROJECT / STUDY

The goal of this project is to develop a comprehensive framework for detecting, monitoring, and correcting biases and stereotypes in interactive persona systems. By addressing the research questions above, the project aims to:

- **Ethical:** Ensure that the interactive persona system does not create harmful stereotypes or biases.
- **User Trust:** Ensure the credibility of the interactive persona system by providing fair representations of diverse user groups.
- **Inclusivity:** Facilitate better design decisions that consider the needs and experiences of all user groups, leading to more inclusive and effective products.

Revised Project Overview

WITHIN SCOPE

The project will involve collaboration with the Natural Language Processing (NLP) team and Interactive Persona System UI design team for system evaluation. The project will focus on evaluating the outputs generated by the NLP team and developing methods to correct any identified biases. The project would propose system improvements for detection, monitoring, and correction of biases and stereotypes. The project will involve building an extension to the interface built by UI design team.

OUTSIDE OF SCOPE

The project does not involve data collection, developing and building the backend, or the design of the UI of interactive persona system.