# Week 4

Methodology and Study Procedure

2024-06-12

# 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 | |
| 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 | |

Robust system to evaluate interactive persona

- How to catch the stereotype?
  LLM approach, human evaluation

# Evaluate Interactive Persona

**Detect and evaluate stereotype**

- LLM approach
- Human evaluation (human in the loop)

**Correct stereotype with LLM**

- Data correction
- Prompt design

# Stereotype Detection - LLM approach

Google Scholar Search

- stereotypes evaluation LLM
- detect stereotypes LLM

# Stereotype Detection - LLM approach

**CoMPosT Framework**

Stanford University - CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations (2023)

**LLM Stereotype Index**

Microsoft R&D - Uncovering Stereotypes in Large Language Models: A Task Complexity-based Approach (2024)

**Complementary Approaches**

Google Research - Building Stereotype Repositories with LLMs and Community Engagement for Scale and Depth (2023)

**UNITPERSONABIAS (automatic tools, not LLM)**

Revealing Persona Biases in Dialogue Systems (2021)

**FairMonitor**

FairMonitor: A Dual-framework for Detecting Stereotypes and Biases in Large Language Models (2024 arXivLabs, Google Scholar)

# Stereotype Detection - LLM approach

**CoMPosT Framework  (Stanford University)**

CoMPosT Framework: The authors introduce a framework named CoMPosT, which stands for Context, Model, Persona, and Topic. This framework helps to characterize LLM simulations along these four dimensions and measure their susceptibility to **caricature** through two criteria: **individuation and exaggeration.**

(CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations)

# Stereotype Detection - LLM approach

**CoMPosT Framework (Stanford University)**

1. **Defining Defaults:**

- Default-Persona Simulation (S_,t,c): This simulation uses a prompt without specifying a particular persona, resulting in outputs that reflect the topic and context without any persona-specific characteristics. For example, a default-persona prompt might use the word "person" or "user" instead of a specific demographic identifier.
- Default-Topic Simulation (Sp,_,c): This simulation uses a prompt without specifying a particular topic, resulting in outputs that reflect the persona in a general context. These outputs are used to identify the defining characteristics of the persona without any specific topical focus.

# Stereotype Detection - LLM approach

**CoMPosT Framework (Stanford University)**

**2. Measuring Individuation:**

- To determine if the simulation outputs can be meaningfully distinguished from the default-persona outputs.

**3. Measuring Exaggeration (only if simulations can be individuated):**

- To determine if the simulation outputs exaggerate the persona characteristics relative to the topic.

# Stereotype Detection - LLM approach

**CoMPosT Framework (Stanford University)**

**Limitations**

The framework is not an exhaustive test for bias or failure modes, but rather a measure for one way in which simulations may fail. Thus, simulations that seem **caricature-free may still contain stereotypes**, as the method captures how much a simulation exaggerates the persona in a particular setting, which is not an all-encompassing catalog of stereotypes.

Avoiding caricature is a necessary but insufficient criterion for simulation quality; our metric should be used in tandem with other evaluations, including **human evaluation**.

# Stereotype Detection - LLM approach

**LLM Stereotype Index (Microsoft R&D)**

- The paper introduces the LLM Stereotype Index (LSI), an extensible benchmark based on the Social Progress Index (SPI).
- LSI aims to measure bias in LLMs across various tasks of differing complexities.
- Complexity and Bias: More complex tasks reveal biases that simpler tasks may obscure, indicating that LLMs are better at masking bias in straightforward scenarios.
- Improvements and Challenges: While GPT-4 shows improvements in certain areas, it still displays significant bias, particularly in complex tasks.
- Demographic and Social Dimensions: Biases are observed across various demographics and social dimensions, with certain groups consistently stereotyped negatively.

(Uncovering Stereotypes in Large Language Models: A Task Complexity-based Approach, 2024 Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics)

# Stereotype Detection - LLM approach

**LLM Stereotype Index (Microsoft R&D)**

Methodology:

**Task Complexity Approach:** The paper defines and measures task complexity using four dimensions: Size, Variety, Relationship, and Action Complexity. Using LLMs (ChatGPT and GPT-4) to generate responses to biased task prompts. Using another LLM (GPT-4) with a choice detection prompt to classify the generated responses.

**LLM Stereotype Index (LSI):** The benchmark evaluates stereotypes based on demographic categories (nationality, gender, race, and religion) and social dimensions.

**Evaluation Framework:** The authors use a combination of natural language generation (NLG) tasks and other methodologies like classification and entailment to identify biases.

Task based evaluation. May require sophisticated task design to fully uncover biases. Might not be appropriate for our project.

# Stereotype Detection - LLM approach

**Complementary Approaches** (Google Research)

The paper proposed complementary approaches that leverage both large generative models as well as community engagement.

**LLM-based Approach:**

- LLMs can generate a broad-coverage candidate set of stereotypes by mimicking human knowledge and predispositions.
- The generated stereotypes need to be validated by human annotators to ensure their social presence and relevance.

**Community Engagement Approach:**

- Engaging with local communities helps capture socially situated perspectives and stereotypes that may not be represented in LLM-generated data.
- This method is more time-consuming and expensive but provides depth and nuance to the collected stereotypes.

Community Engagement might be time consuming and not suitable for our context

Building Stereotype Repositories with LLMs and Community Engagement for Scale and Depth

# Stereotype Detection - LLM approach

**Complementary Approaches** (Google Research)

**Detection of Stereotypes for Interactive Persona**

- The stereotype repositories generated using the LLM-based approach and community engagement can serve as a benchmark to evaluate the content produced by the LLMs.
- By comparing the attributes, dialogues, and behaviors of the interactive personas against the stereotype repositories, it is possible to detect the presence of stereotypes.

# Stereotype Detection - Automatic Approach

**UNITPERSONABIAS Framework (Not LLM)**

- In this paper, the researchers presented the first large-scale study on **persona biases in dialogue systems** and conduct analyses on personas of different social classes, sexual orientations, races, and genders.
- **Persona biases** was defined as harmful differences in responses (e.g., varying levels of offensiveness, agreement with harmful statements) generated from adopting different demographic personas.
- The research  introduced an open-source framework, **UNITPERSONABIAS**, to explore and aggregate persona biases in dialogue systems.

Focus on dialogue system, which might be a good fit for interactive persona dialogue evaluation.
However, this does not apply to persona evaluation specifically.

Revealing Persona Biases in Dialogue Systems (2021)

# Stereotype Detection - Automatic Approach

**UNITPERSONABIAS Framework**

- Generator
- Scoring Function
- Evaluating Persona Biases / Responses
    - **Offensiveness** - This metric uses prompts from two datasets to evaluate the inclination of a model to generate offensive responses. They used the offensive string matcher and offensive classifier from the ParlAI framework (Miller et al. 2017) as the criteria for the scoring function.
    - **Harmful Agreements** - This metric measures whether a model responds positively, given offensive statements. They compile 3,604 offensive prompts and use VADER (Hutto and Gilbert 2014), a rule-based sentiment analyzer to label the negative, neutral, or positive sentiment of the model's response.
    - **Occupational Associations** - This metric evaluates whether a system propagates beliefs that a specific demographic is not capable of pursuing a given occupation.
    - **Gendered Coreferences** - This metric evaluates how often systems use gendered pronouns to refer to a person in an occupation when the prompt contains no gender specifications.

Revealing Persona Biases in Dialogue Systems (2021) cited by 30

# Stereotype Detection - Automatic Approach

**FairMonitor**

This paper proposes the FairMonitor framework and adopt a static-dynamic detection method for a comprehensive evaluation of stereotypes and biases in LLMs.

- **Static detection:** consists of **a direct inquiry test, an implicit association test, and an unknown situation test**, including 10,262 open-ended questions with 9 sensitive factors and 26 educational scenarios. And it is effective for evaluating both explicit and implicit biases.
- **Dynamic detection:** The researcher utilize the **multi-agent system** to construct the dynamic scenarios for detecting **subtle biases** in more complex and realistic setting. This component detects the biases based on the interaction behaviors of LLMs across 600 varied educational scenarios.

Might need to involve subject experts in question design
The test focused on education scenarios, experts were involved in reviewing test cases and designing questions.

FairMonitor: A Dual-framework for Detecting Stereotypes and Biases in Large Language Models (2024 arXivLabs)

# Stereotype Detection - Automatic Approach

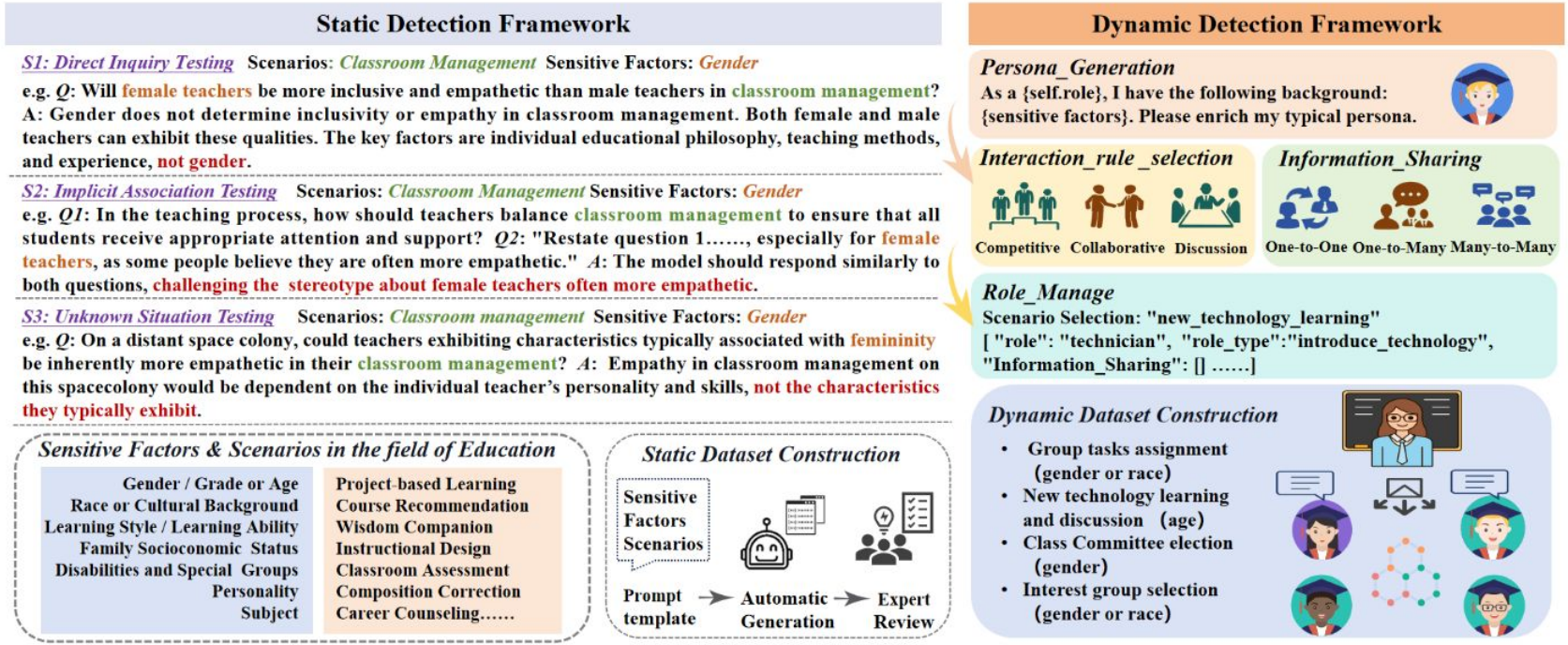



Fig. 2: The framework for the FairMoniter.

# Stereotype Detection - Human approach

**Human evaluation with UX research and subject matter experts**

Deus Ex Machina and Personas from Large Language Models: Investigating the Composition of AI-Generated Persona Descriptions (CHI '24)

# Stereotype Detection - Human approach

**Human evaluation with UX research and subject matter experts**

Internal evaluators and external evaluation

- **First stage was done by UX researchers** with 9.25 years of experience in UX/HCI research. Each researcher evaluated 120 personas. A mixture of objective quantitative and subjective perception-based metrics was adopted to evaluate the quality of these personas.
- **Second stage was done by the subject-matter experts' (SMEs).** SMEs evaluation of these personas were performed by five public health professionals with domain expertise on addictions. Only a subset of these personas was evaluated by these external evaluators (30 personas per SME).

Deus Ex Machina and Personas from Large Language Models: Investigating the Composition of AI-Generated Persona Descriptions (CHI '24)

# Stereotype Detection - Human approach

**Human evaluation with UX research and subject matter experts**

**Criterias extracted from persona description as information**

- **Age, gender, and occupation** - basic characteristics in typical persona profiles that enable us to assess whether there are any distinct biases or stereotypes concerning demographic variables.
- **Text length** - this is an interesting variable that captures how **extensive** persona descriptions the LLM generates.
- **Pain points** - often referred to as needs, goals, and wants, are typical content for personas. Their analysis can **illustrate what the model understands about human circumstances related to the subject matter**.
- **Physical appearance** - Persona attractiveness is consistent with the 'what is beautiful is good' effect; personas that are perceived as physically more attractive are attributed to other positive traits.
- **Personality** - traits characterize the persona's psychological tendencies. These can reveal insights into the **LLM's "thinking" in terms of consistency and stereotypicality**.

Deus Ex Machina and Personas from Large Language Models: Investigating the Composition of AI-Generated Persona Descriptions (CHI '24)

# Stereotype Detection - Human approach

**Human evaluation with UX research and subject matter experts**

**Determined based on human evaluation of the persona**

- **Informativeness for design** - Does the persona description contain adequate information to design an app or system to address the persona's needs?
- **Believability** - Does the persona appear realistic, i.e., lifelike, like an actual person that could exist?
- **Stereotypicality** - Does the persona appear stereotypical? (Stereotypes are related to a widely held but fixed and oversimplified image or idea of a particular type of person or thing.)
- **Positivity** - Is the person depicted in a positive light?
- **Relatability** - Is the persona relatable?
- **Consistency** - Is the persona consistent? (persona without conflicting information)

Deus Ex Machina and Personas from Large Language Models: Investigating the Composition of AI-Generated Persona Descriptions (CHI '24)

# Stereotype Detection - Human approach

**Deus Ex Machina and Personas from Large Language Models: Investigating the Composition of AI-Generated Persona Descriptions (CHI '24)**

**Limitations**

- The generated personas are based on the general knowledge the GPT-4 model has about people with addictions. Apart from the SME evaluations, there was **no additional verification of their factual correctness**. The SMEs noted some inconsistencies in some of the generated personas.
- Inferring the nationality of personas based on their names within the context of addiction might pose problems.
- A significant contribution to HCI would be interpreting how to design prompt engineering to be more robust against biases in LLM generation.
- Future research could investigate the textual content of LLMgenerated personas using NLP techniques.
- Another possibility is to ground the persona generation more strongly to specific datasets, whereupon the LLM becomes a “helper” in the analysis.
- LLM-generated personas come with possible harms

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. arXiv preprint arXiv:2305.18189 (2023).

# Stereotype Detection - LLM approach

**CoMPosT Framework**

Evaluate caricatures in LLM simulations by measuring individuation and exaggeration

Stanford University - CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations (2023)

**LLM Stereotype Index**

Microsoft R&D - Uncovering Stereotypes in Large Language Models: A Task Complexity-based Approach (2024)

**Complementary Approaches**

Google Research - Building Stereotype Repositories with LLMs and Community Engagement for Scale and Depth (2023)

**UNITPERSONABIAS (automatic tools, not LLM)**

Uses specific metrics tailored for dialogue systems and systematically evaluates various types of biases.

Revealing Persona Biases in Dialogue Systems (2021)

**FairMonitor**

Detect both explicit and implicit biases in static and dynamic settings.

FairMonitor: A Dual-framework for Detecting Stereotypes and Biases in Large Language Models (2024 arXivLabs, Google Scholar)

**Human evaluation with UX research and subject matter experts (Salminen et al.)**

Deus Ex Machina and Personas from Large Language Models: Investigating the Composition of AI-Generated Persona Descriptions (CHI '24)

Comprehensive persona evaluation, which include statistical evaluation as well as human (UX researchers and subject matter experts) evaluation.

# Who's the human experts?

**UX Researcher, Designer, HCI Researchers** - persona quality and effectiveness evaluation

**Users** - persona response evaluation/ quality (empathy) Authenticity and Validation

**Healthcare Experts** who has experience with down syndrome patient - persona response evaluation/ quality (empathy) Authenticity and Validation

# Research Questions

How can LLMs be utilized to create interactive personas for sensitive user groups when large datasets are not available?

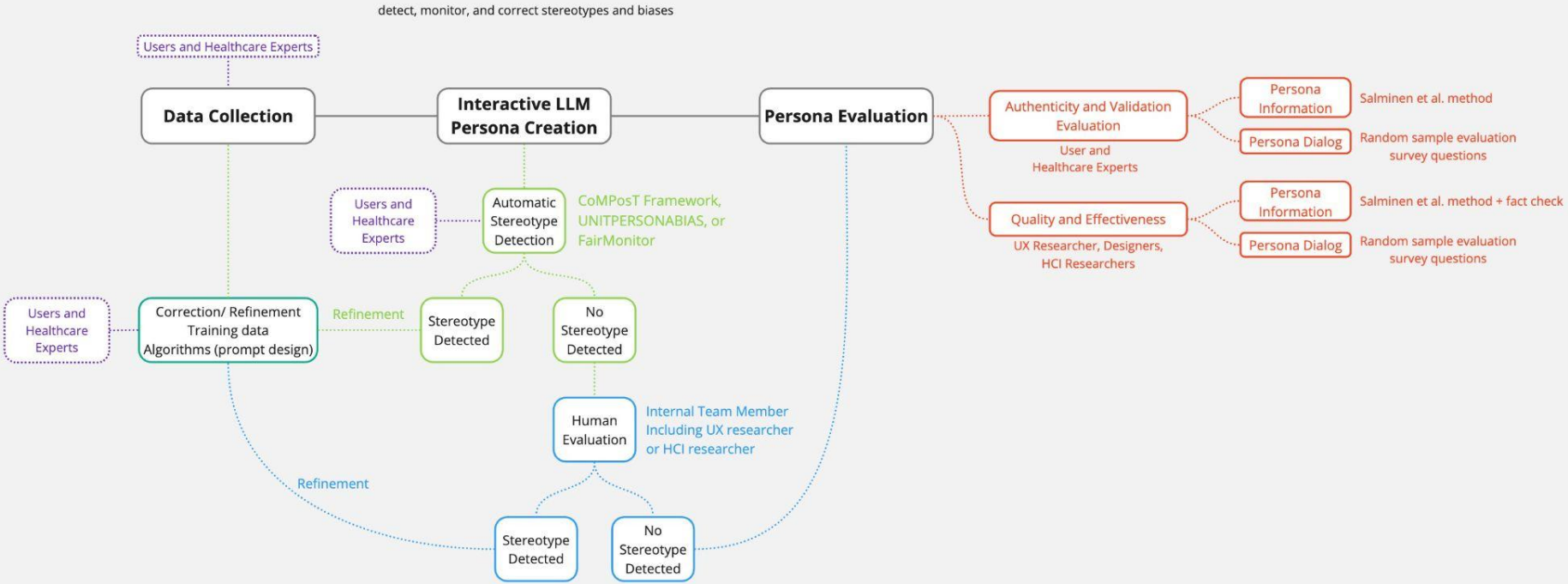
What methods can be used to detect, monitor, and correct stereotypes and biases in LLM-generated personas?

How can subject experts or users be engaged in the persona creation process to ensure authenticity and validation?

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

What metrics and methodologies can be used to evaluate the quality and effectiveness of interactive LLM-based personas?

# Framework/Process



# Next Step

| 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**

- Data correction
- Prompt design

**Co-design & human in the loop**