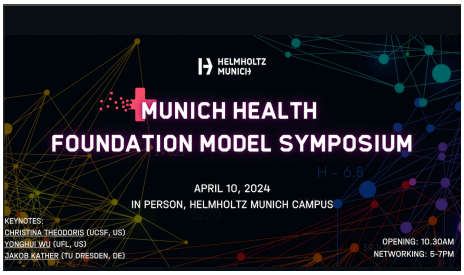

Contribution ID: **34** Type: **Talk**

VLMGender: A Benchmark for Evaluating Gender Bias in Vision-Language Assistants

Pre-trained large language models (LLMs) have been reliably integrated with visual input for multimodal tasks. The widespread adoption of instruction-tuned image-to-text vision-language assistants (VLAs) like LLaVA and MiniGPT, necessitates evaluating social biases. We propose a comprehensive prompt-image benchmark measuring gender bias for VLAs in personality traits, skills, and occupations. On this benchmark, we evaluate 16 popular open-source vision-language assistants, varying from 1.7B to 34B parameters. Results consistently show a tendency to attribute positive adjectives to females and negative adjectives to males. Additionally, many models exhibit a bias towards associating work-relevant soft skills with females. Furthermore, we discover that the vision module amplifies gender bias patterns present in the language module. Our research underscores the need for pre-deployment gender bias assessment in VLAs and advocates for the development of debiasing strategies to ensure equitable societal outcomes.

Primary author: Mr GIRRBACH, Leander

Co-authors: Dr ALANIZ, Stephan (Helmholtz Munich); Ms HUANG, Yiran (Helmholtz Munich); AKATA, Zeynep (Helmholtz Munich)

Session Classification: Break + Posters session

Track Classification: Poster