Nolwenn Bernard, PhD – Persona Blueprints: Decoding Intersectional Personas in LLMs

Nolwenn Bernard, PhD
is a postdoctoral researcher at the Faculty of Computer Science and Engineering Science. Her research focuses on user simulation and its applications in information access, particularly for evaluating information systems.

Beschreibung des Fellowvorhabens

- Die Nutzung von Large Language Models (LLMs) ist rasant gewachsen, was zu ihrer Anwendung in verschiedenen Bereichen – von der Forschung bis zur Industrie – geführt hat. Daher sind ihre gesellschaftlichen Auswirkungen erheblich. Damit einher gehen Befürchtungen bezüglich verzerrter Darstellungen und systemischer Biases, die zu einer unfairen Behandlung von Einzelpersonen oder Gruppen führen können.
- In jüngster Zeit ist im Kontext des Informationszugangs ein Wandel vom durchschnittlichen hin zum spezifischen Nutzerverhalten bei Simulationen auf Basis großer Sprachmodelle zu beobachten. Der Simulator wird dazu aufgefordert, als jemand mit einer vordefinierten Persona zu agieren (d. h. Rollenspiel), die in der Literatur typischerweise entlang verschiedener Dimensionen definiert wird (z. B. demografische Merkmale, Persönlichkeitseigenschaften oder sozialer/kultureller Hintergrund).
- Es besteht kein Konsens darüber, welche Attribute für die unterschiedlichen Dimensionen berücksichtigt werden sollten und wie diese Attribute modelliert werden können. Darüber hinaus ist unklar, ob die Persona lediglich als sprachliche Maske fungiert oder ob sie spezifische stereotype neuronale Schaltkreise im Simulator aktiviert.
- Die Motivation dieses Vorhabens besteht darin, besser zu verstehen, wie Personas intern in LLM-basierten Simulatoren repräsentiert sind, um Maßnahmen zur Minderung von Verzerrungen zu erleichtern, die Kontrolle simulierten Verhaltens zu ermöglichen sowie die Interpretierbarkeit der Simulatoren und der Simulationsergebnisse zu verbessern.

Forschungsfragen:

1. Können wir die Neuronen und Schaltkreise in LLMs identifizieren, die mit persona-abhängigem Verhalten zusammenhängen? Können wir eine generische Sonde entwickeln, um eine solche Identifikation durchzuführen?
2. Wie werden intersektionale Attribute einer Persona in LLMs repräsentiert? Werden sie unabhängig voneinander oder in verflochtener Weise dargestellt?

Description of the fellow project

- The use of large language models (LLMs) has been growing rapidly, leading to their adoption in various application domains from research to industry. Therefore, their social implications are significant, with concerns relating to the exhibition and amplification of systemic biases that can lead to unfair treatment of individuals or groups.
- Recently, a shift from average to specific user behavior regarding large language model-based simulation in the context of information access can be observed. The simulator is prompted to act as someone with a predefined persona (i.e., role play), in the literature typically defined along different dimensions (e.g., demographic, personality traits, or social/cultural background)
- There is no consensus on which attributes should be considered for the different dimensions and how these attributes should be modeled. Furthermore, it remains unclear if the persona acts as a linguistic mask or if it activates specific stereotypical neural circuits in the simulator.
- The motivation behind this project is to better understand how personas are internally represented in LLM-based simulators to facilitate mitigations of biases, control of simulated behavior, and interpretability of the simulators and simulation results.

Research questions:

1. Can we identify the neurons and circuits in LLMs that are related to persona-dependent behavior? Can we develop a generic probe to perform such identification?

1. How are intersectional attributes of a persona represented in LLMs? Are they represented independently or in an entangled manner?

Literatur

Cintas, C., Rateike, M., Miehling, E., Daly, E., & Speakman, S. (2025). Localizing Persona Representations in LLMs. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8(1), 630-642. https://doi.org/10.1609/aies.v8i1.36577

Gefördert durch:

Bundesministerium für Forschung, Technologie und Raumfahrt

Technology Arts Sciences TH Köln