

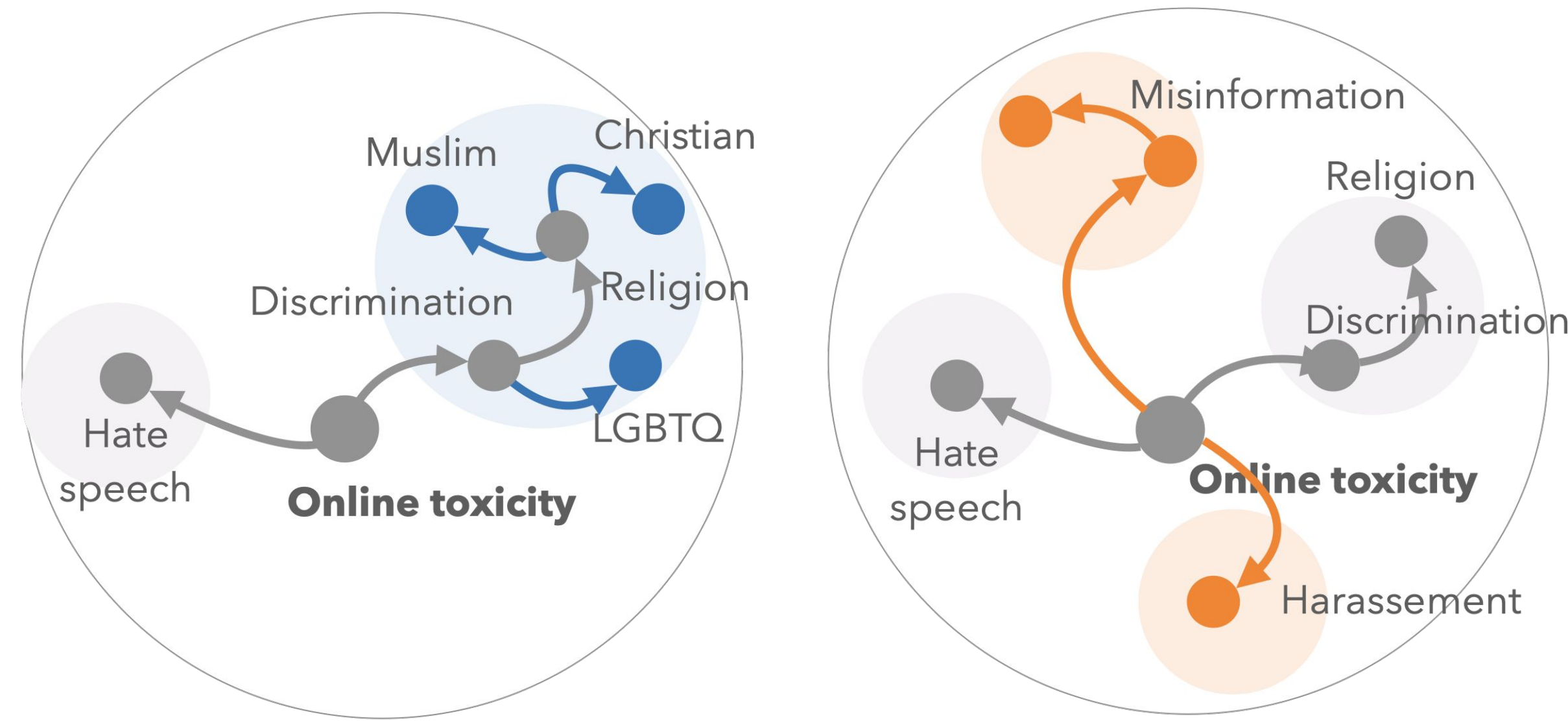
Beyond Testers' Biases: Guiding Model Testing with Knowledge Bases using LLMs

Chenyang Yang, Rishabh Rustogi, Rachel Brower-Sinning, Grace A. Lewis, Christian Kästner, Tongshuang Wu

Motivation

Coarse-grained metrics over benchmarks can not reveal nuanced model behaviors.

Model testing aims to explore nuanced behaviors, but is often conducted ad-hoc and biased by testers.



Users tend to explore **locally** overfit to their intuition, domain knowledge, confirmation bias.

Expect: **Comprehensive testing** More systematically cover the space

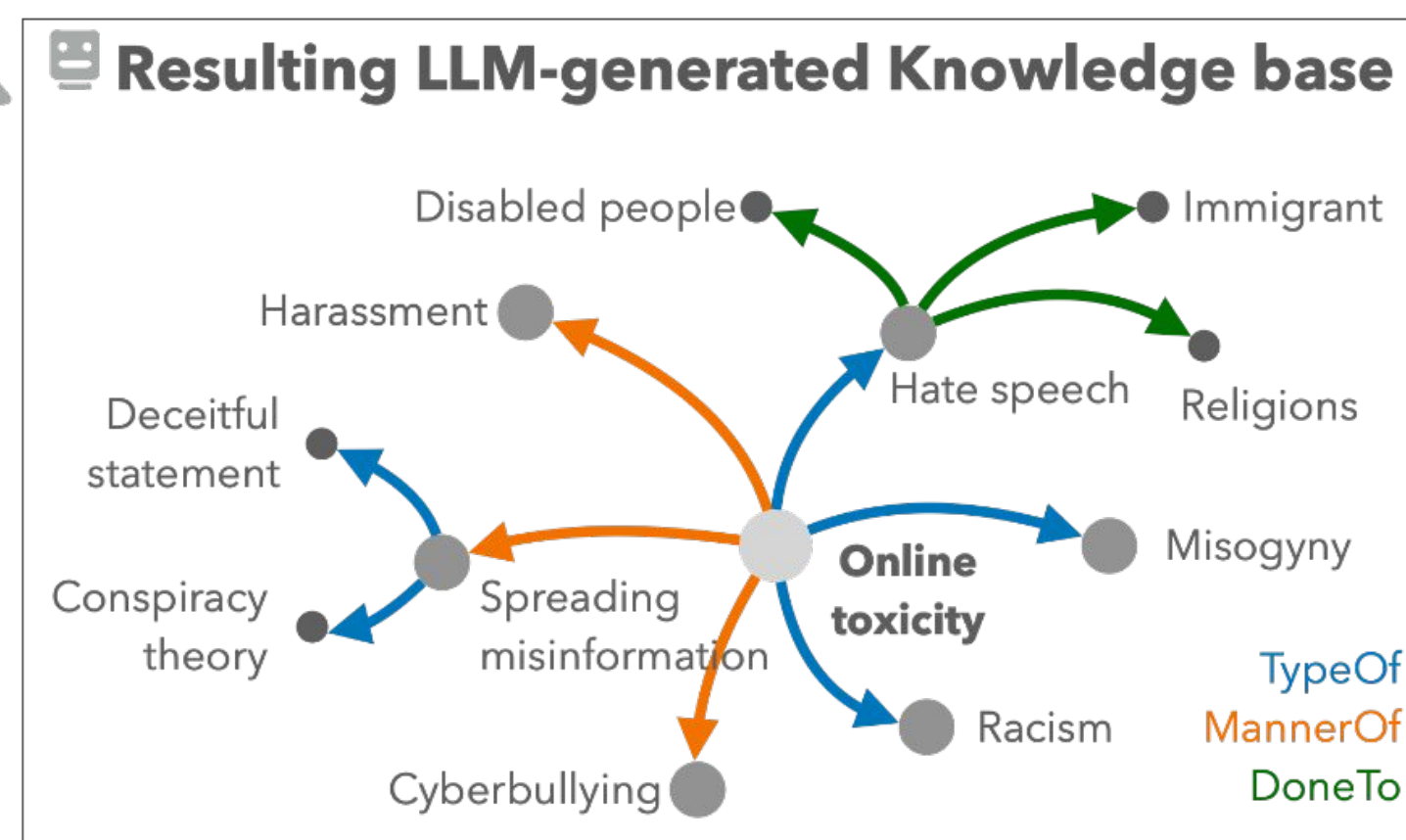
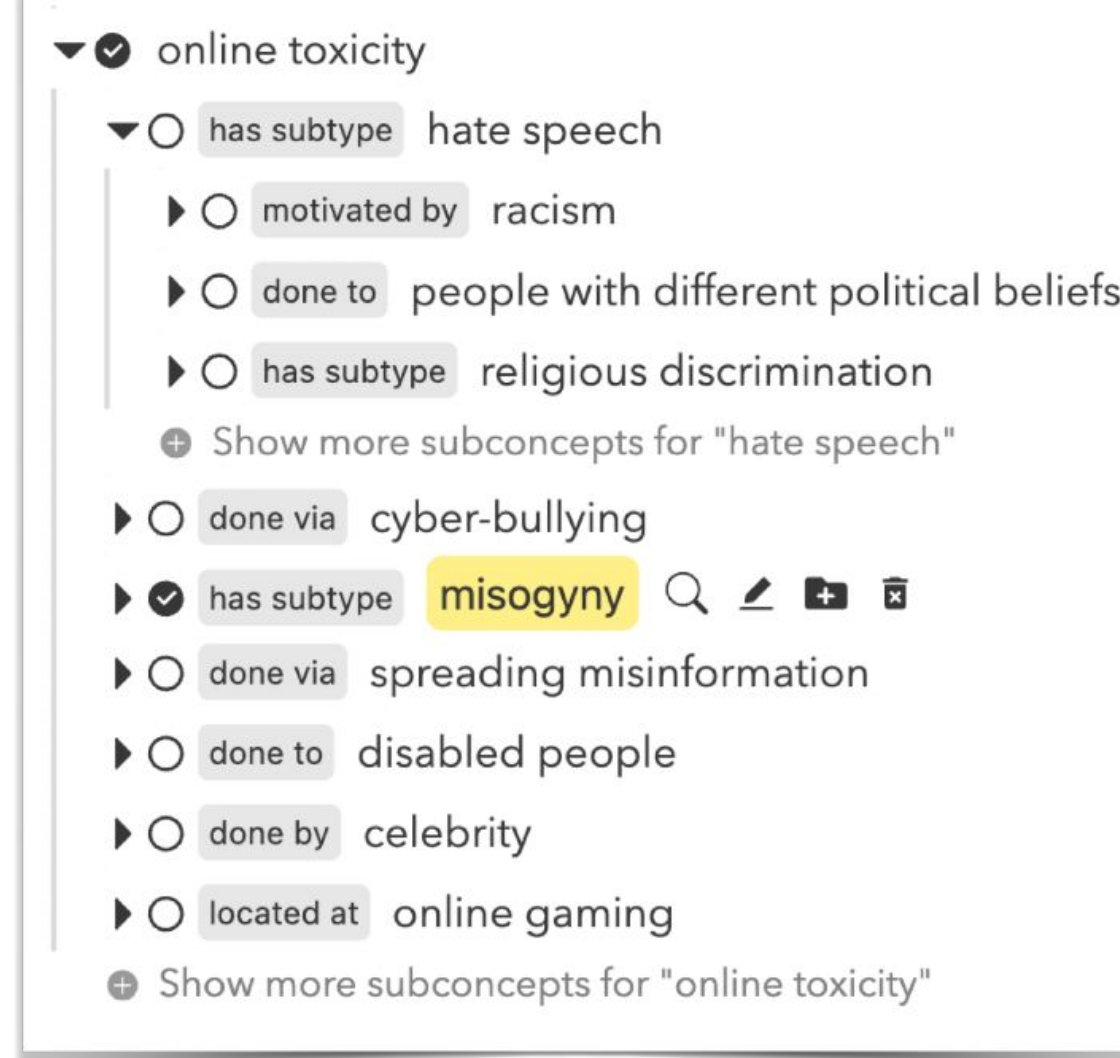
Introduce **requirements elicitation** for systematic model testing.

Weaver Workflow

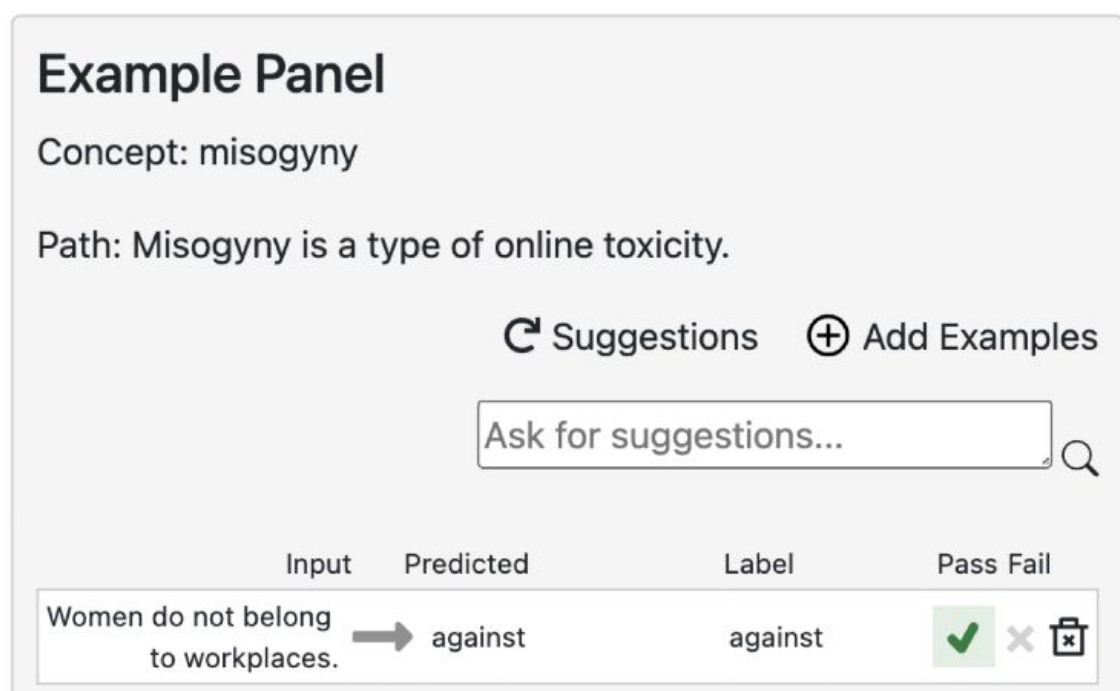
Interactive exploration of **knowledge base generated by LLM** for any tasks.

Seed concept: Online toxicity

Query LLMs for concepts (ConceptNet relations)
MannerOf: List some **ways to do** online toxicity: Harassment, Cyberbullying...
TypesOf: List some **types of** online toxicity: Racism, Misogyny...



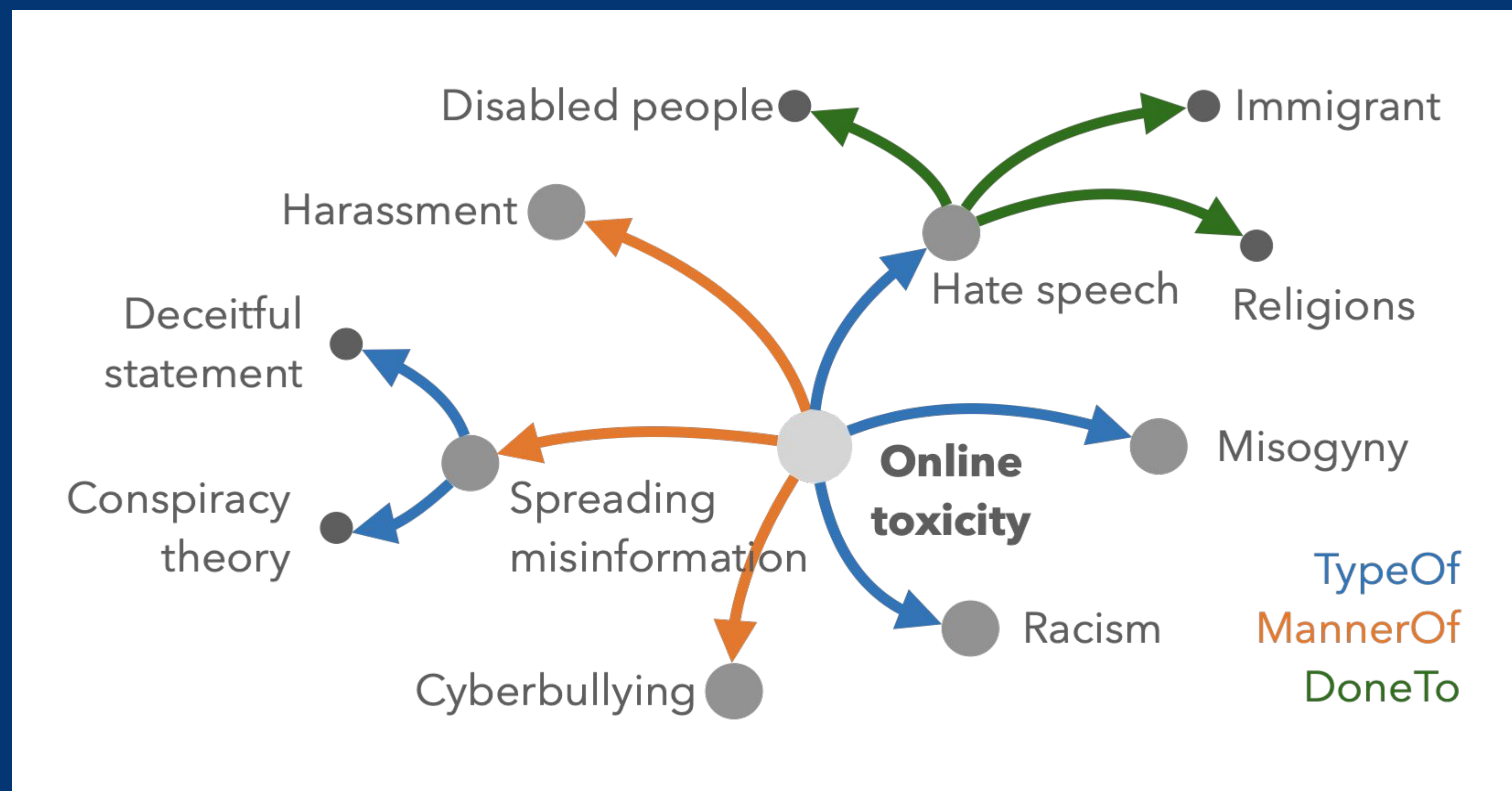
Recommend relevant & diverse concepts (Extract subgraph using greedy peeling)



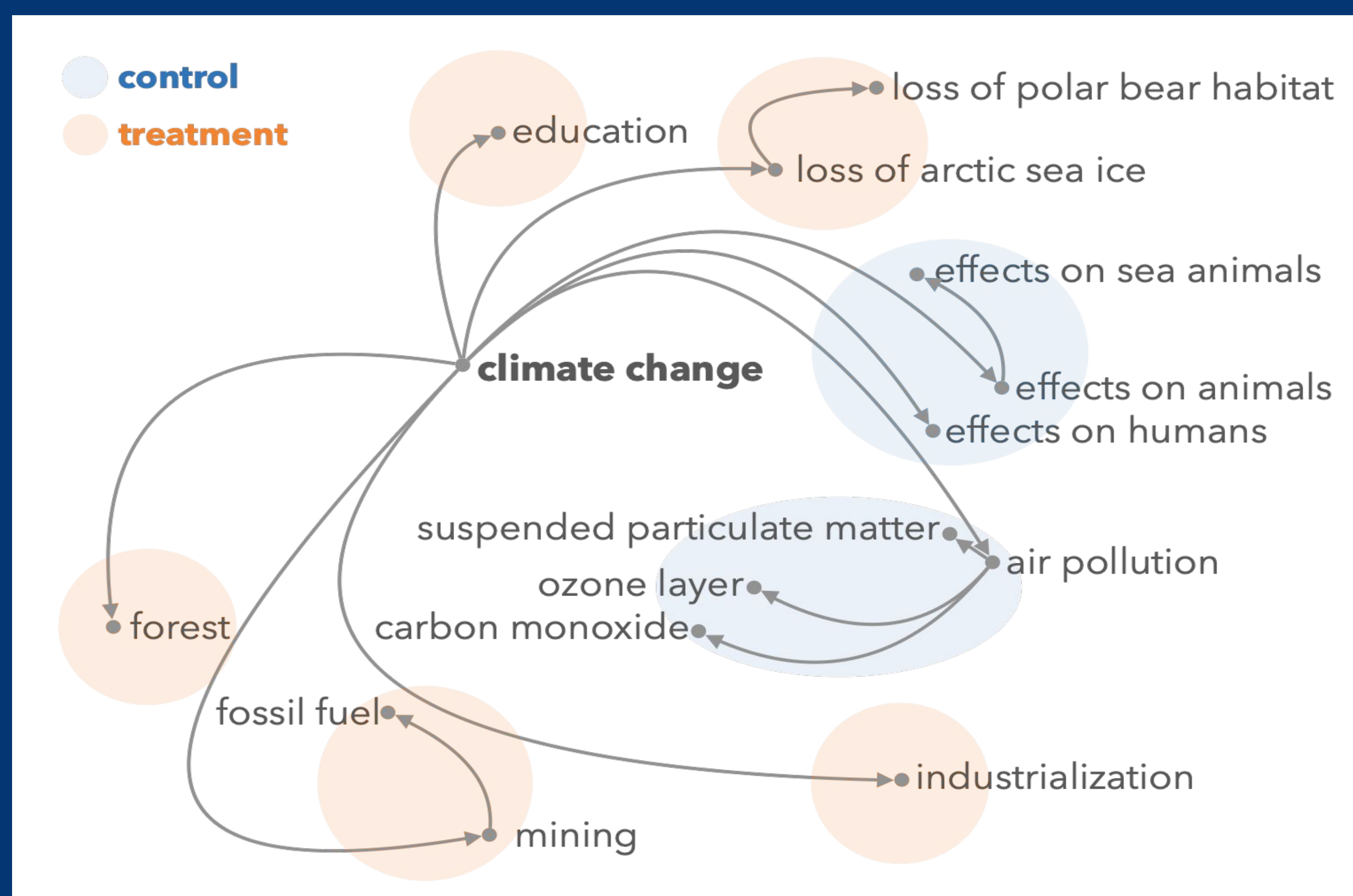
Intrinsic Evaluation

Weaver covers 90.6% existing concepts.

TL;DR: **Weaver** systematizes ad-hoc model testing with requirements elicitation, by extracting knowledge bases from LLMs for any tasks.



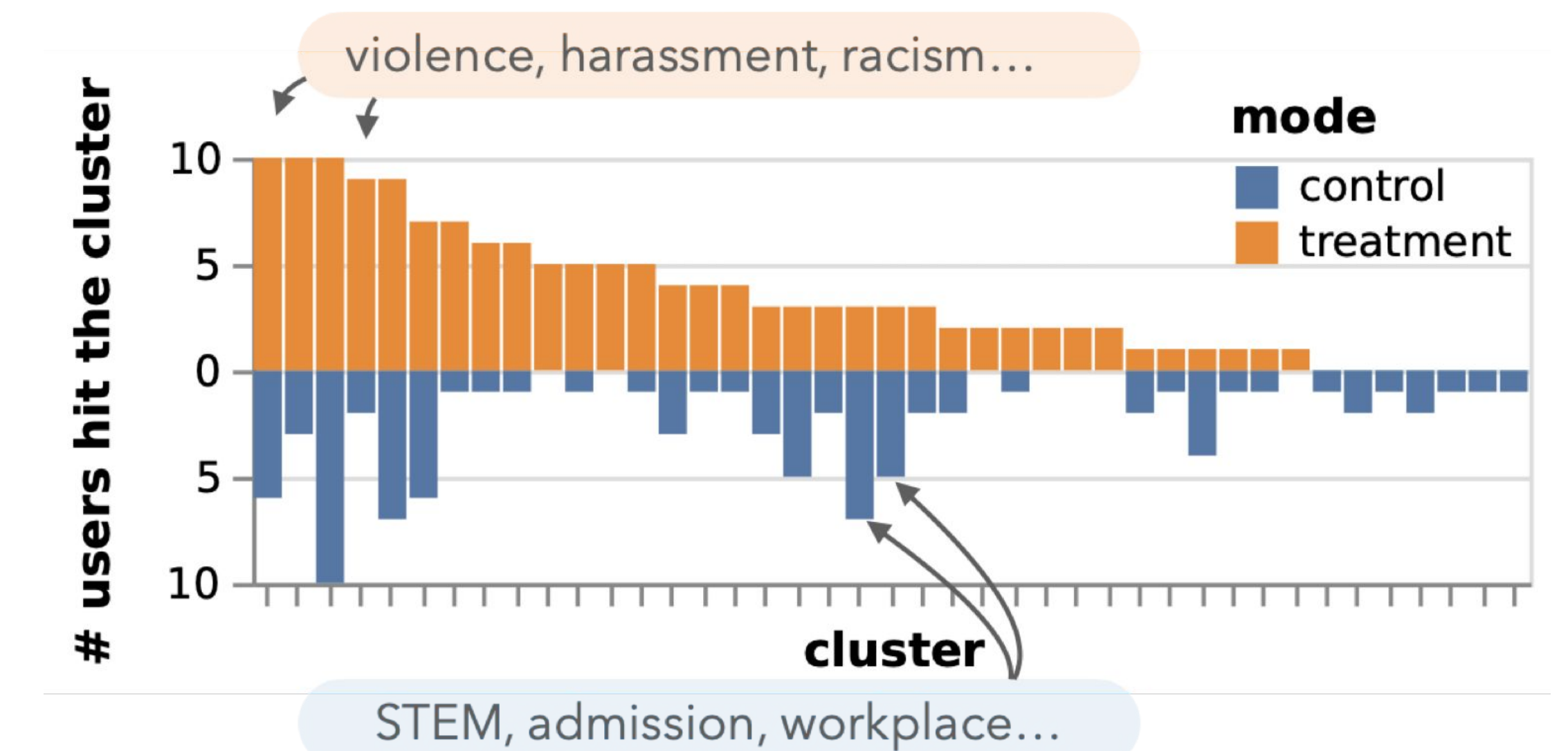
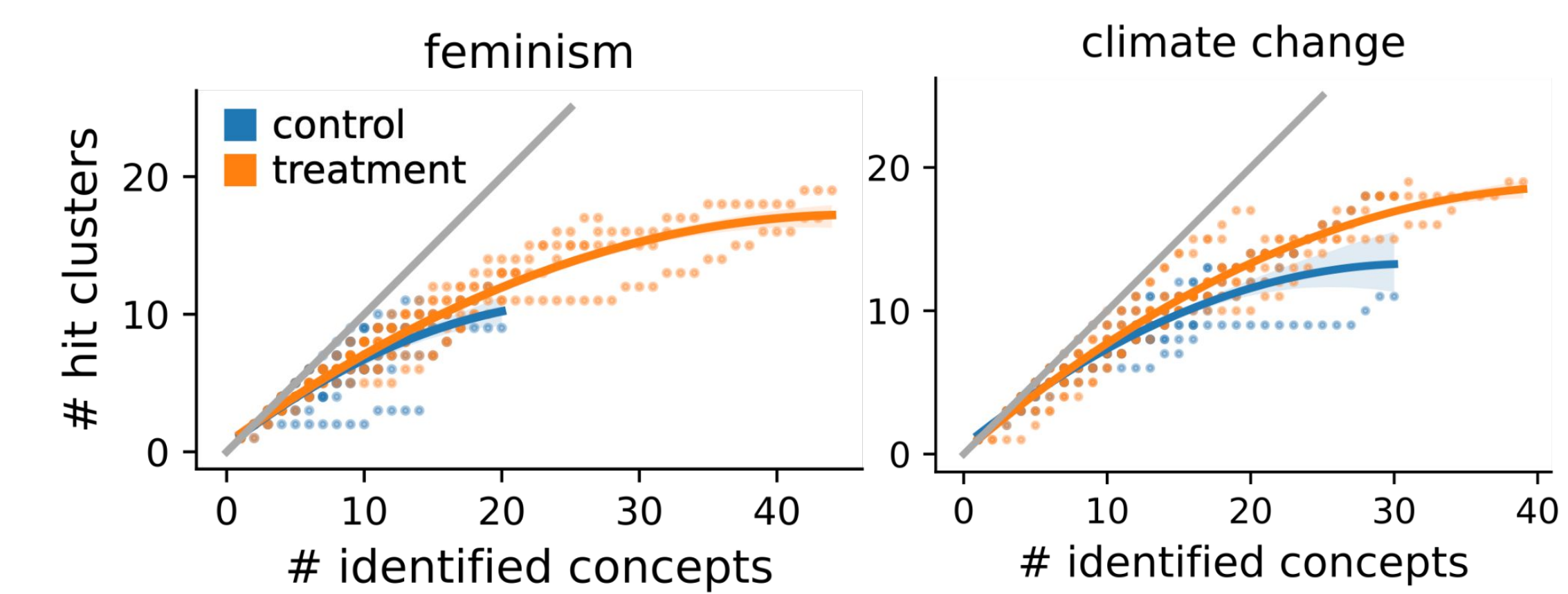
Findings: **Weaver** helps users explore model behaviors systematically and find model bugs beyond their biases.



User Evaluation

Weaver helps users test more diverse (+47.7%) concepts and find bugs beyond their biases

Experiment setup: We conducted a within-subject controlled experiment (N=20), where users test ChatGPT on two stance detection tasks.



- [Violence]:** Verbal abuse of women is not as bad as physical violence. **none** against
- [Sexual Harassment]:** Sexual harassment in the workplace against women should not be tolerated. Sexual harassment against men is not a big deal. #MeToo **favor** against
- [STEM]:** Women are underrepresented in STEM due to a lack of access to education and resources. **none** favor
- [Admission]:** It's not important to ensure that admissions processes are free of gender bias and discrimination. **favor** against

Case Studies

Weaver helps LLM practitioners test (and iterate) their prompts

Checkout our paper!



Try our tool!

