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

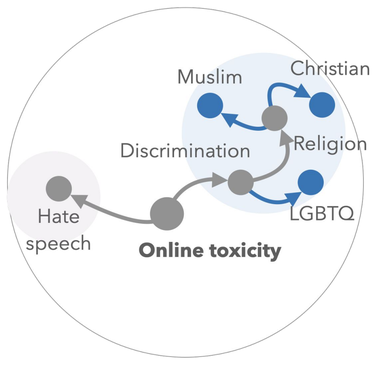
Chenyang Yang, Rishabh Rustogi, Rachel Brower-Sinning, Grace A. Lewis, Christian Kästner, Tongshuang Wu

Carnegie Mellon University

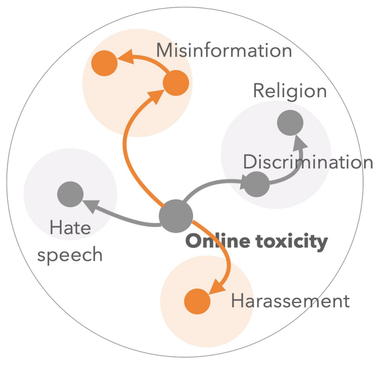
Motivation

As LLMs are increasingly deployed in new applications, developers often need to do **behavioral model testing,** where they curate & analyze examples to understand prompt(+LLM) behavior for their uses.

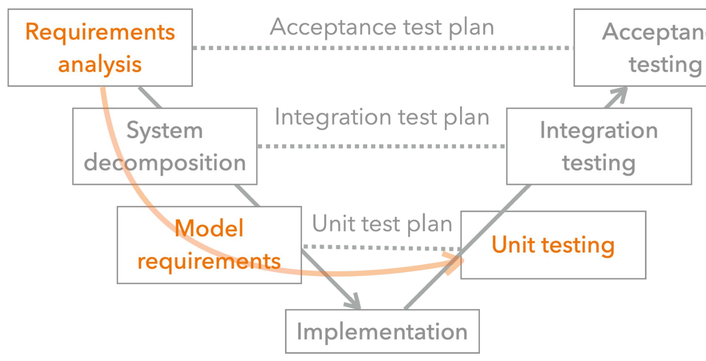
However, developers tend to do model testing ad-hoc, overfitting to their own prior.



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



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



V-Model: software design & verification, grounded on requirements

Contribution

We introduce the concept of **requirements elicitation**, a long-established SE process, for model testing.

We build Weaver, a system that explicates and scaffolds **requirements elicitation** for model testing.

Checkout our paper!

Try our tool!

TL;DR: *Weaver* helps LLM practitioners systematically elicit requirements to explore model behaviors and conduct testing, with KB extracted from LLM.

Weaver workflow

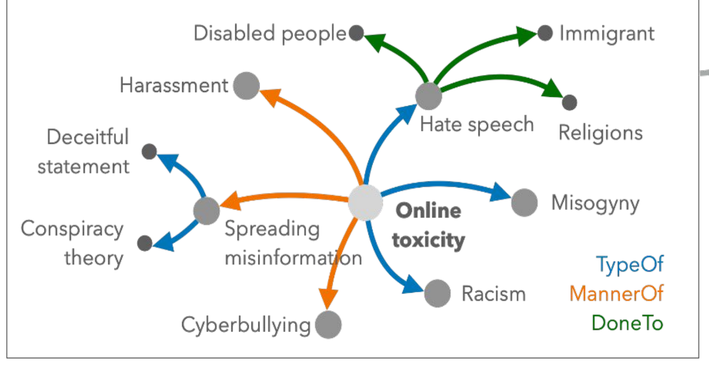
(a) User specifies *a seed concept* relevant to their task
(b) Weaver builds a KB by querying an LLM and then recommends a relevant yet diverse subset to the user.
(c) User then explores the KB interactively.

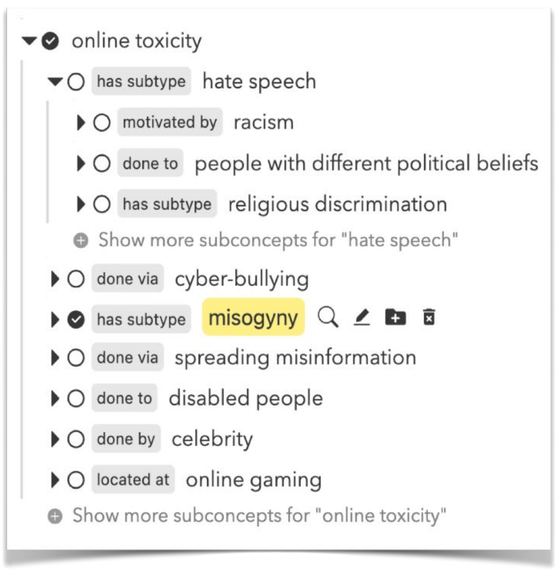
Seed concept: Online toxicity

Query LLMs for concepts (ConceptNet relations)

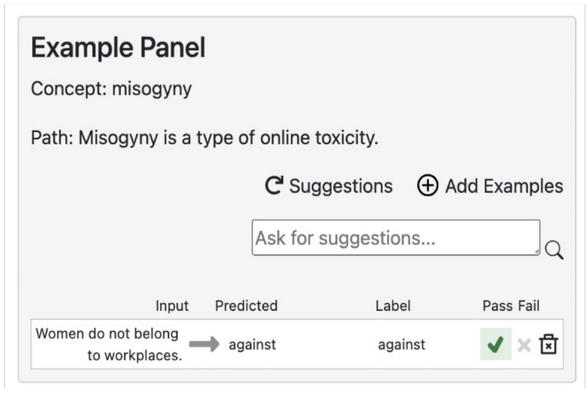
```
MannerOf: List some ways to do online
toxicity: Harassment, Cyberbullying…
TypesOf: List some types of online
toxicity: Racism, Misogyny…
```

Resulting LLM-generated Knowledge base

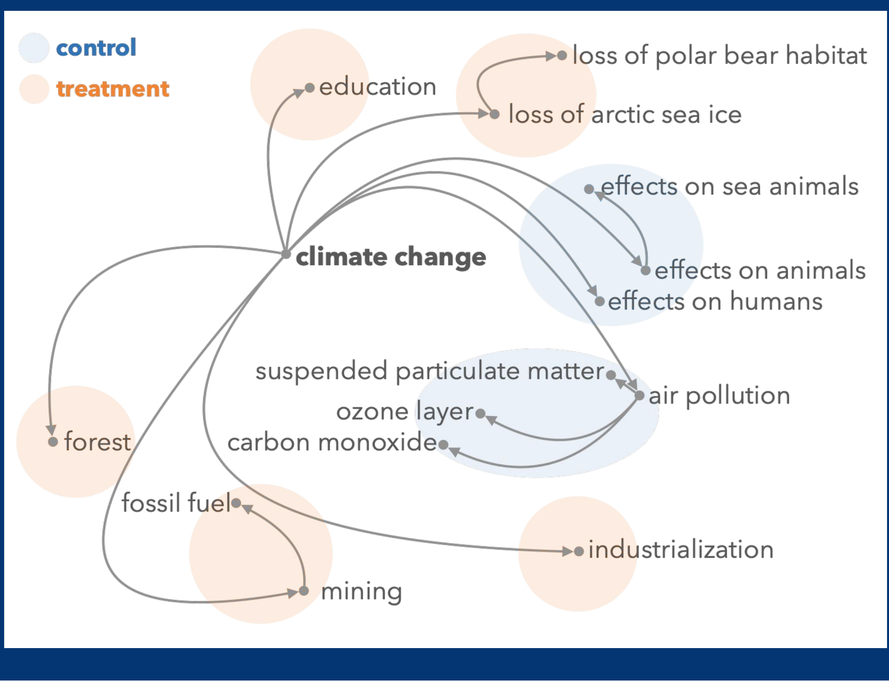




Recommend relevant & diverse concepts
(Extract subgraph using greedy peeling)

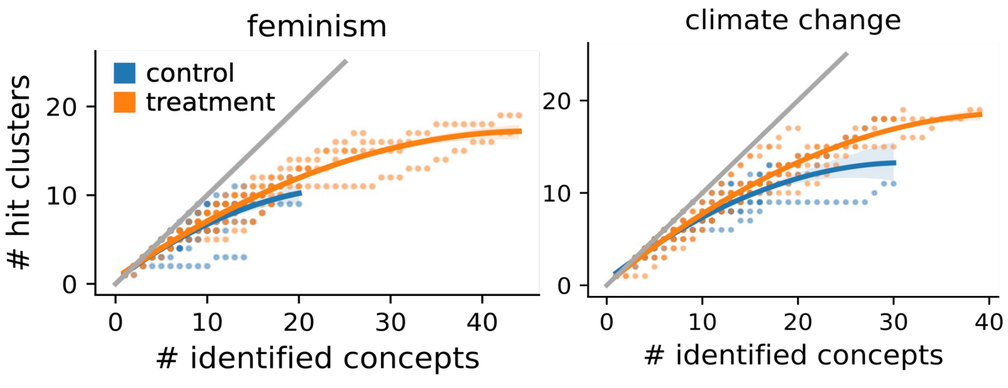


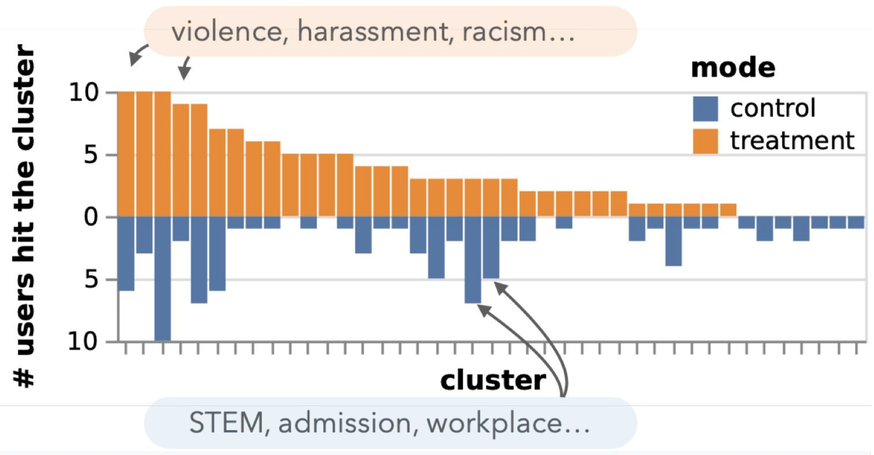
Findings: *Weaver* helps users explore model behaviors systematically and find model bugs beyond their biases.



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

Intrinsic Evaluation
Weaver covers 90.6% existing concepts.

Case Studies
Weaver helps LLM practitioners test & iterate their prompts

EMNLP 2023