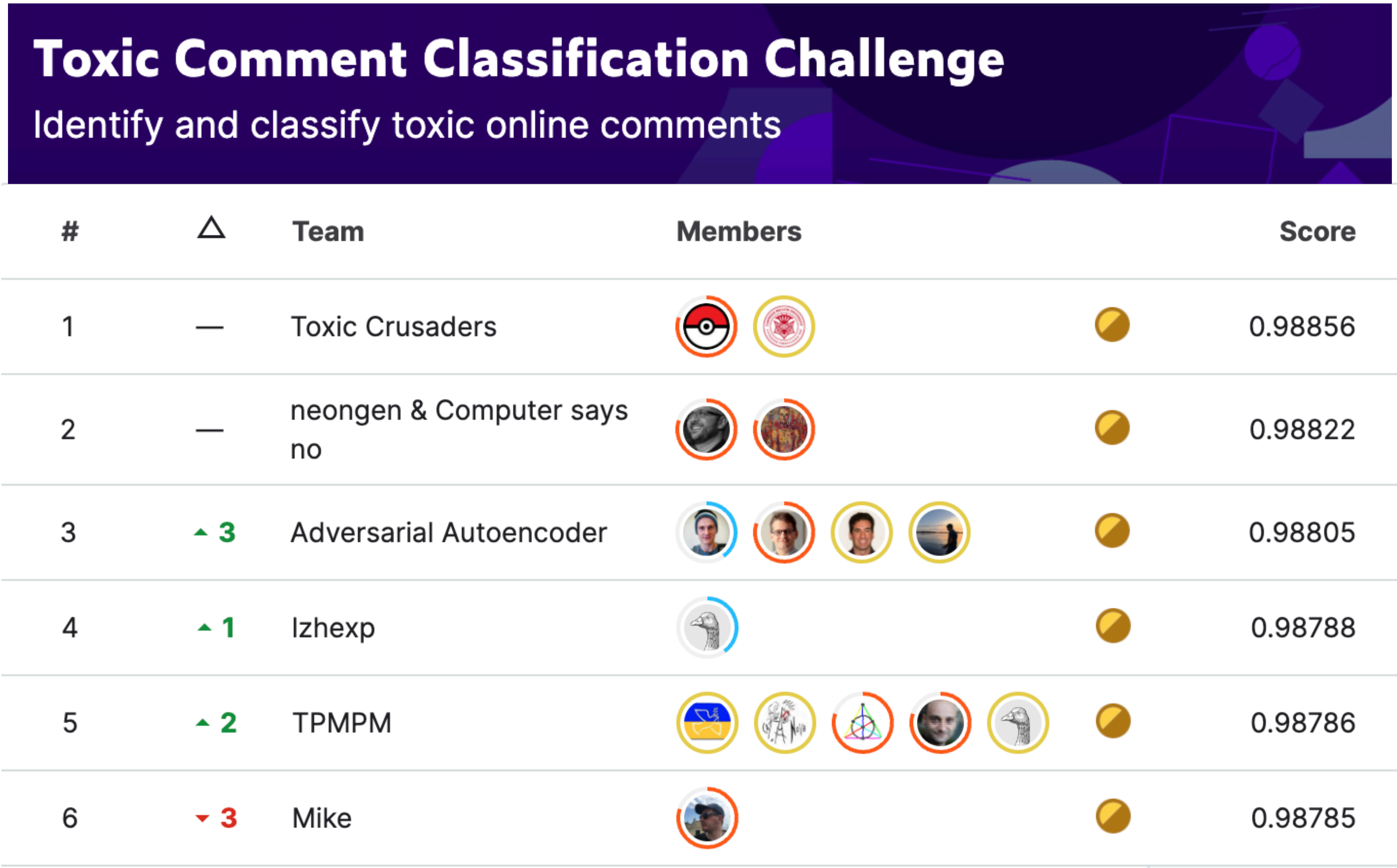
Beyond Testers' Biases: Guiding Model Testing with Knowledge Bases using LLMs

Chenyang Yang, Rishabh Rustogi, Rachel Brower-Sinning, Grace Lewis,
Christian Kaestner, Tongshuang (Sherry) Wu

Coarse-grained Model Evaluation is not Enough

Model evaluations use **a single score** to compare and rank different ML models.

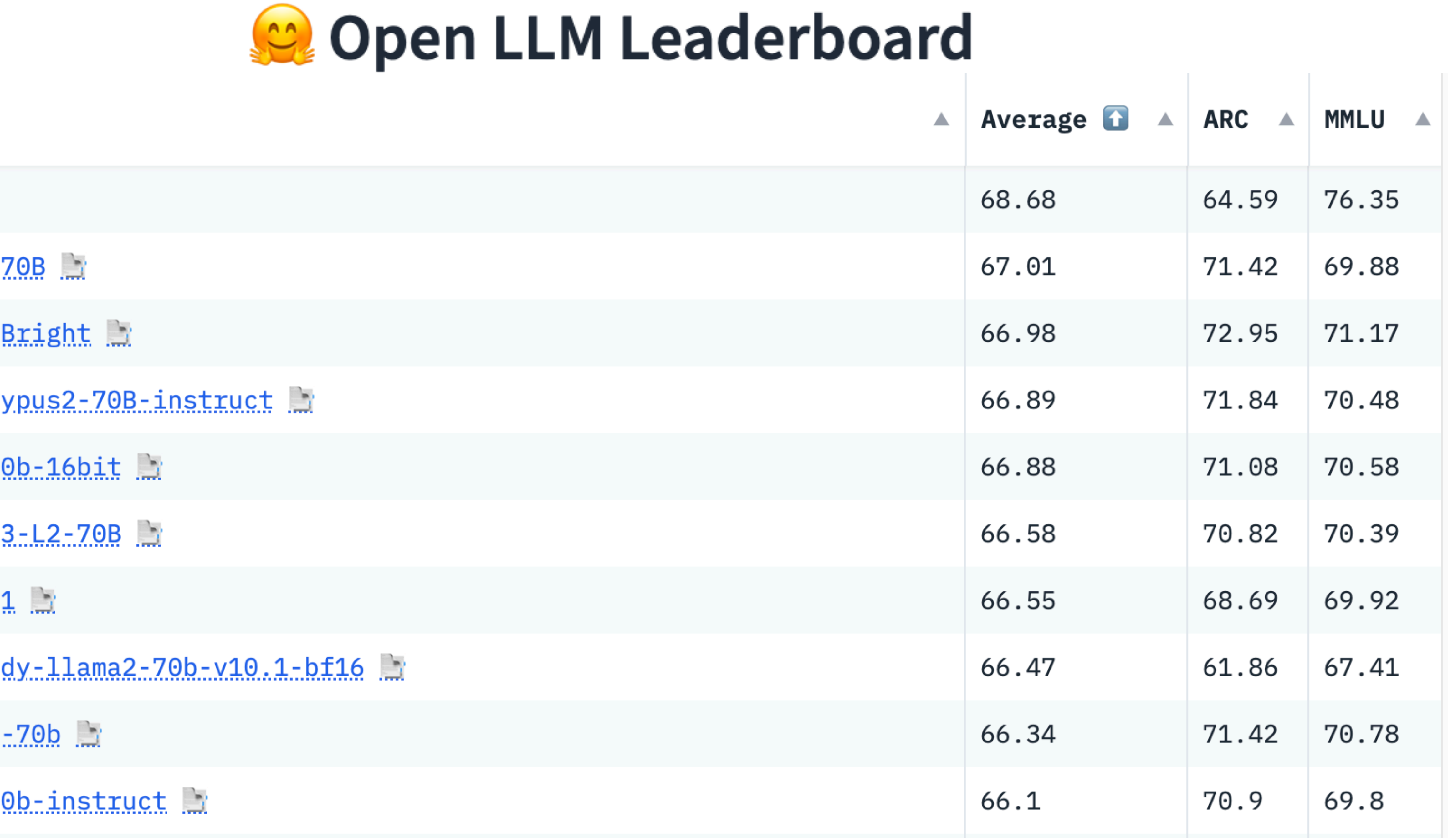
Coarse-grained evaluations can not provide insights for **models' strengths and weaknesses**, which are useful for model analysis, debugging, and selection.



Toxic Comment Classification Challenge

Identify and classify toxic online comments

#	△	Team	Members	Score
1	—	Toxic Crusaders		0.98856
2	—	neongen & Computer says no		0.98822
3	▲ 3	Adversarial Autoencoder		0.98805
4	▲ 1	Izhexp		0.98788
5	▲ 2	TPMPM		0.98786
6	▼ 3	Mike		0.98785



🤗 Open LLM Leaderboard

	Average	ARC	MMLU
	68.68	64.59	76.35
70B	67.01	71.42	69.88
Bright	66.98	72.95	71.17
ypus2-70B-instruct	66.89	71.84	70.48
0b-16bit	66.88	71.08	70.58
3-L2-70B	66.58	70.82	70.39
1	66.55	68.69	69.92
dy-llama2-70b-v10.1-bf16	66.47	61.86	67.41
-70b	66.34	71.42	70.78
upstage/Llama-2-70b-instruct	66.1	70.9	69.8

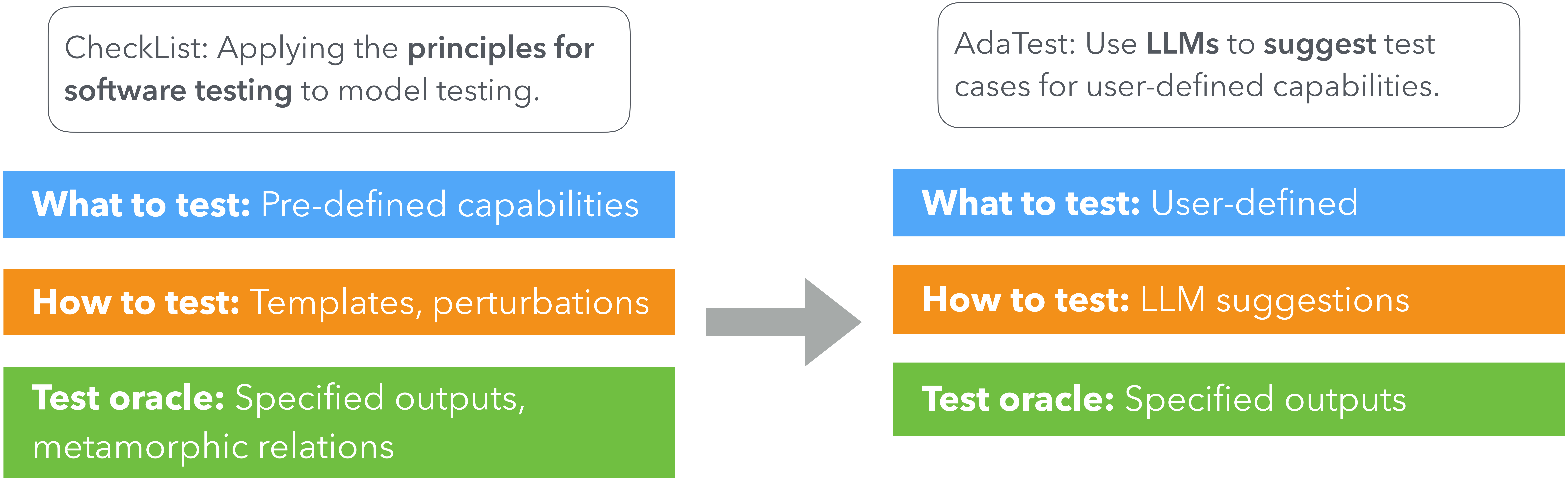
Beyond Accuracy: Behavioral Model Testing

Capabilities	Descriptions	Examples
Vocab/POS	important words or word types for the task.	template('This is a {adj:mask} movie.')
Named entities	appropriately understanding named entities.	perturb('{John} doesn't like the movie', change_name)
Negation	understand the negation words.	template('The food is not {adj:mask}.')
Taxonomy	synonyms, antonyms, etc.	perturb('How can I become more {optimistic}?', antonym)
Robustness	to typos, irrelevant changes, etc.	perturb('@SouthwestAir no {thanks}', replace_char)
Coreference	resolve ambiguous pronouns, etc.	
Fairness	not biasing towards certain gender/race groups.	
Semantic Role Labeling	understanding roles such as agent, object, etc.	
Logic	handle symmetry, consistency, and conjunctions.	
Temporal	understand order of events.	template('I used to hate the {noun:mask}, but now I like it')

CheckList: Applying the **principles for software testing** to model testing. Create test cases for **concrete model behavior**.

[ACL'20] Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

Beyond Accuracy: Behavioral Model Testing



CheckList: Applying the **principles for software testing** to model testing.

What to test: Pre-defined capabilities

How to test: Templates, perturbations

Test oracle: Specified outputs, metamorphic relations

AdaTest: Use **LLMs** to **suggest** test cases for user-defined capabilities.

What to test: User-defined

How to test: LLM suggestions

Test oracle: Specified outputs

Different work varies on these three dimensions

Model Testing is Ad-hoc and Biased

CheckList: Applying the **principles for software testing** to model testing.

What to test: Pre-defined capabilities

How to test: Templates, perturbations

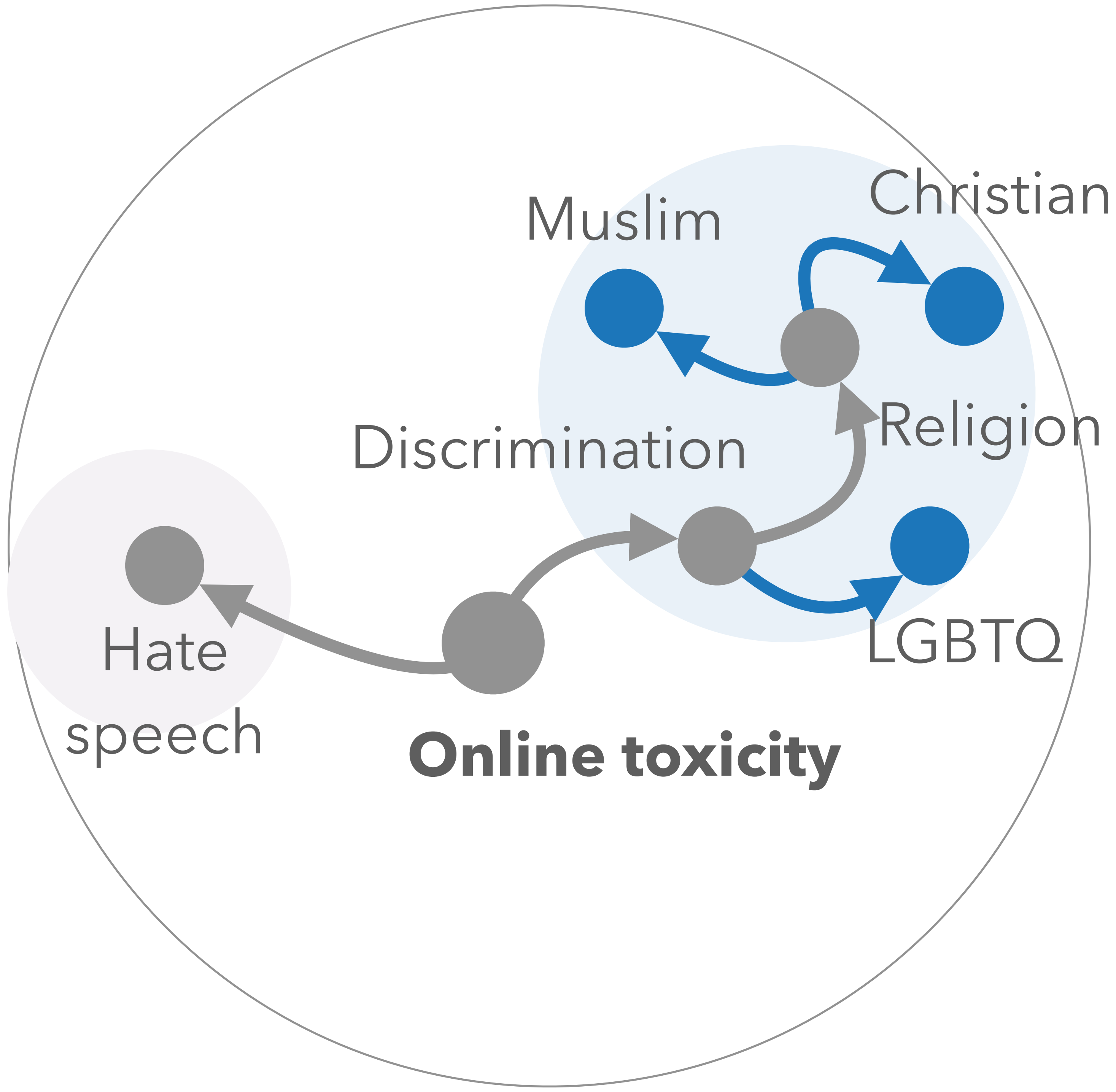
AdaTest: Use **LLMs** to **suggest** test cases for user-defined capabilities.

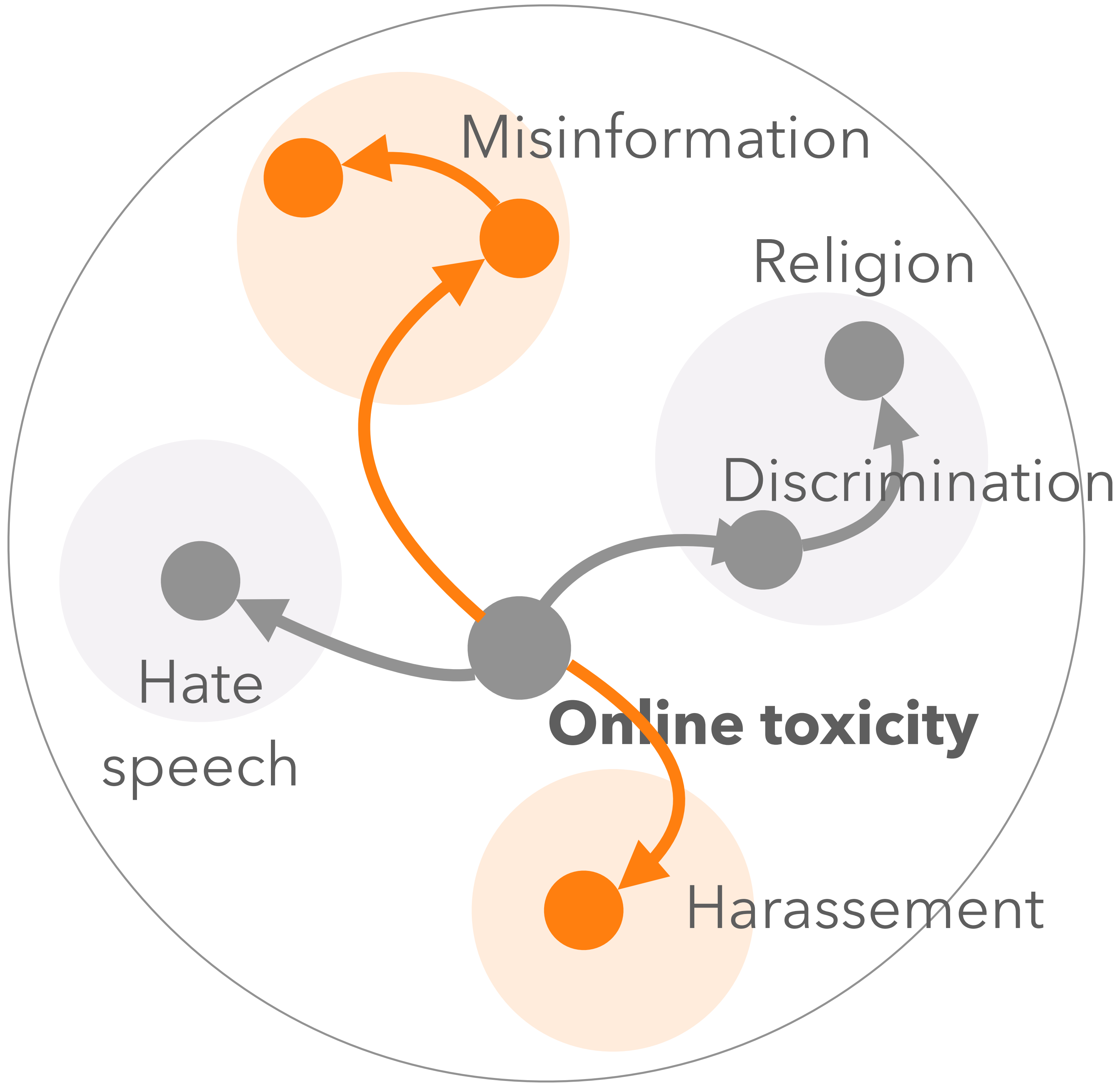
What to test: User-defined

How to test: LLM suggestions

Existing model testing methods focus on **how to test**, exploring different test generation methods. But how do testers know **what to test**?

Model Testing is Ad-hoc and Biased





Users tend to explore **locally**

overfit to their intuition, domain knowledge, confirmation bias.

Expect: **Comprehensive testing**

More systematically cover the space beyond individual biases

Model Testing is Ad-hoc and Biased

Muslim
Christian
Discrimination
Religion

Misinformation
Religion

How to move from **local exploitation** towards **global exploration**?

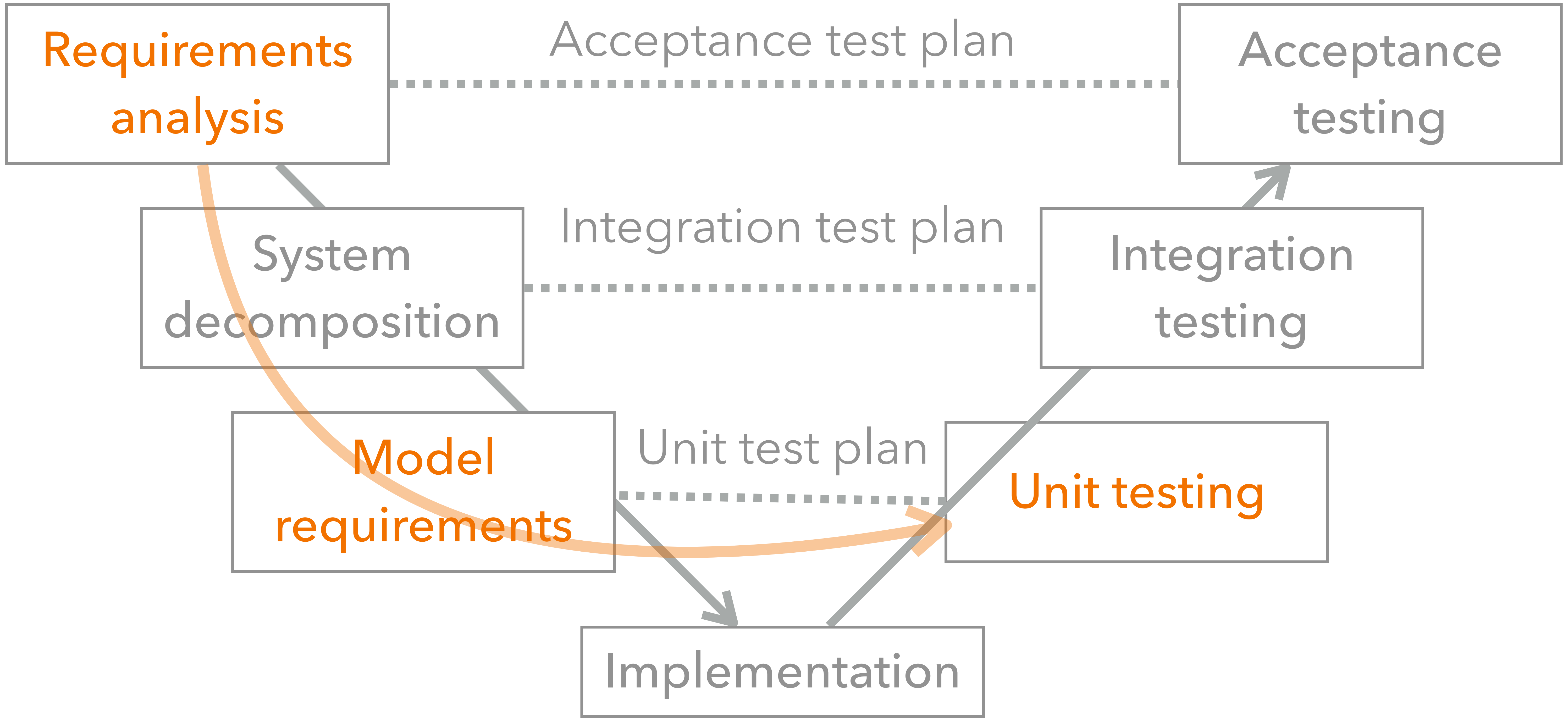
Users tend to explore **locally**

overfit to their intuition, domain knowledge, confirmation bias.

Expect: **Comprehensive testing**

More systematically cover the space beyond individual biases

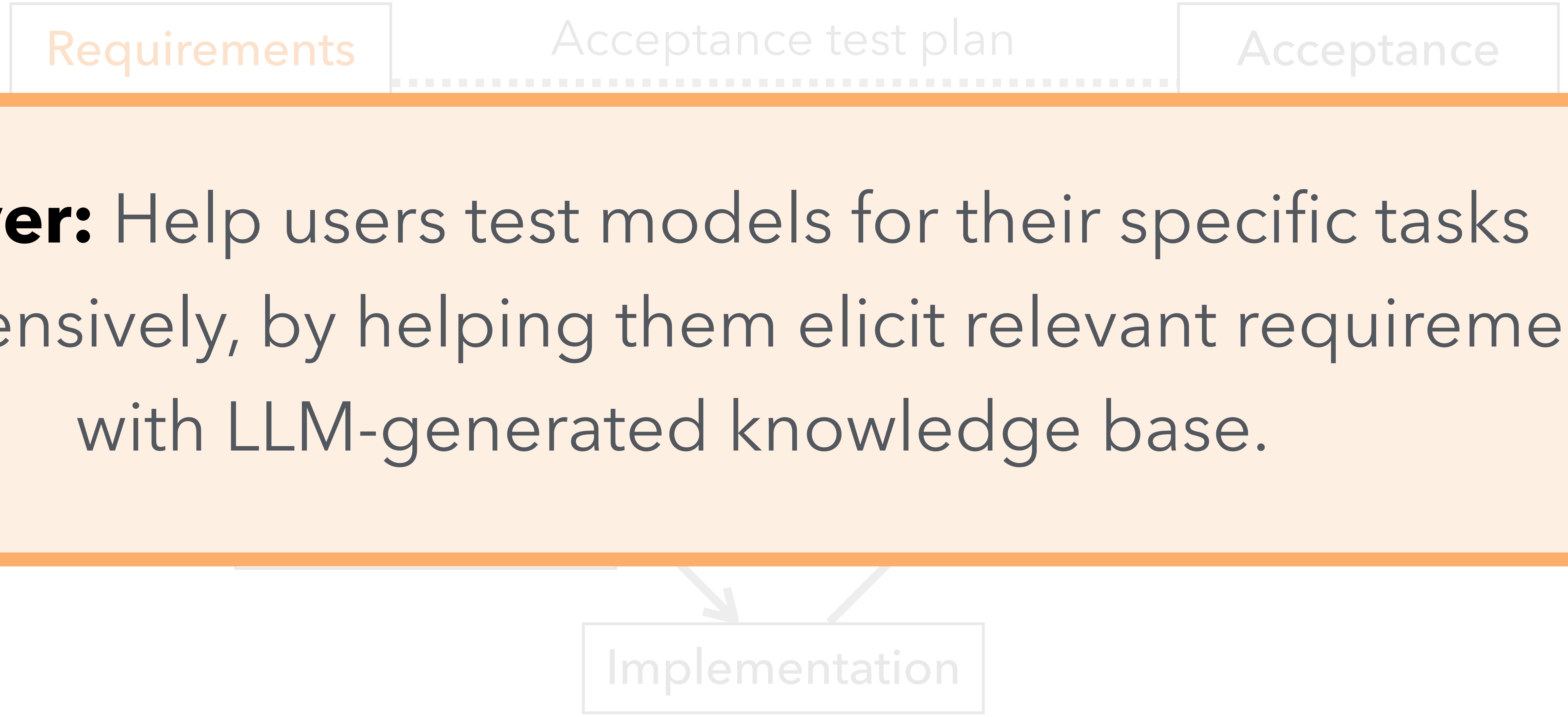
SE to NLP: Requirements Engineering



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

SE to NLP: Requirements Engineering

Weaver: Help users test models for their specific tasks comprehensively, by helping them elicit relevant requirements with LLM-generated knowledge base.



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

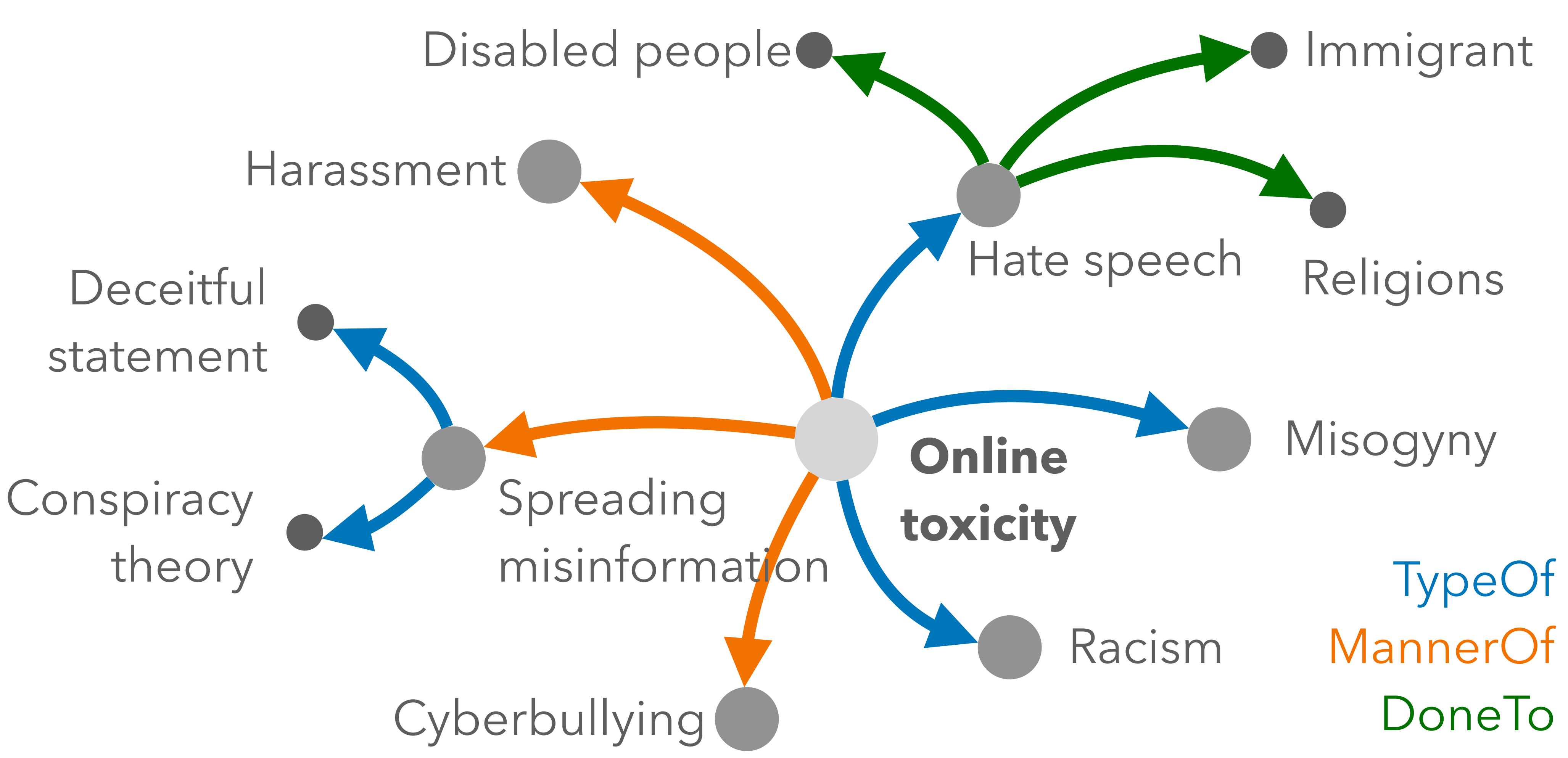
Weaver Workflow

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



Intuition: LLMs have **parametric knowledge** for various domains, tasks, and topics.

Traditional **knowledge base relation** helps **extract the knowledge** comprehensively.

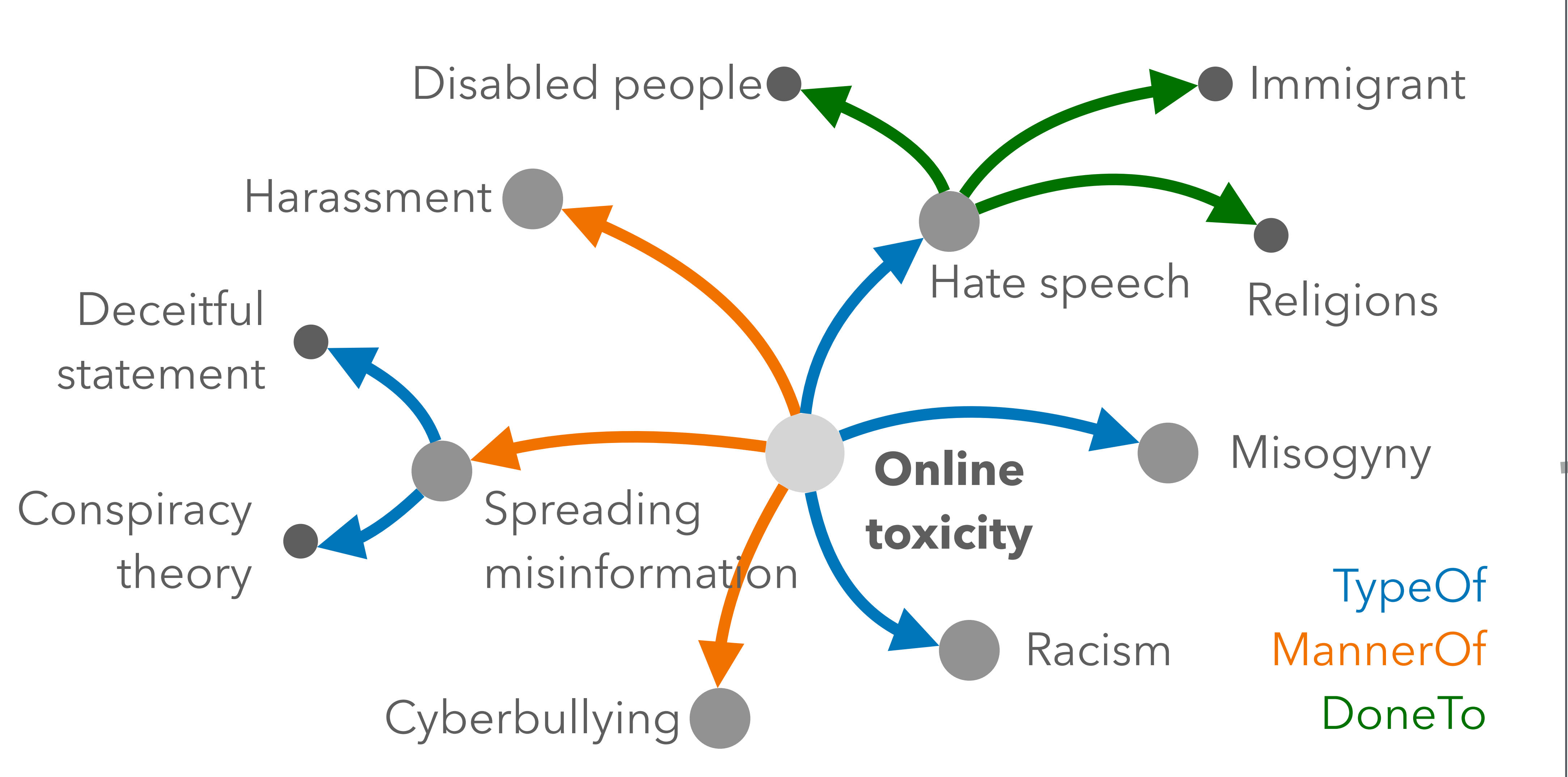
Weaver Workflow

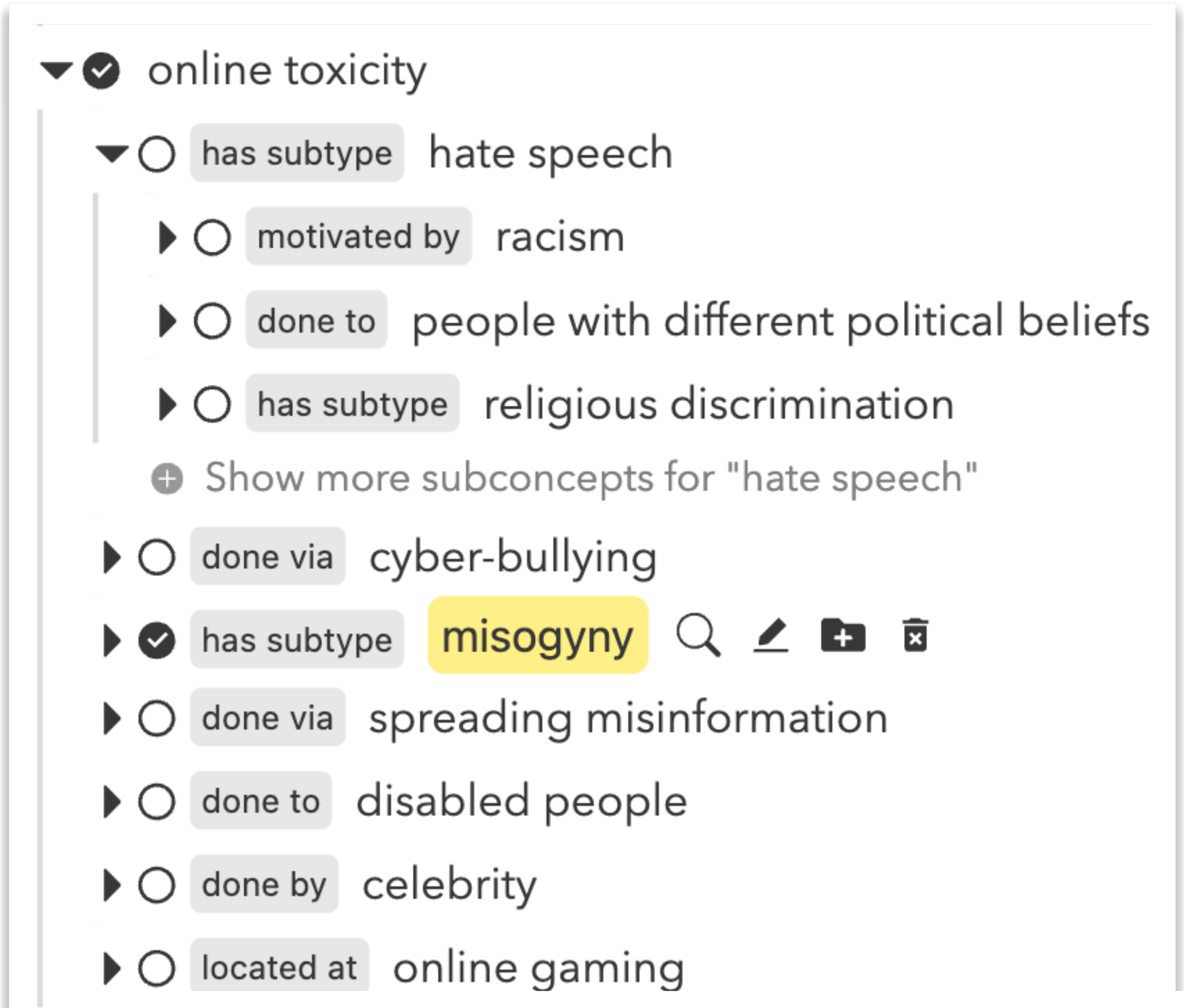
Seed concept: Online toxicity

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

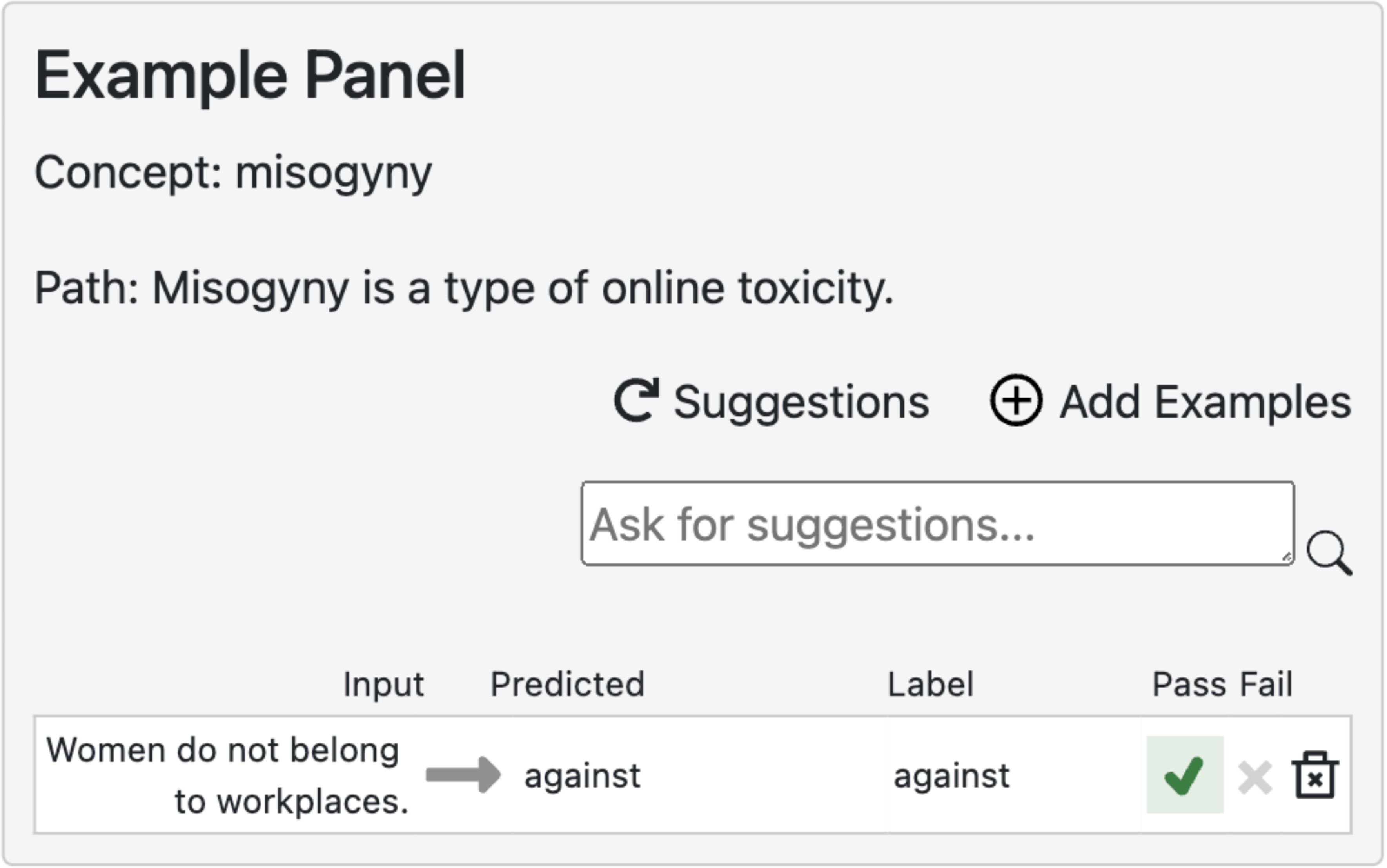




- ▾ ● online toxicity
  - ▾ ○ has subtype hate speech
    - ▸ ○ motivated by racism
    - ▸ ○ done to people with different political beliefs
    - ▸ ○ has subtype religious discrimination
    - ⊕ Show more subconcepts for "hate speech"
  - ▸ ○ done via cyber-bullying
  - ▸ ● has subtype misogyny
  - ▸ ○ done via spreading misinformation
  - ▸ ○ done to disabled people
  - ▸ ○ done by celebrity
  - ▸ ○ located at online gaming
  - ⊕ S

Reco

(Extrac



Example Panel

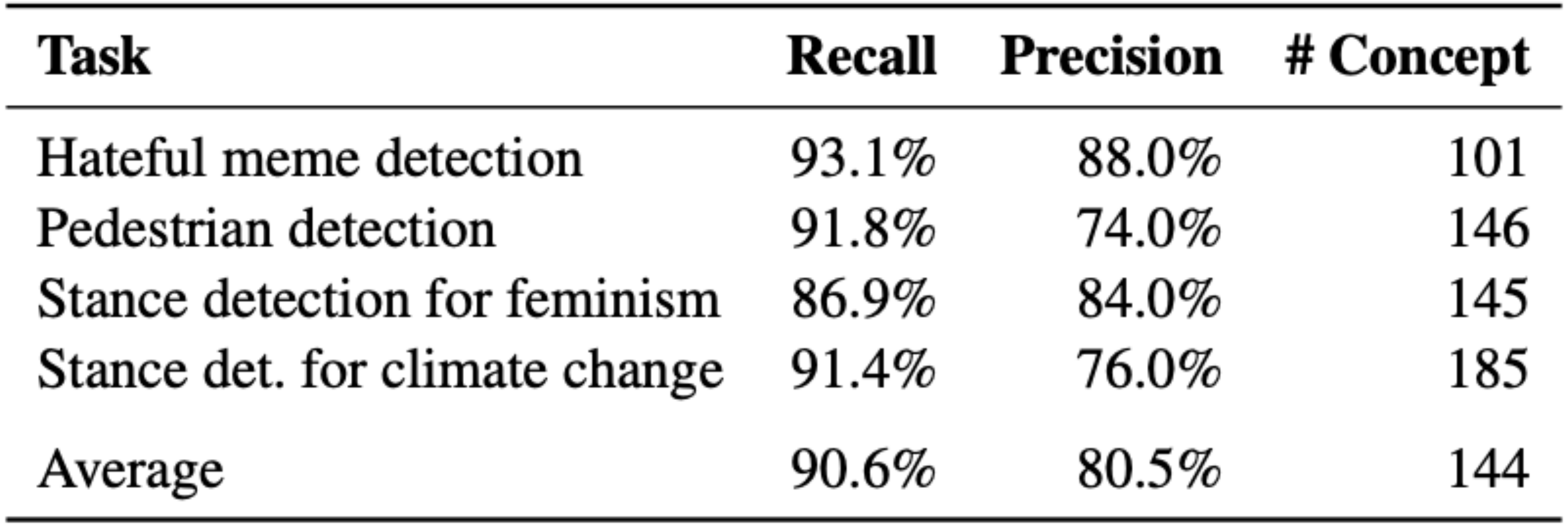
Concept: misogyny

Path: Misogyny is a type of online toxicity.

Input	Predicted	Label	Pass Fail
Women do not belong to workplaces. →	against	against	✓ ✗

Weaver recommends important concepts

Comparing **Weaver concepts** vs.
gold concepts identified from existing dataset analysis and user studies



Task	Recall	Precision	# Concept
Hateful meme detection	93.1%	88.0%	101
Pedestrian detection	91.8%	74.0%	146
Stance detection for feminism	86.9%	84.0%	145
Stance det. for climate change	91.4%	76.0%	185
Average	90.6%	80.5%	144

Weaver covers most of the important concepts even when we only grow the knowledge base to the second layer!

Weaver supports systematic bug finding

We conducted a within-subject controlled experiment (N=20) to see whether Weaver helps users…

Explore more concepts?

Explore more diverse concepts?

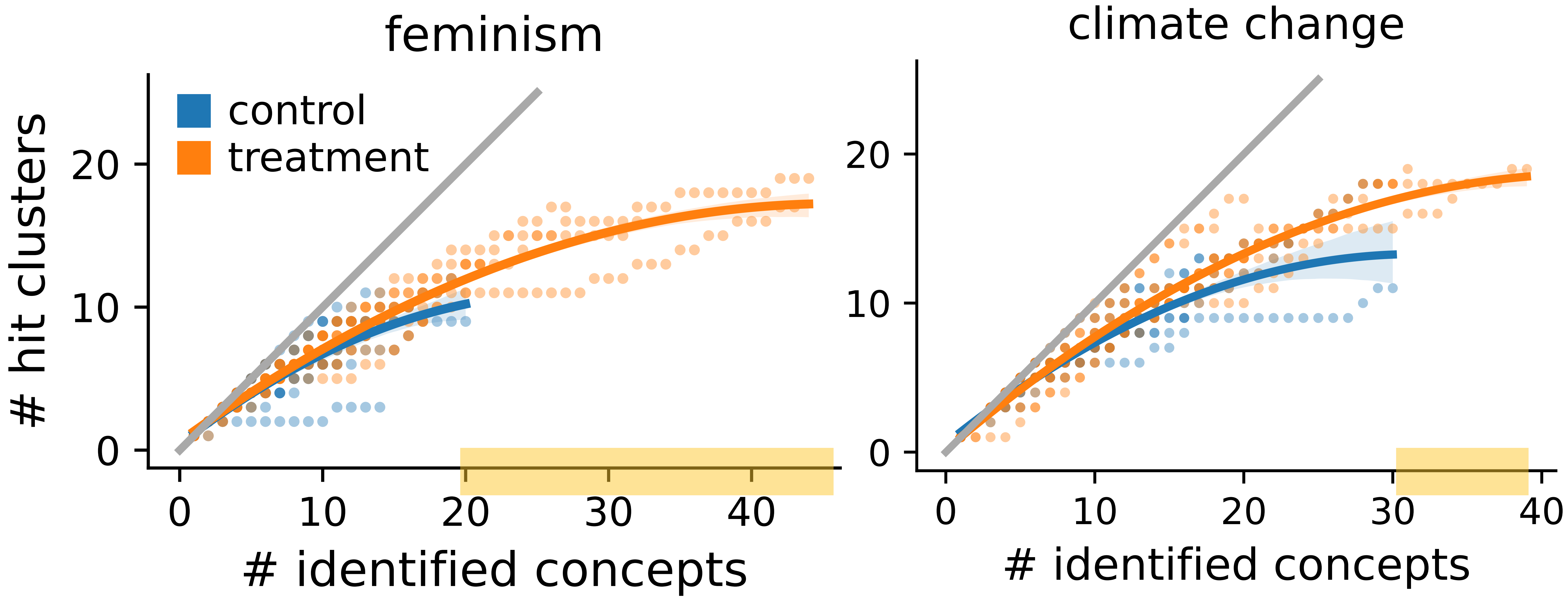
Mitigate their biases?

ChatGPT

Stance detection for feminism
Stance det. for climate change

Weaver supports systematic bug finding

vs. Manually adding concepts while exploring model errors (on LLM-generated inputs), Weaver…

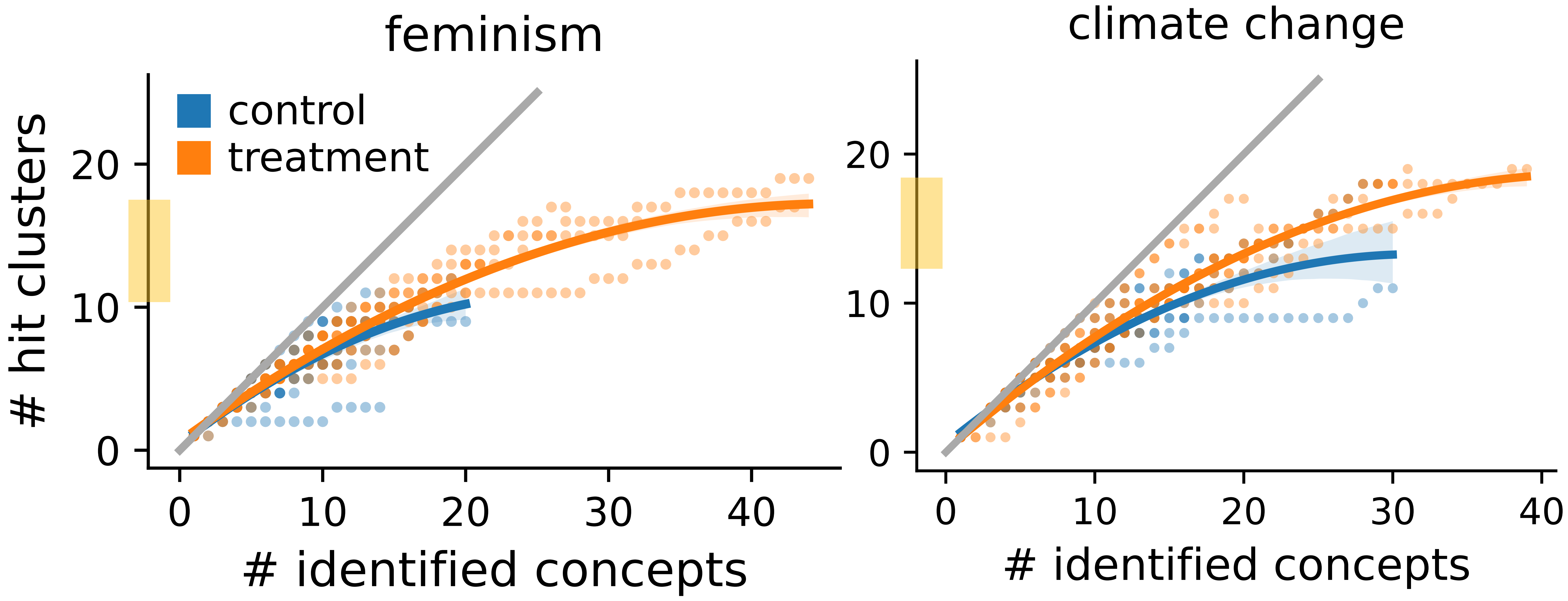


Helps **identify 57.5% more concepts** in the same amount of time

Weaver supports systematic bug finding

vs. Manually adding concepts while exploring model errors (on LLM-generated inputs), Weaver…

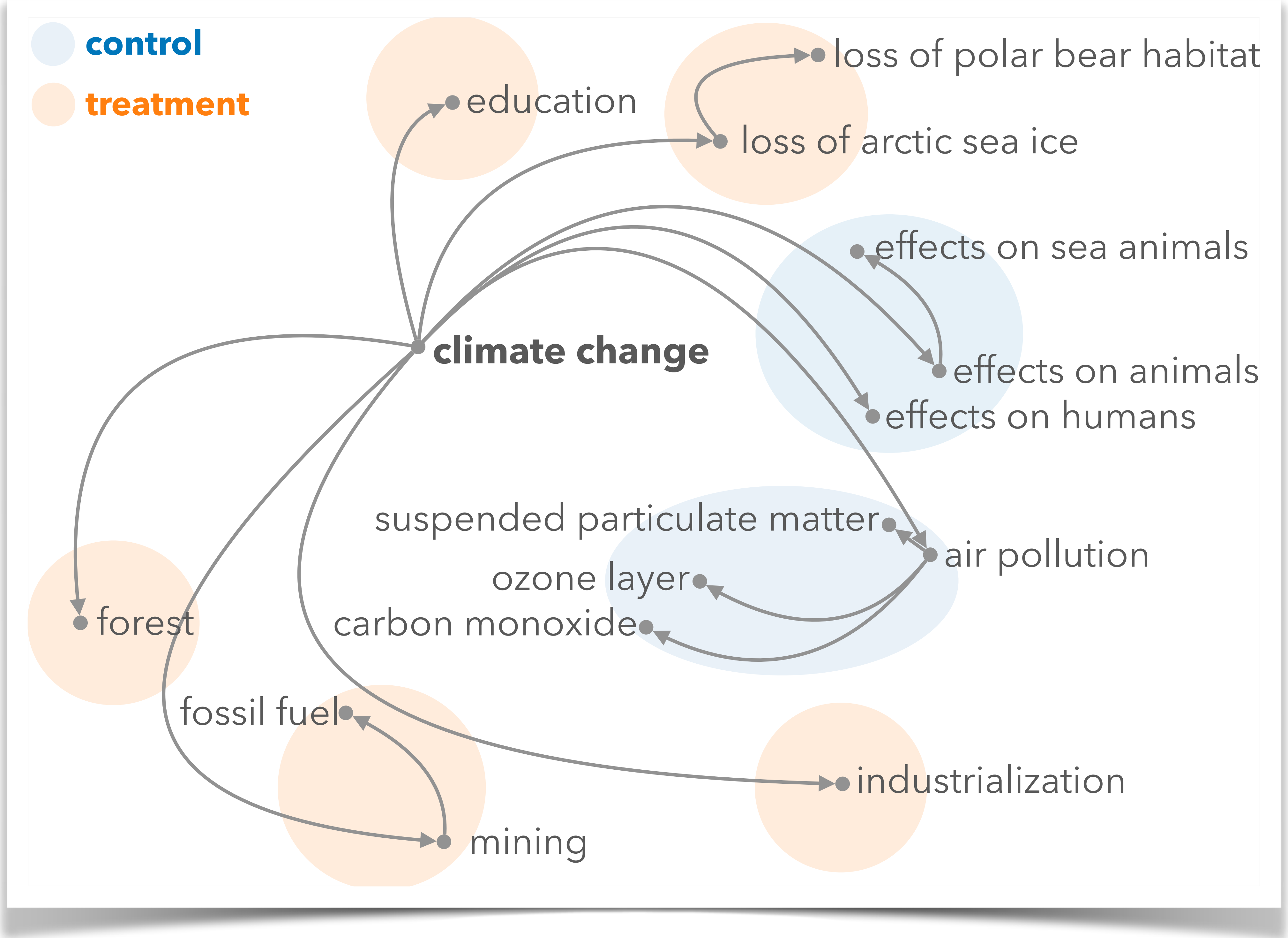
These concepts are more diverse
(covered 47.7% more clusters)



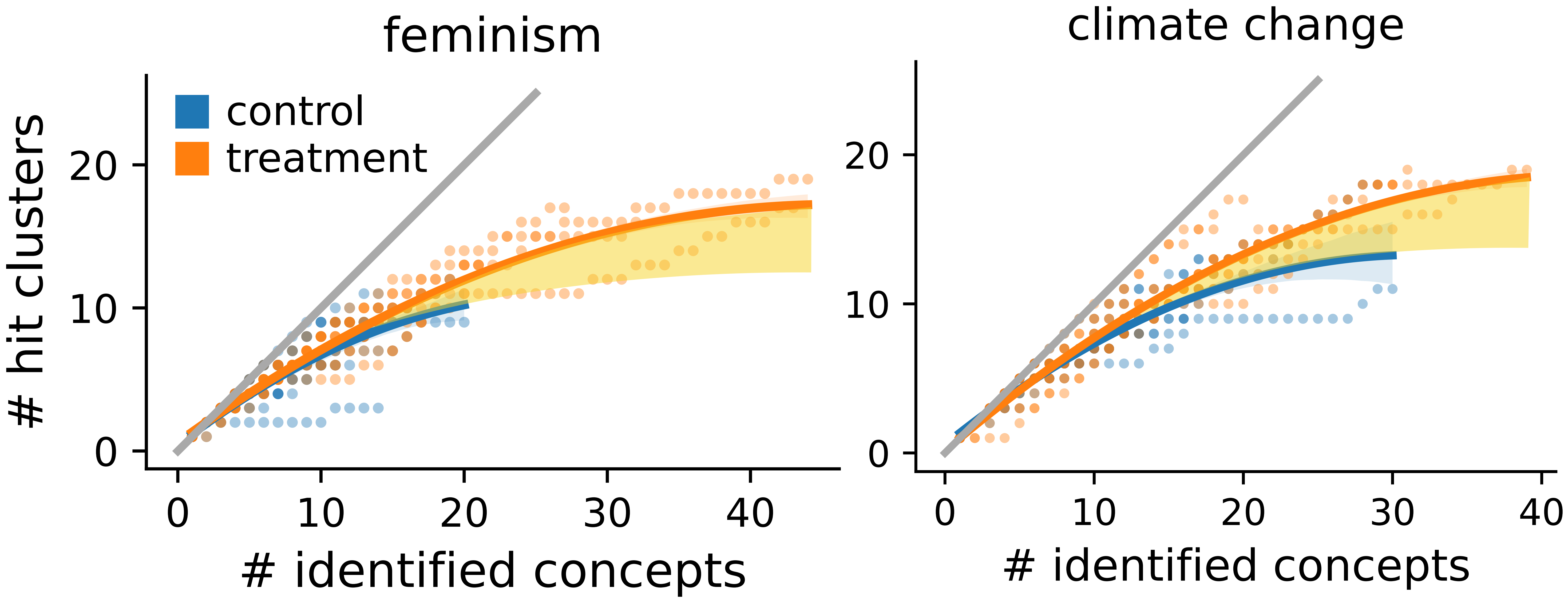
Helps **identify 57.5% more concepts** in the same amount of time

Weaver supports systematic bug finding

vs. Manually adding concepts while exploring model errors (on LLM-generated inputs), Weaver…



Enables users to **continuously discover distinct concepts** (vs. control: focus on refining existing concepts)

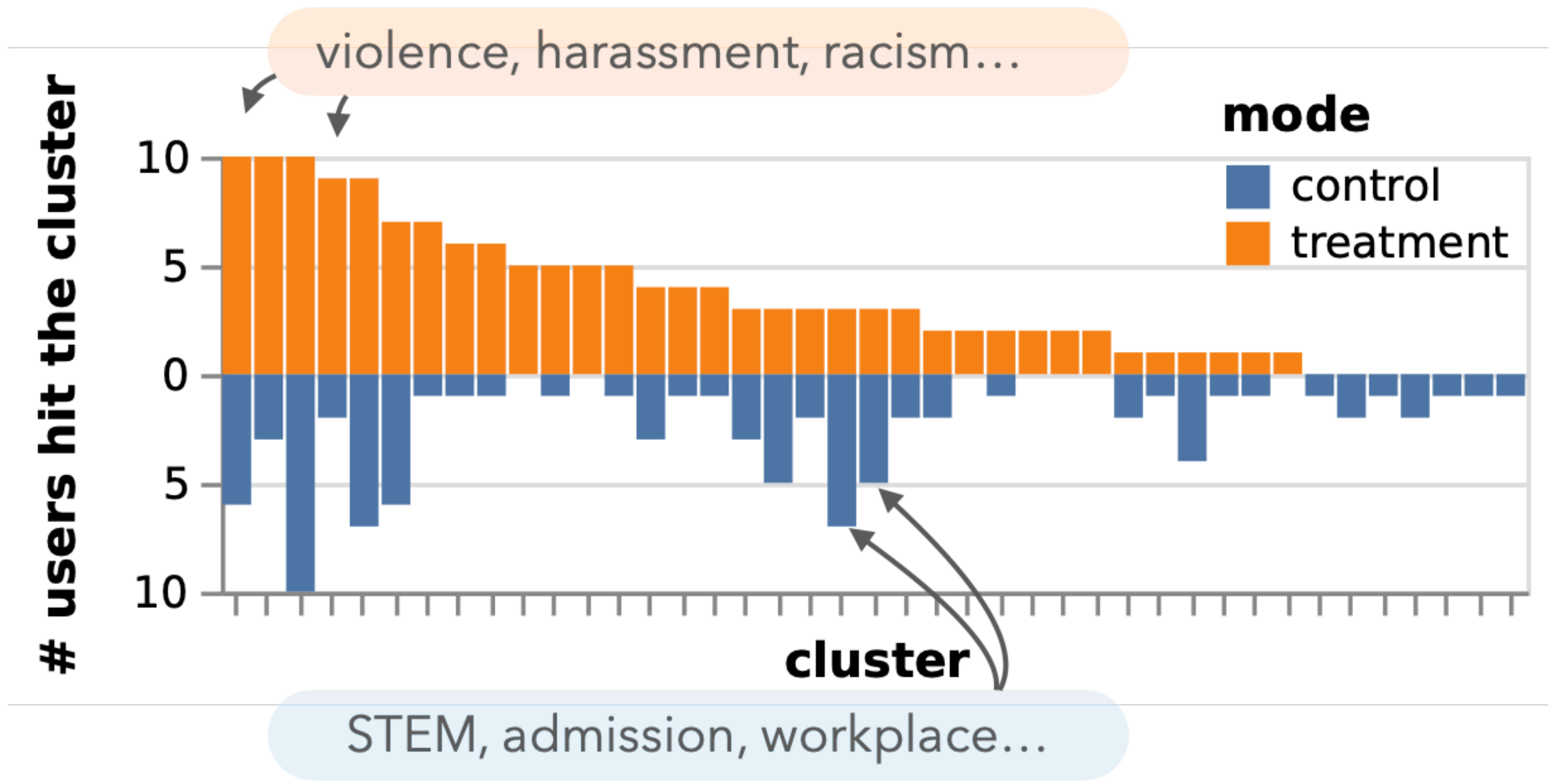


Helps **identify 57.5% more concepts** in the same amount of time

Weaver supports systematic bug finding

vs. Manually adding concepts while exploring model errors (on LLM-generated inputs), Weaver…

Enables users to explore concepts beyond their biases



[**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

Weaver helps practitioners test (and iterate) prompts

We conducted two case studies to see whether Weaver is useful in real-world settings.

C1: Prompt LLMs to summarize transcripts

C2: Prompt LLMs to explain code

Weaver helps practitioners find new bugs

"Summaries are chronological even
when reordering is desired"

"Specific challenges that novice programmers might
have in comprehending [domain] code"

Weaver can help early-stage development

Weaver helps practitioners test (and iterate) prompts

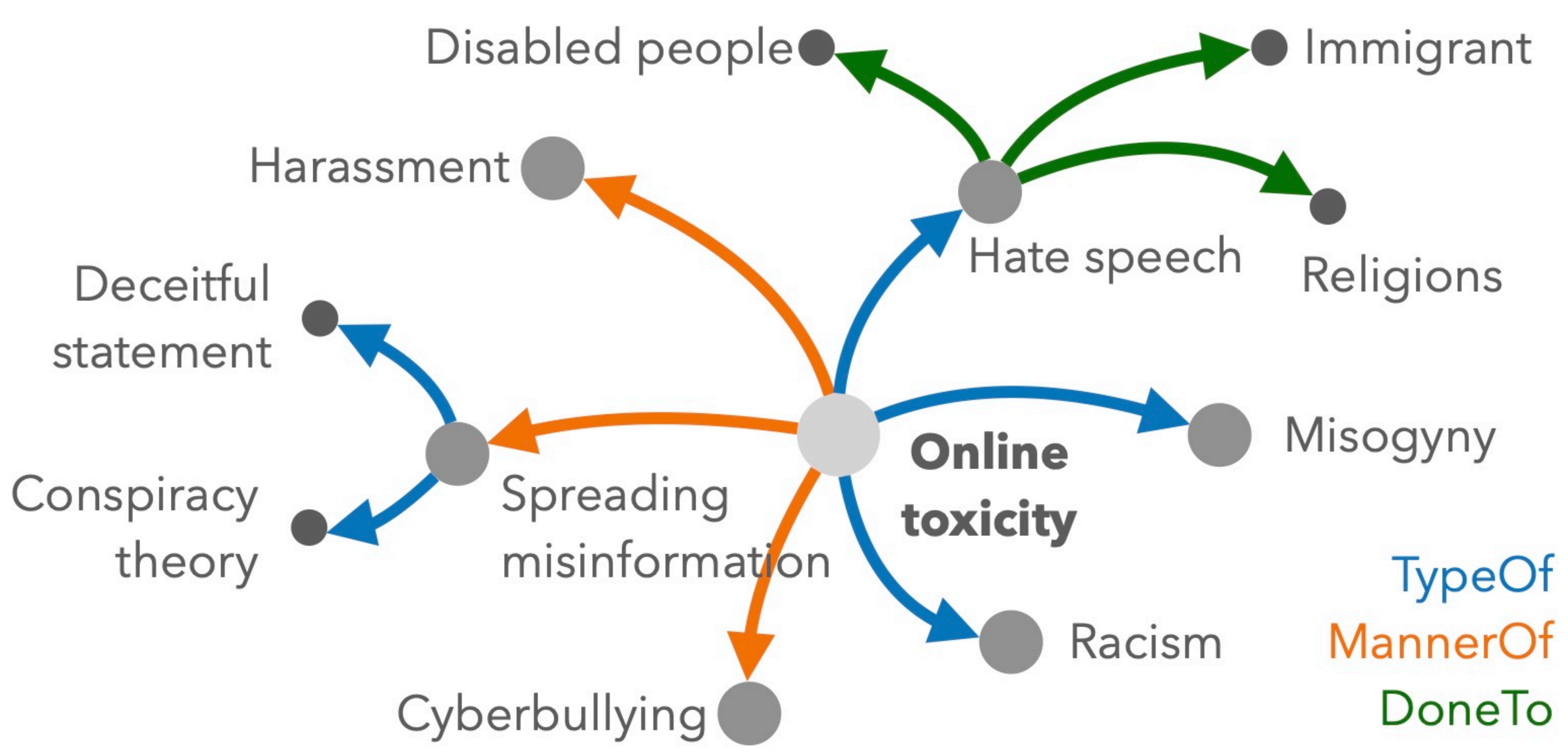
We conducted two case studies to see whether Weaver is useful in real-world settings.

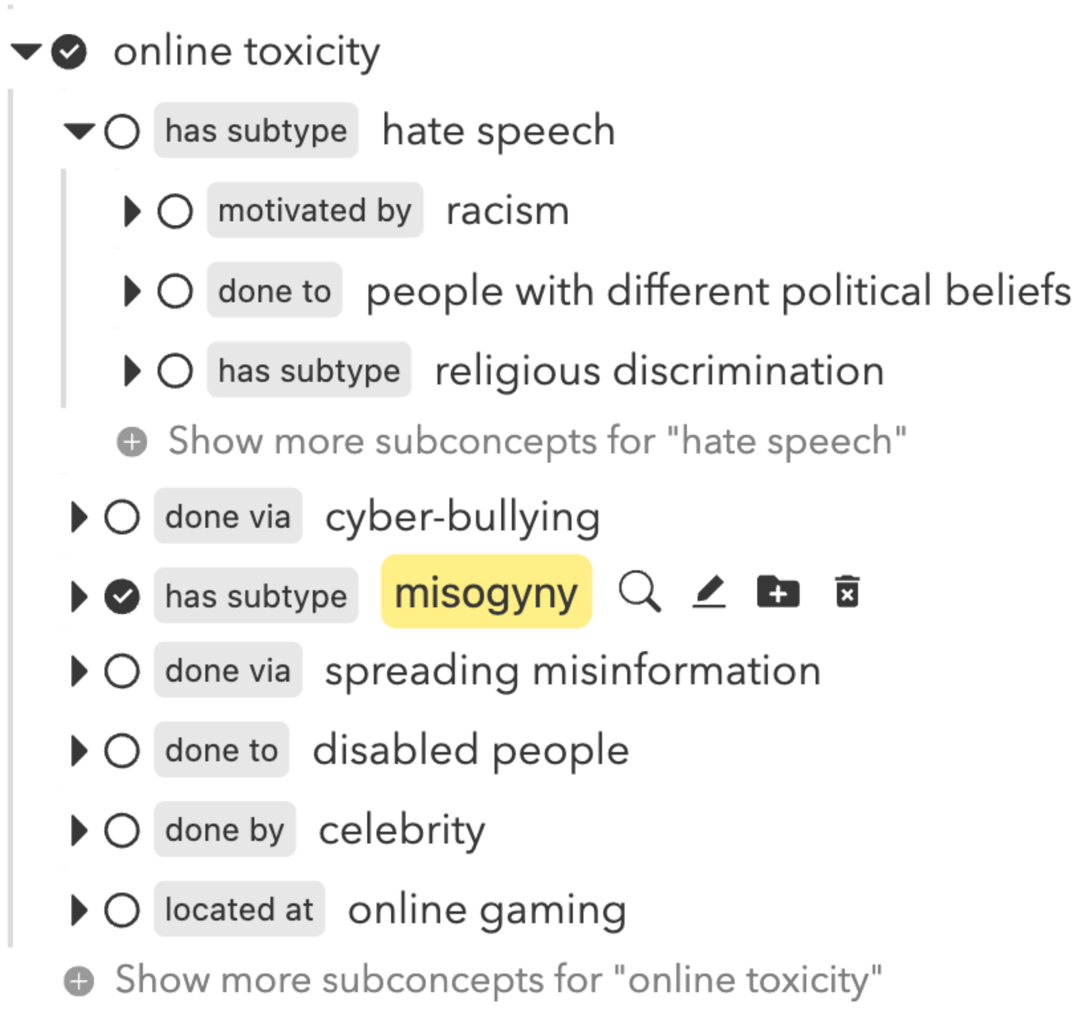
C1: Prompt LLMs to summarize transcripts

C2: Prompt LLMs to explain code

Looking for more users!

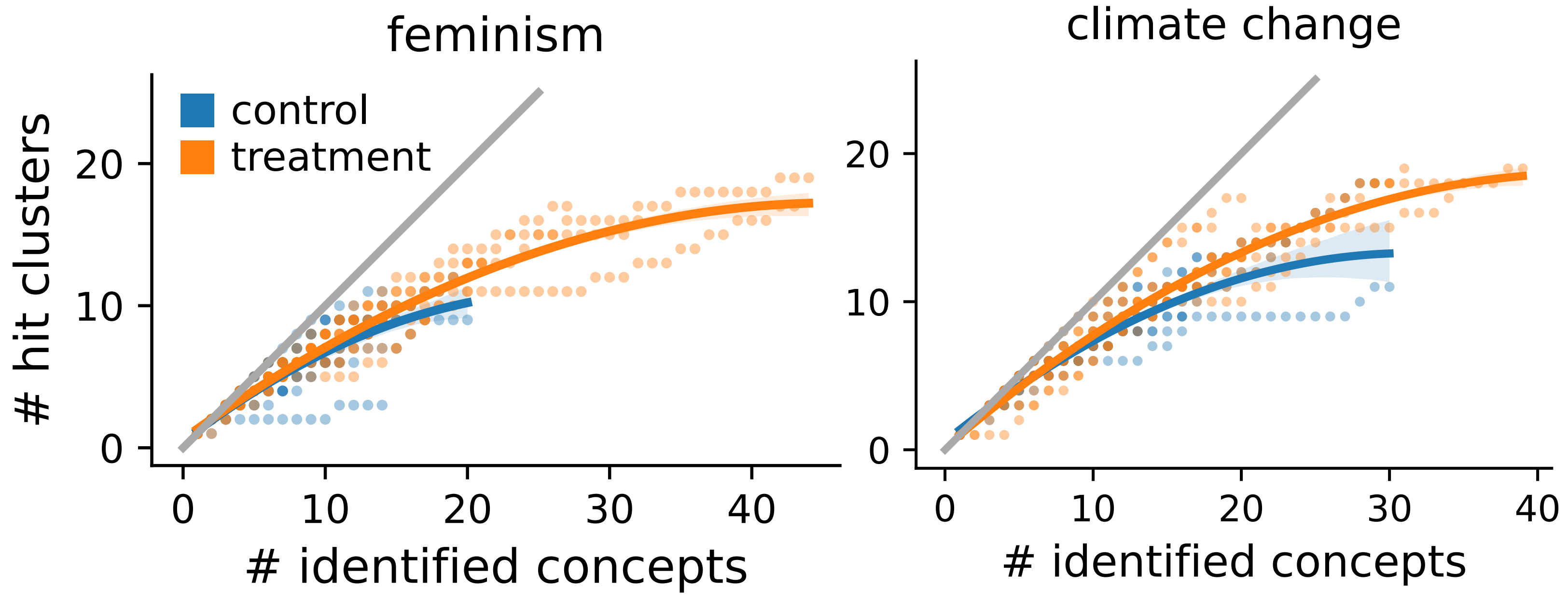
Takeaways





- ▾ ● online toxicity
  - ▾ ○ has subtype hate speech
    - ▸ ○ motivated by racism
    - ▸ ○ done to people with different political beliefs
    - ▸ ○ has subtype religious discrimination
    - ⊕ Show more subconcepts for "hate speech"
  - ▸ ○ done via cyber-bullying
  - ▸ ● has subtype misogyny
  - ▸ ○ done via spreading misinformation
  - ▸ ○ done to disabled people
  - ▸ ○ done by celebrity
  - ▸ ○ located at online gaming
  - ⊕ Show more subconcepts for "online toxicity"

Weaver: **Extract knowledge base from LLM** for interactive requirements elicitation for **systematic model testing**



[**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

Weaver helps users & practitioners…

identify **more** concepts and **more diverse** concepts
find bugs **beyond their biases**
test (and iterate) **prompts in domain applications**