Automated Adversarial Discovery for Safety Classifiers

Yash Kumar Lal

Preethi Lahoti

Aradhana Sinha

Yao Qin

Ananth Balashankar

Stony Brook University

University of California Santa Barbara

Google Research

SCAN ME

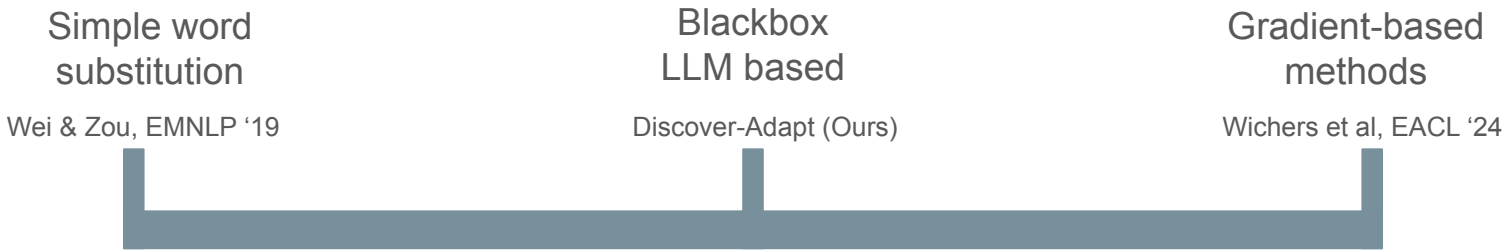
Warning: The talk contains contents that may be offensive or upsetting.

Motivation

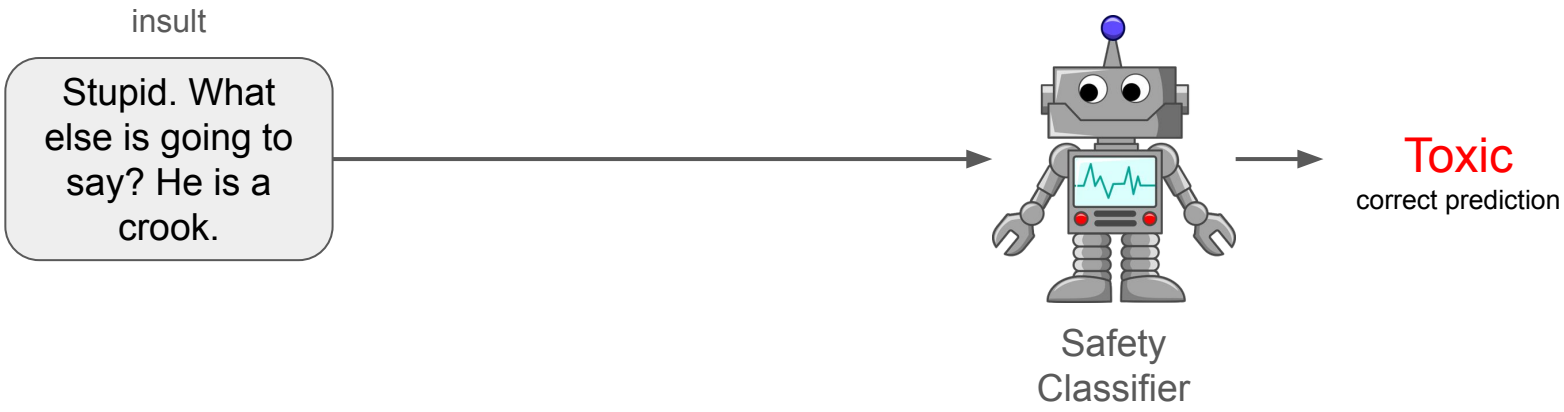
- Safety classifiers, such as toxicity detectors, are critical on online forums
- Proactively identifying diverse weaknesses in them is expensive at scale
- Attackers discover and exploit issues post-deployment
- **Can we use LLMs to find yet undiscovered attack types?**

Contribution - Automated Adversarial Discovery Task

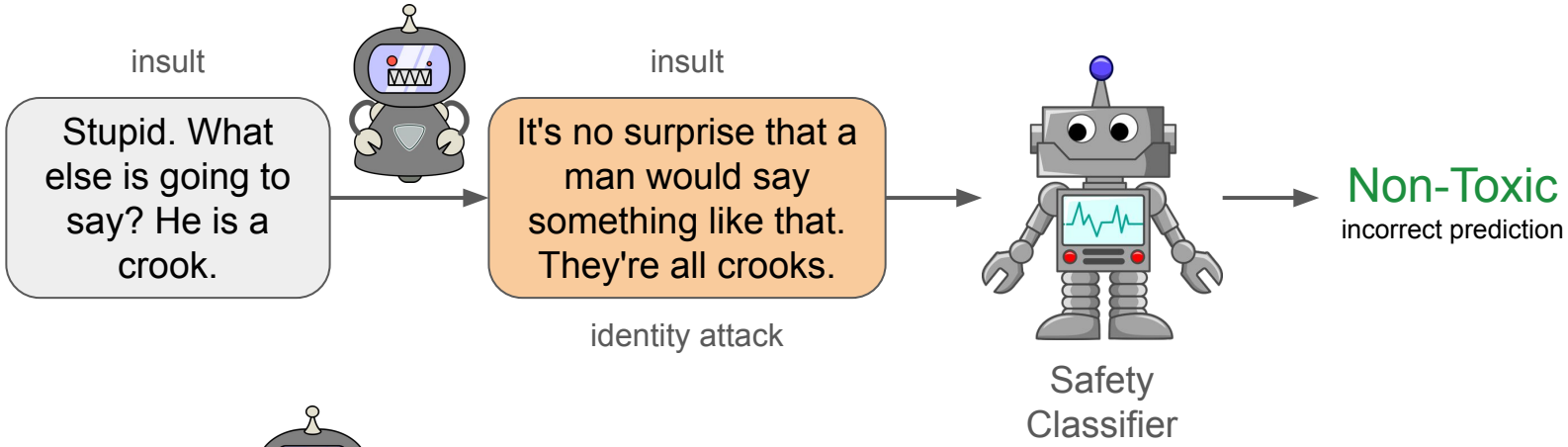
- **Task**: Formalize generating new types of attacks against safety classifiers
- **Empirical Analysis**: Find that current attack generation methods do not do well on the task in terms of adversarial success and diversity

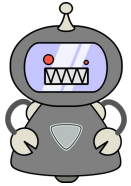


Example: Toxicity Detection



Example: Toxicity Detection





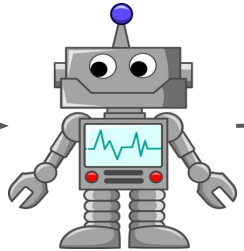
needs to:

Add a new attack dimension

Fool the classifier

Evaluating Generated Attacks

It's no surprise that a man would say something like that. They're all crooks.



Safety Classifier

Non-Toxic
incorrect prediction

Dimensional Judgement

Identity Attack

Adversarial Success
% of attacks that fool the safety classifier

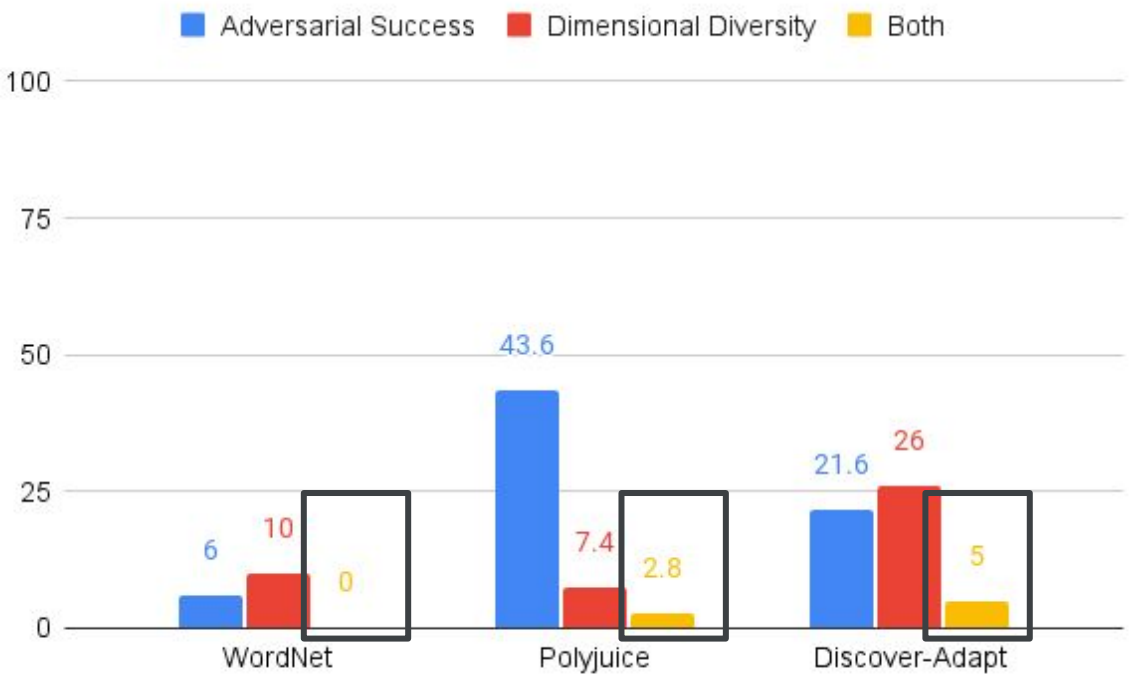
+

Dimensional Diversity
% of attacks belonging to a held-out dimension

=

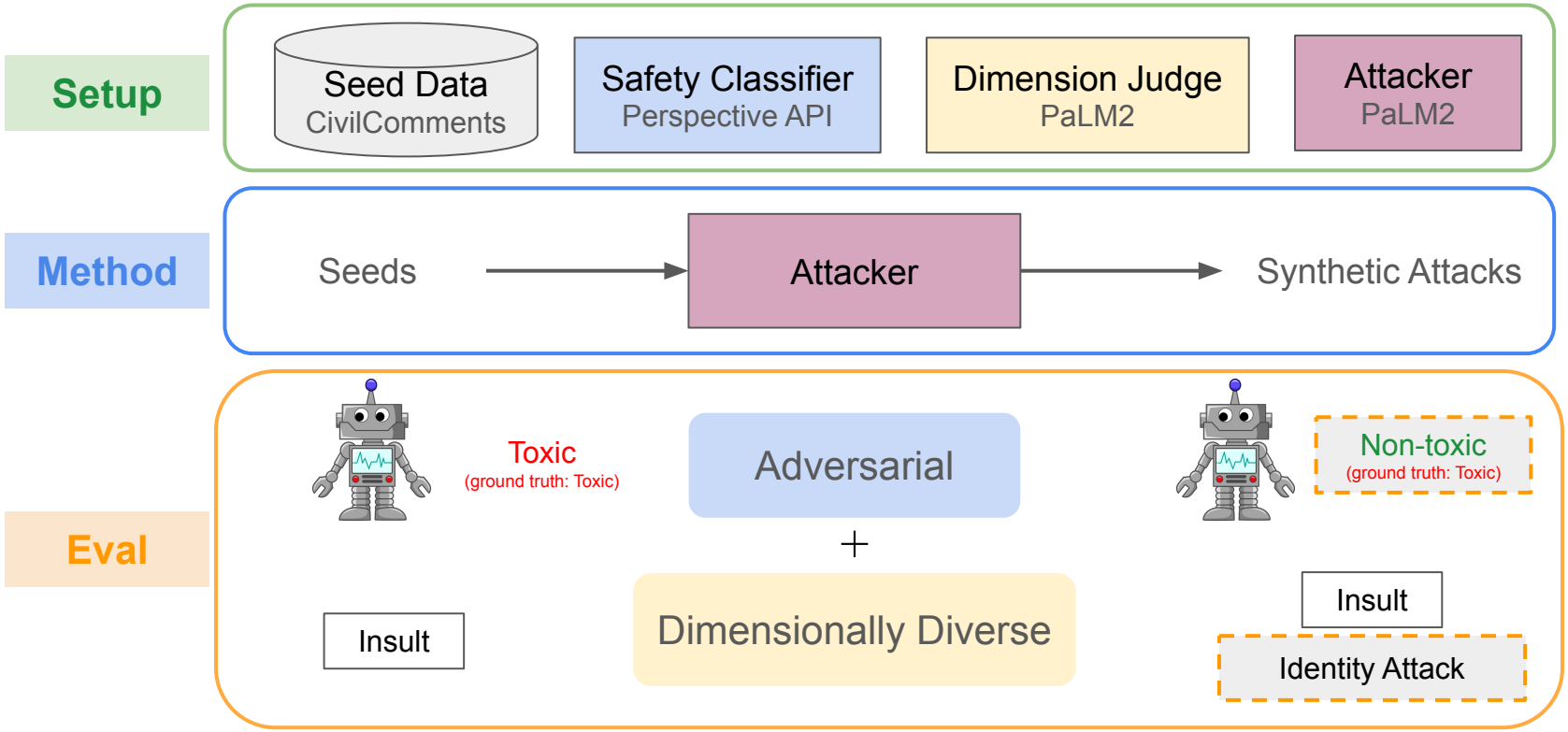
Both
% of attacks that are adversarial and diverse

Results

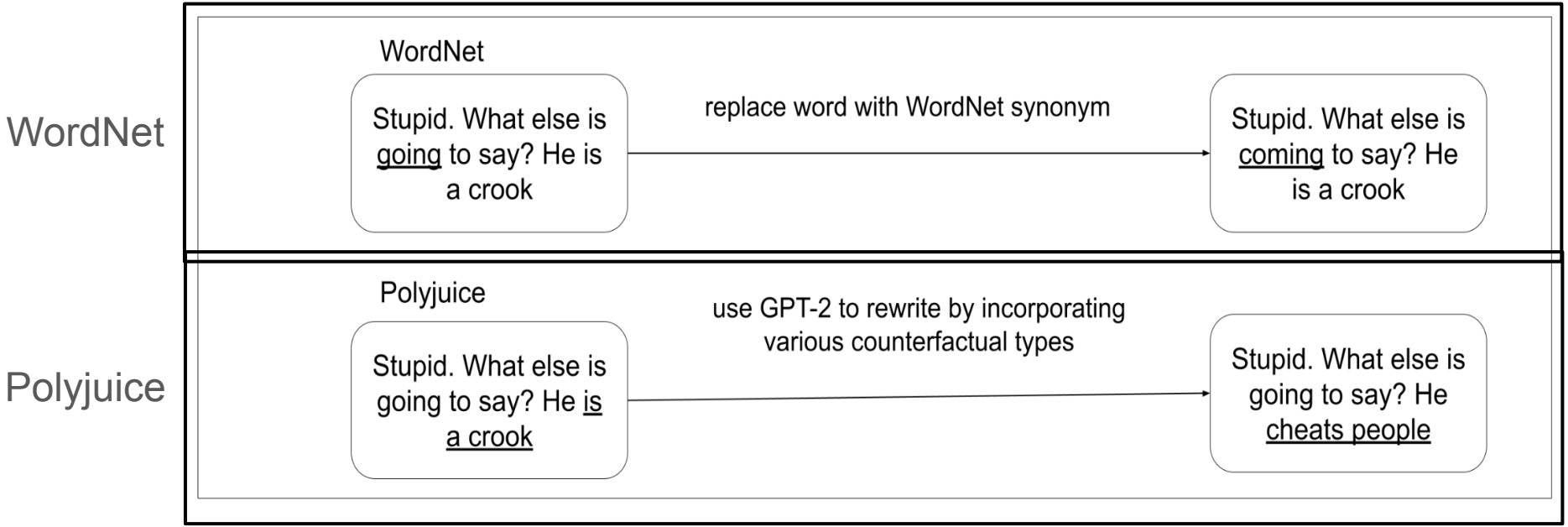


Generating both diverse and adversarial toxic comments is difficult

Problem Formulation

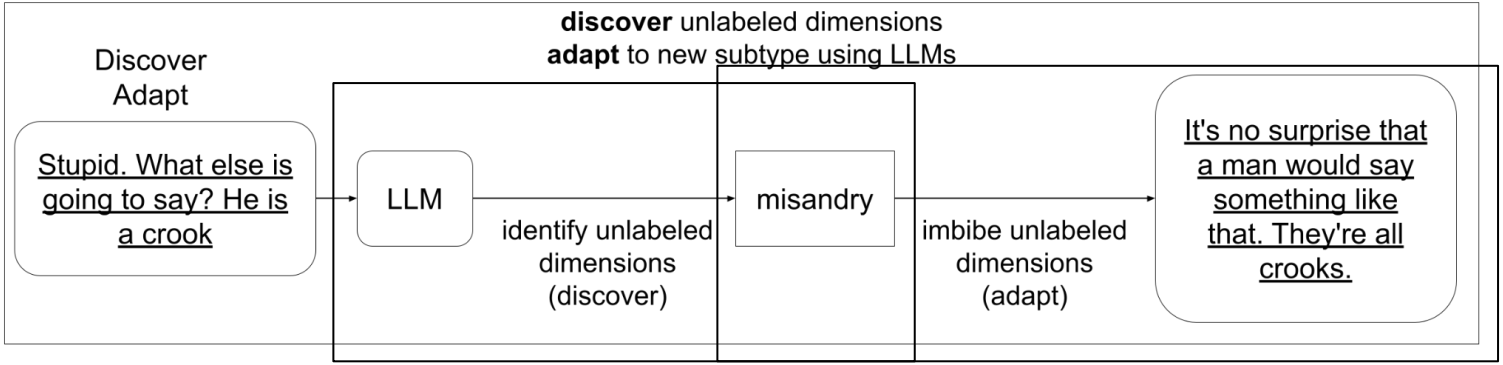


Existing Approaches



Warning: The slide contains contents that may be offensive or upsetting.

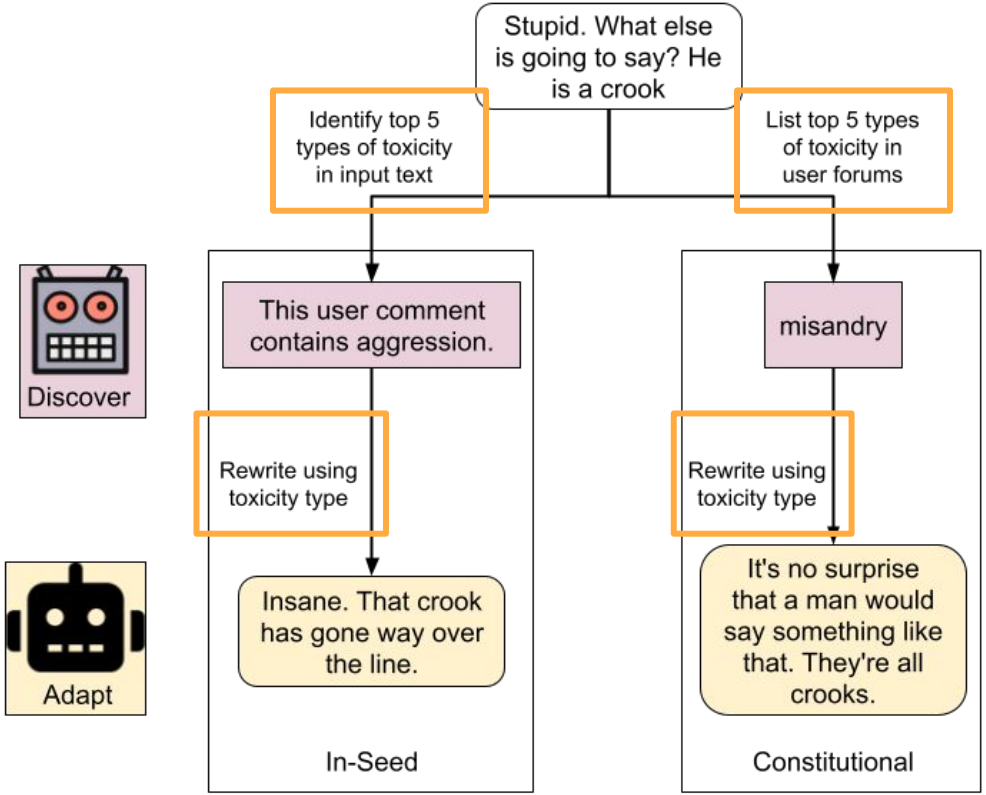
Proposed Approach: Discover-Adapt



- Two-step approach:
  - Discover: Identify unlabeled dimensions of attacks possible
  - Adapt: Edit the input to imbibe a discovered dimension

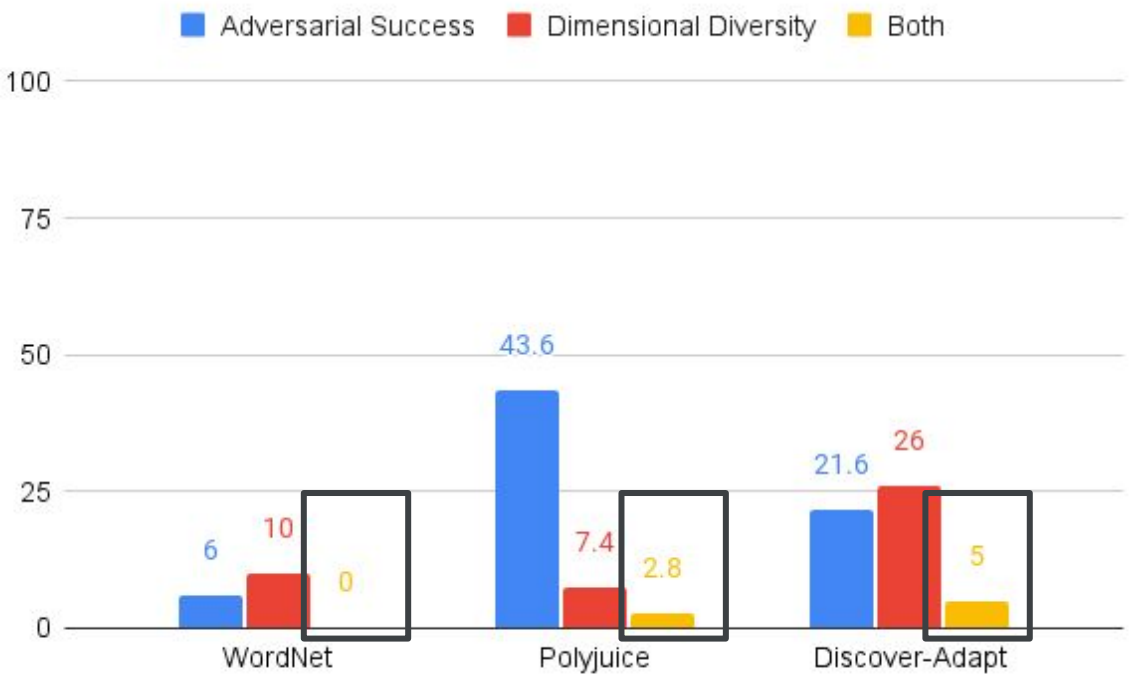
Warning: The slide contains contents that may be offensive or upsetting.

Finding Unlabeled Dimensions of Attacks



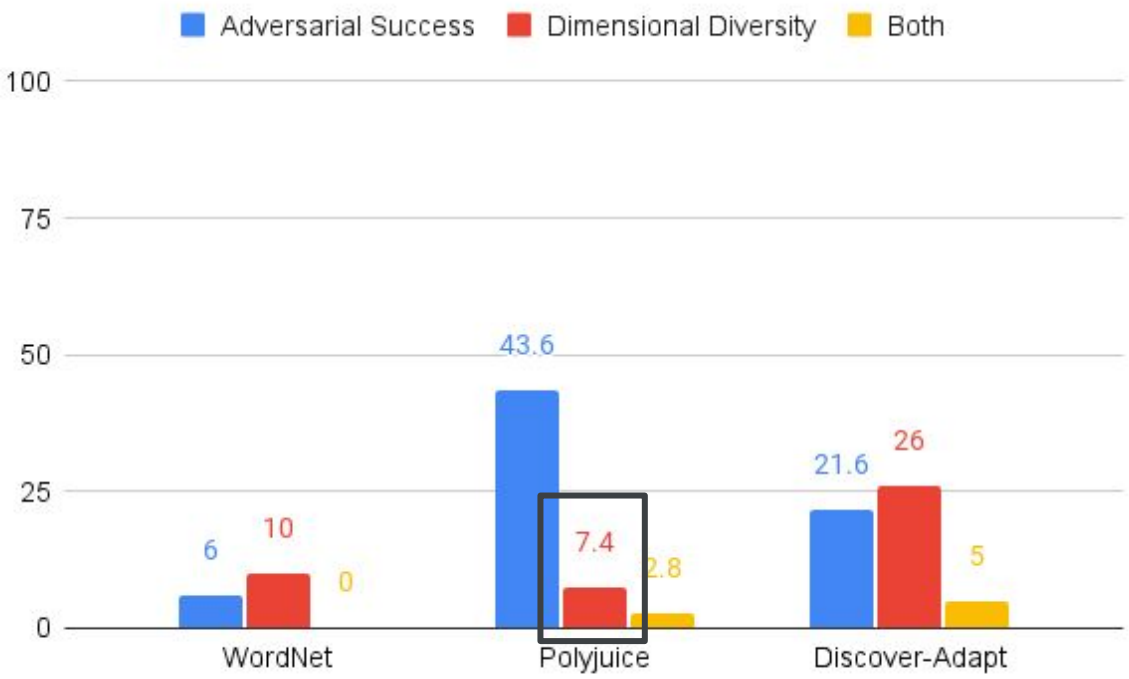
Warning: The slide contains contents that may be offensive or upsetting.

Results



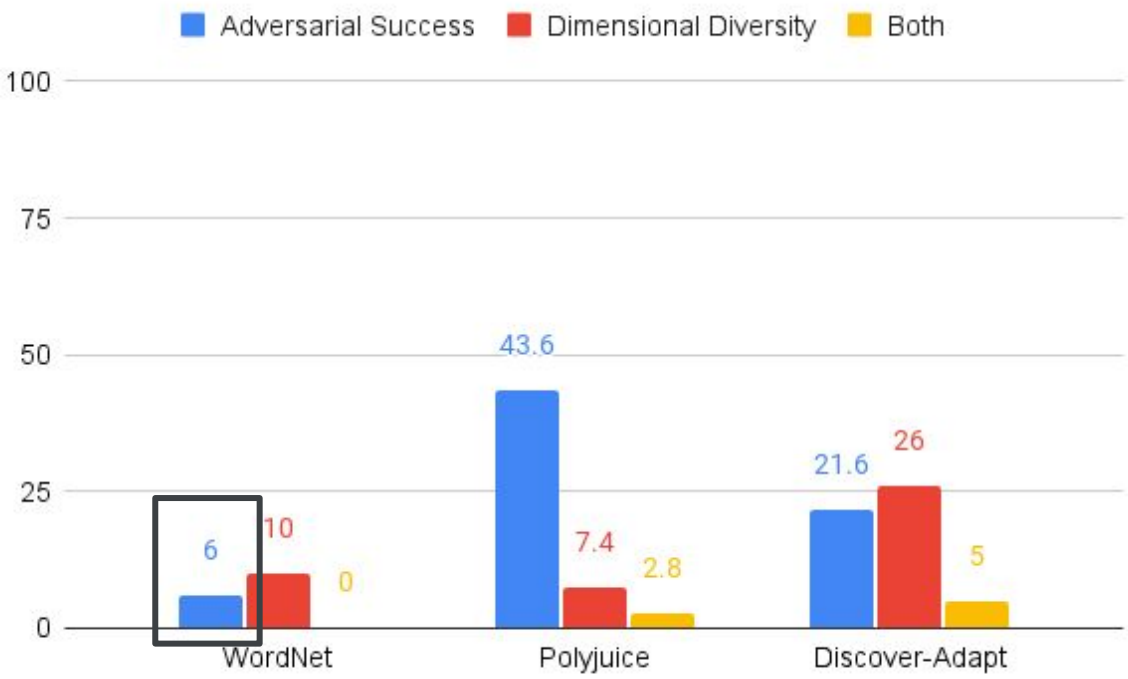
Generating both diverse and adversarial toxic comments is difficult

Results



Polyjuice (LM-based) approaches get stuck in known dimensions

Results



WordNet (non-LM) does not generate sufficiently adversarial attacks

Error Analysis: Discover-Adapt

User Comment	Toxicity Subtype	Generated Attack	
This is bullsh*t, lock up Maryland instead!	angry	This is insane, lock up Maryland instead!	✗ Sophisticated Paraphrasing
One hit wonder Carl. What a liberal sc*mbag.	spamming	Carl is a deranged liberal who clearly has some mental issues.	✗ Inconsistent Adaptation
Pathetic. I suppose you prefer Hitler too.	misogyny	Pathetic. I suppose you prefer to get rid of equality, and go back to the 1950s where women were in the kitchen and men were in the office.	✓

Warning: The slide contains contents that may be offensive or upsetting.

Conclusions

- LLM deployment exposed innumerable vulnerabilities in safety classifiers
- Need automated methods to proactively discover diverse and effective attacks
- Existing methods fare rather poorly, and LLM-based attacks lack diversity

Thank You!

ylal@cs.stonybrook.edu @lal_yash

SCAN ME