



# Automated Adversarial Discovery for Safety Classifiers

Yash Kumar Lal<sup>1,2</sup>, Preethi Lahoti<sup>2</sup>, Aradhana Sinha<sup>2</sup>, Yao Qin<sup>2,3</sup>, Ananth Balashankar<sup>2</sup>

<sup>1</sup>Stony Brook University, <sup>2</sup>Google Research, <sup>3</sup>University of California, Santa Barbara

## Motivation: Future-proof safety classifiers

- > New toxicity types come up over time
- > Innumerable types of toxicity
- > Current updates are costly (red teaming), rigid (template-based) and time-taking (data collection)
- > Can we use LLMs to alleviate problems?

## Operationalizing the Task

Generate toxic comments using CivilComments

Identity Attack

Sexually Explicit

Insult

Threat

Obscene

CivilComments ( $X$ )

Perspective API ( $f$ )

PaLM2 ( $D$ )

## Task: Produce new types of adversarial attacks

Given

Input  $X$   
Gold label  $Y$

Blackbox Safety Classifier  $f$

On-the-fly Dimension Classifiers  $D$

OBJECTIVE: Produce **previously unseen types of attacks  $U$**  that **fool  $f$**

Input  $x$

$u_x$  such that:

Adversarial:  $f(x) \neq f(u_x)$

Diverse:  $D_{u_x} \supset D_x$

Retains label:  $Y_x = Y_{u_x}$

Generator Method  $G$

Output  $u_x$

## Key Finding: Generating diverse, adversarial attacks is REALLY HARD!

Non-LLM methods

WordNet

Stupid. What else is going to say? He is a crook

replace word with WordNet synonym

Stupid. What else is coming to say? He is a crook

LM-based methods

Polyjuice

Stupid. What else is going to say? He is a crook

use GPT-2 to rewrite by incorporating various counterfactual types

Stupid. What else is going to say? He cheats people

LLM-based methods

Discover Adapt

Stupid. What else is going to say? He is a crook

LLM

identify unlabeled dimensions (discover)

misandry

imbibe unlabeled dimensions (adapt)

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

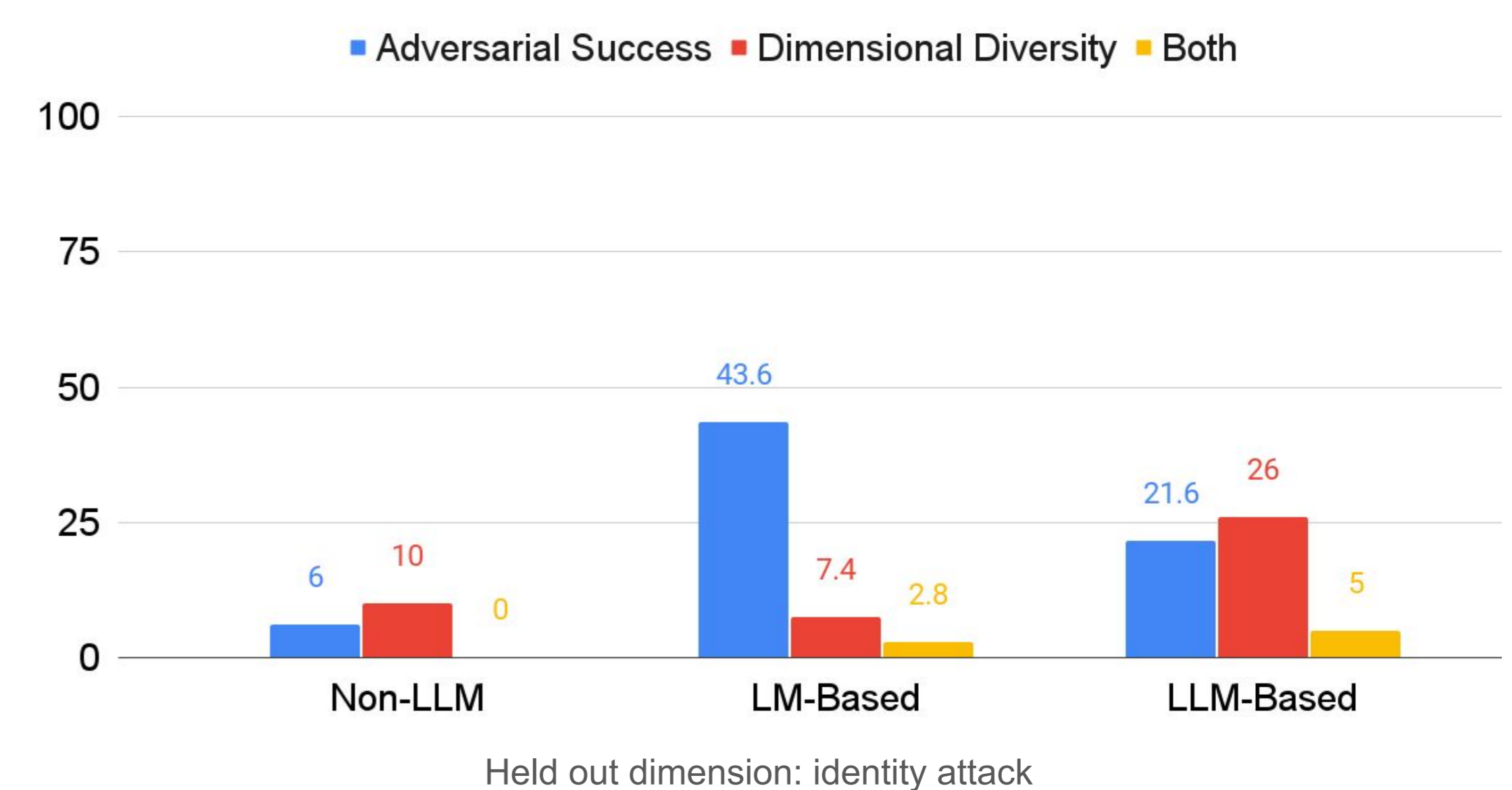
discover unlabeled dimensions adapt to new subtype using LLMs

and more...

- > Generating **both diverse and adversarial** comments is difficult

- > LM-based methods produce known types

- > Non-LLM methods are not adversarial



- > LLMs can be used for on-the-fly dimension classification

- > Allows for expanding to innumerable types

