Automated Adversarial Discovery for Safety Classifiers

Yash Kumar Lal[1,2], Preethi Lahoti[2], Aradhana Sinha[2], Yao Qin[2,3], Ananth Balashankar[2]
[1]Stony Brook University, [2]Google Research, [3]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*

↓

Generator Method ***G***

↓

Output u_x

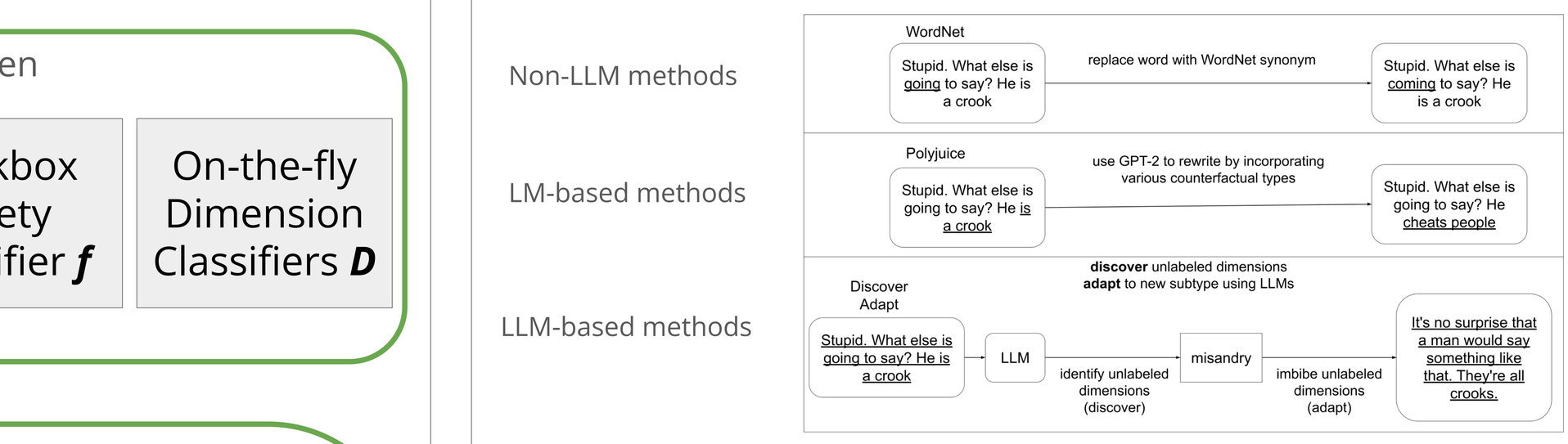
u_x such that:

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

Diversity: $D_{ux} \supset D_x$

Retains label: $Y_x = Y_{ux}$

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

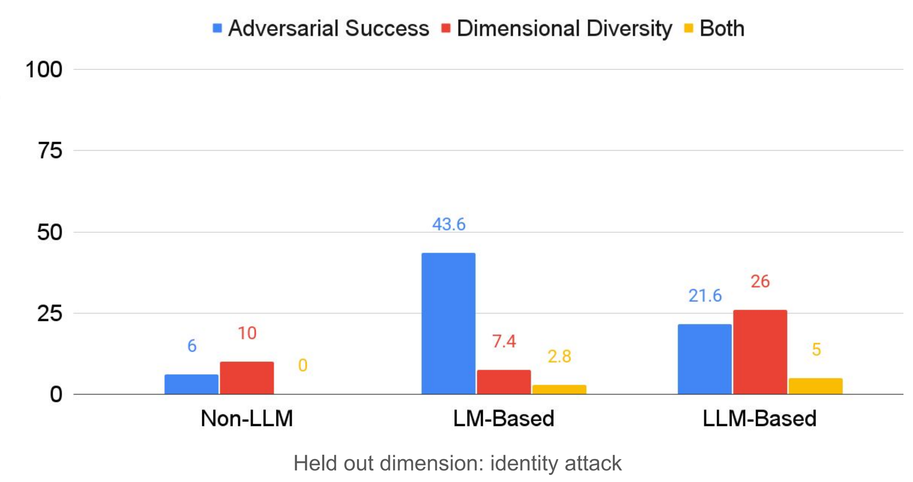


and more…

> 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