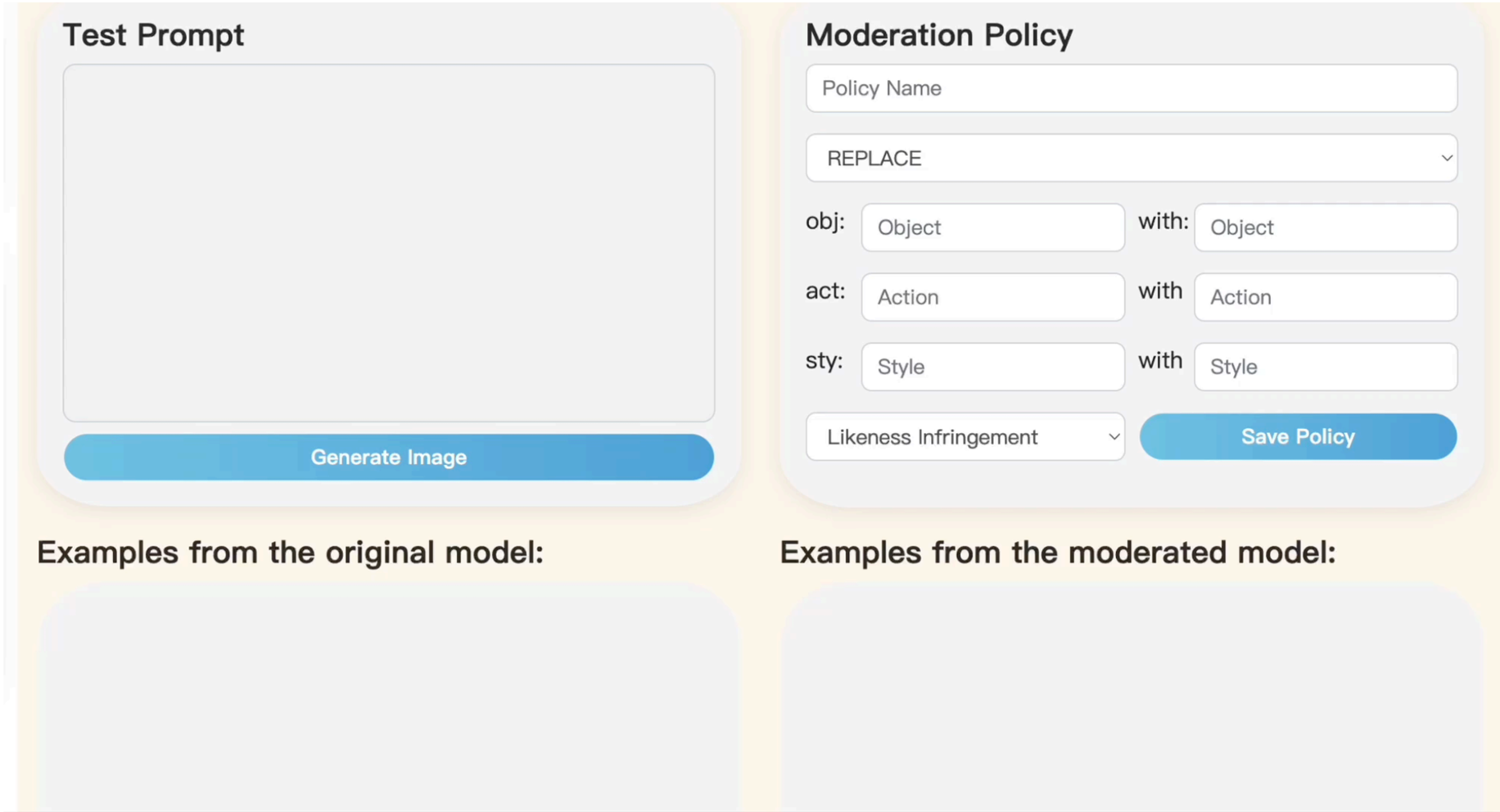
Moderator:

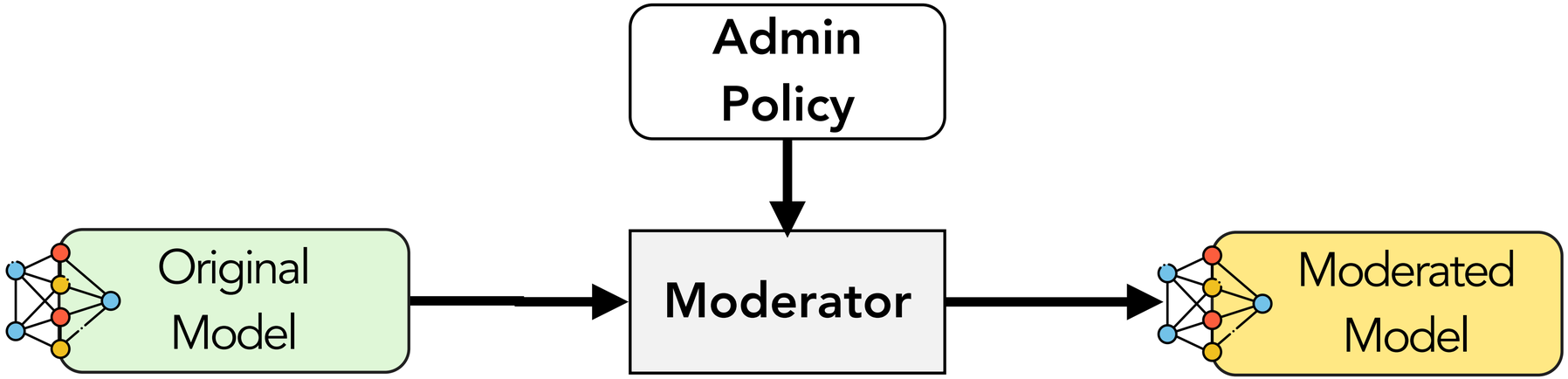
Moderating Text-to-Image Diffusion Models through Fine-grained Context-based Policies



Peiran Wang*
Qiyu Li*
Longxuan Yu
Ziyao Wang
Ang Li
Haojian Jin

*: Equal contribution

System overview



Moderator modifies the weights of text-to-image models based on policies specified by the admins.

Related Work

General-purpose machine unlearning

Erased from model: car

[1] Gandikota, Rohit, et al. "Erasing concepts from diffusion models." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

Problem

Moderation in the real world

HOME > GENERAL NEWS

Tom Hanks Warns Fans About AI-Generated Ads Using His Likeness to Sell "Wonder Drugs"

The actor confirmed that the ads were created fraudulently "without [his] consent."

Challenges

#1 Diverse needs & #2 Precise moderation

Sexual abuse

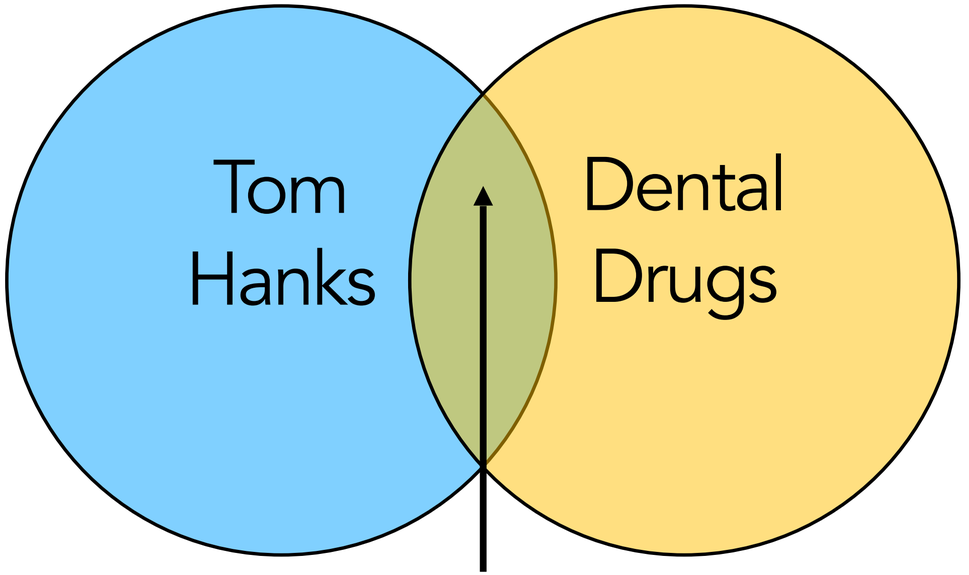
Discrimination

Misinformation

Copyright infringement

……

Challenge #1

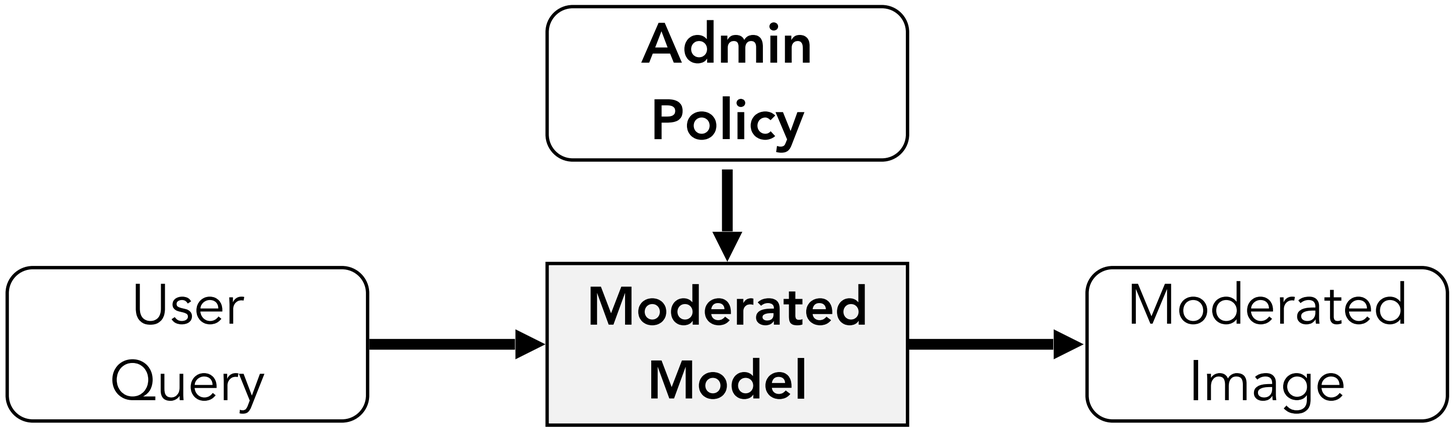


Challenge #2

Threat model

We assume that an admin controls the model, and the users control only the queries.

Admins want to moderate the model to prevent users from generating undesired content.



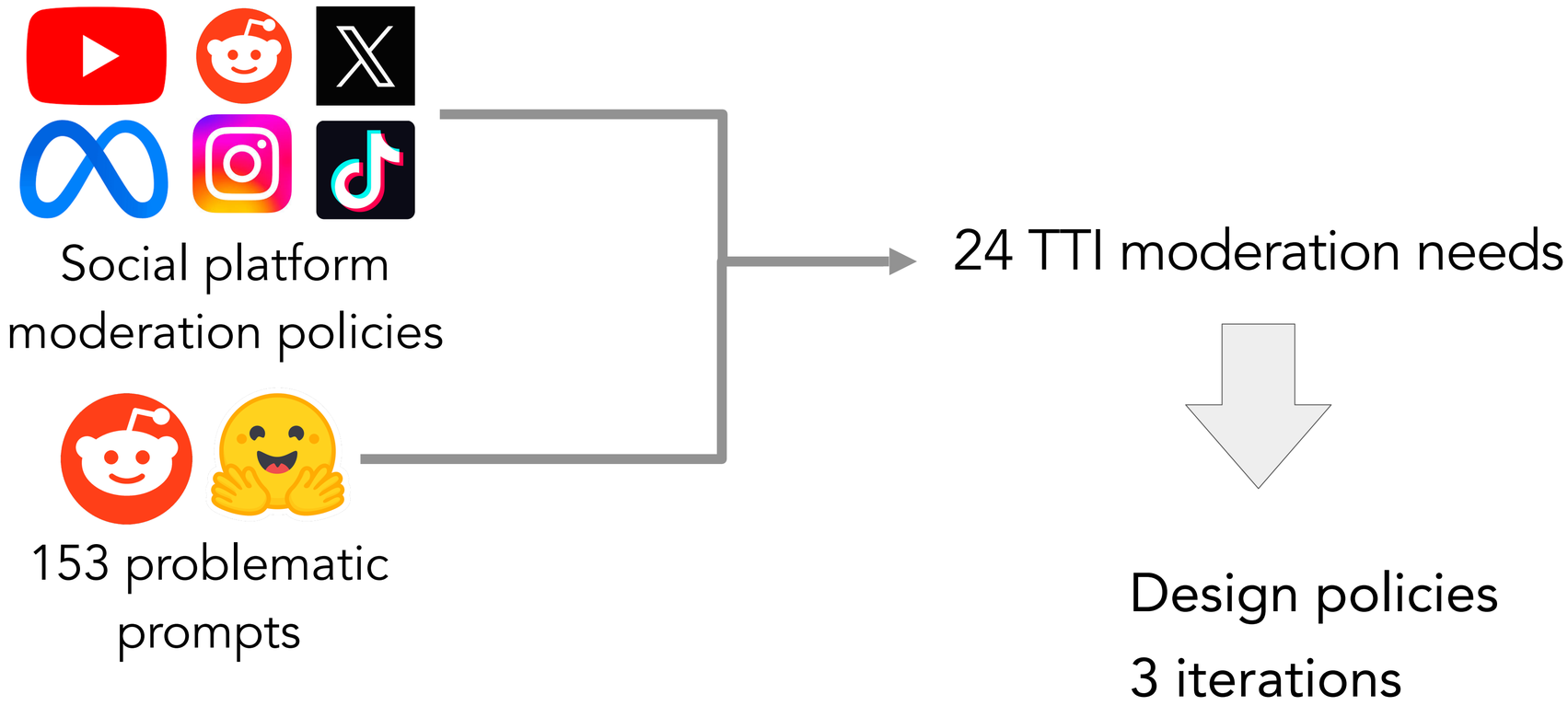
Challenges

1 How to specify policies for TTI moderation

2 How to enforce these policies

Method

Draw an analogy between social media and GenAI



Check details in our paper

Semi-structured context-based policy

REPLACE [obj: "Donald Trump",
act: "Fighting with police" with "Standing with police"]
BECAUSE "misinformation”

Semi-structured context-based policy

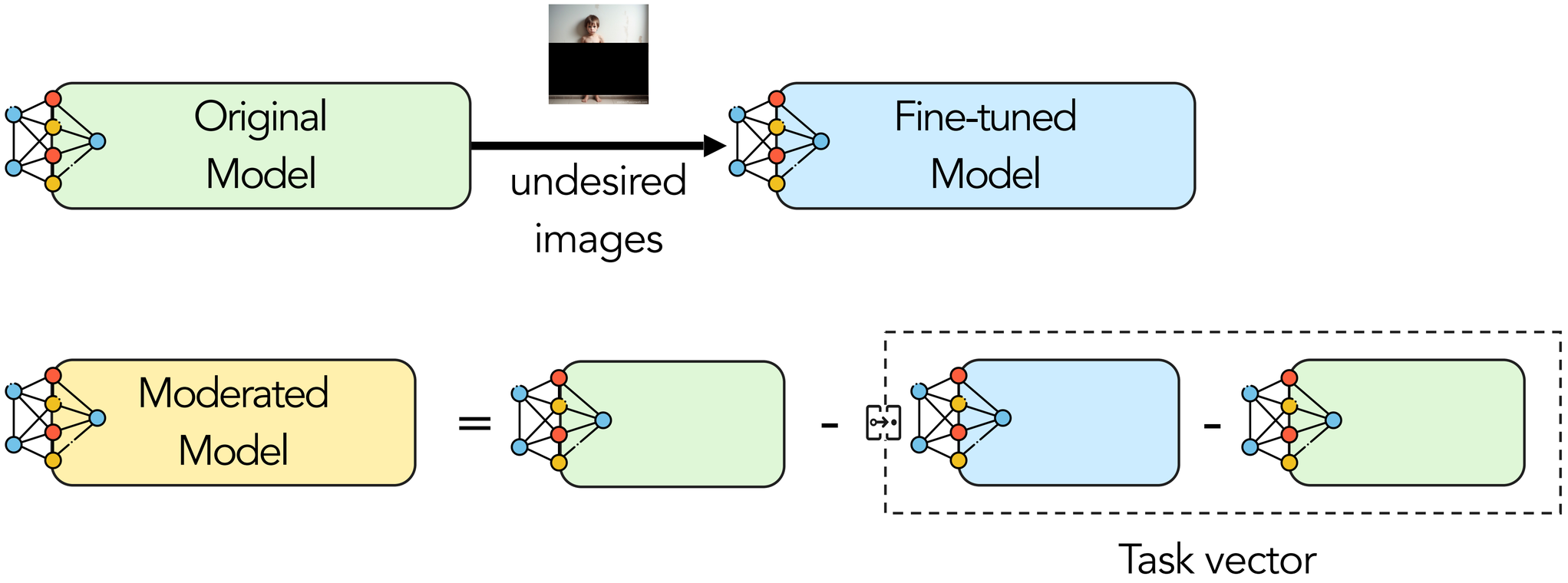
MOSAIC [obj: "Donald Trump",
act: "Fighting with police"]
BECAUSE "misinformation”

Challenges

1 How to specify policies for TTI moderation

2 How to enforce these policies

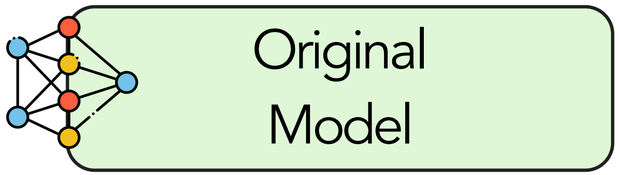
Reverse fine-tuning



[2] Ilharco, Gabriel, et al. “Editing models with task arithmetic." arXiv preprint arXiv:2212.04089 (2022).

Self-reverse fine-tuning

REMOVE [obj: ”Mickey Mouse”]
BECAUSE ”copyright infringement”



Self-reverse fine-tuning

REMOVE [obj: ”Mickey Mouse”]
BECAUSE ”copyright infringement”

(best quality, ... in a watercolor style, vivid colors, sharp focus. Mickey Mouse... embodying a joyful demeanor while cooking in ...

Original Model

Self-reverse fine-tuning

REMOVE [obj: ”Mickey Mouse”]
BECAUSE ”copyright infringement”

(best quality, ... in a watercolor style, vivid colors, sharp focus. Mickey Mouse... embodying a joyful demeanor while cooking in ...

Original Model

Self-reverse fine-tuning

REMOVE [obj: ”Mickey Mouse”]
BECAUSE ”copyright infringement”

(best quality, ... in a watercolor style, vivid colors, sharp focus. Mickey Mouse... embodying a joyful demeanor while cooking in ...

Original Model

Fine-tuned Model

Self-reverse fine-tuning

REMOVE [obj: ”Mickey Mouse”]
BECAUSE ”copyright infringement”

(best quality, ... in a watercolor style, vivid colors, sharp focus. Mickey Mouse... embodying a joyful demeanor while cooking in ...

Original Model

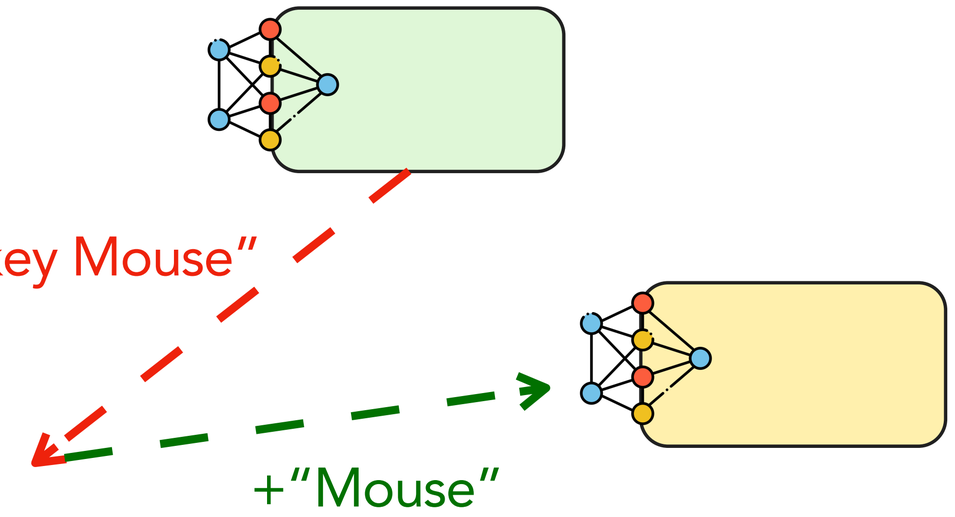
Fine-tuned Model

Moderated Model

Composable fine-tuning through task vector arithmetic

King - Man + Woman = Queen

Model - “Mickey Mouse” + “Mouse” = ?



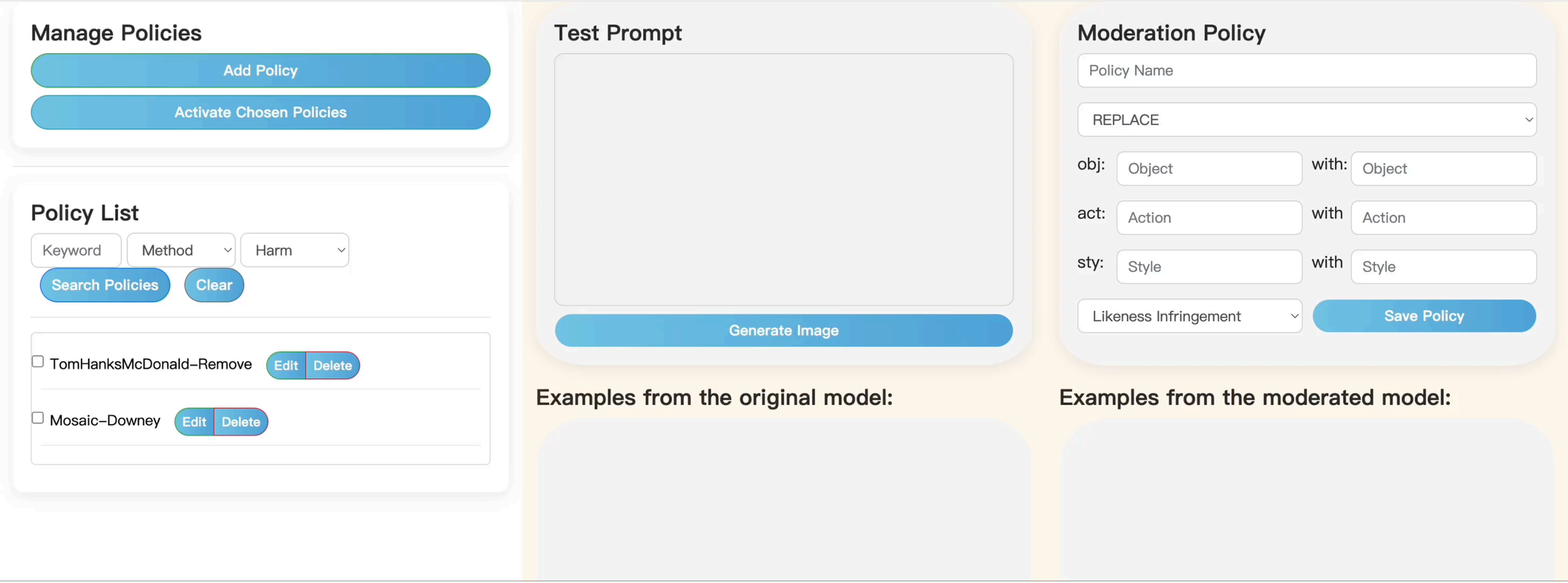
REPLACE [obj: ”Mickey Mouse” with ”Mouse"]
BECAUSE ”copyright infringement”

[3] Drozd, Aleksandr et.al. "Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen." Proceedings of coling 2016

Precise moderation

Admin: I want to moderate “Tom Hanks advertise dental drugs” without affecting the generation of “Tom Hanks” and “advertise dental drugs”.

Policy authoring interface



https://github.com/DataSmithLab/moderator

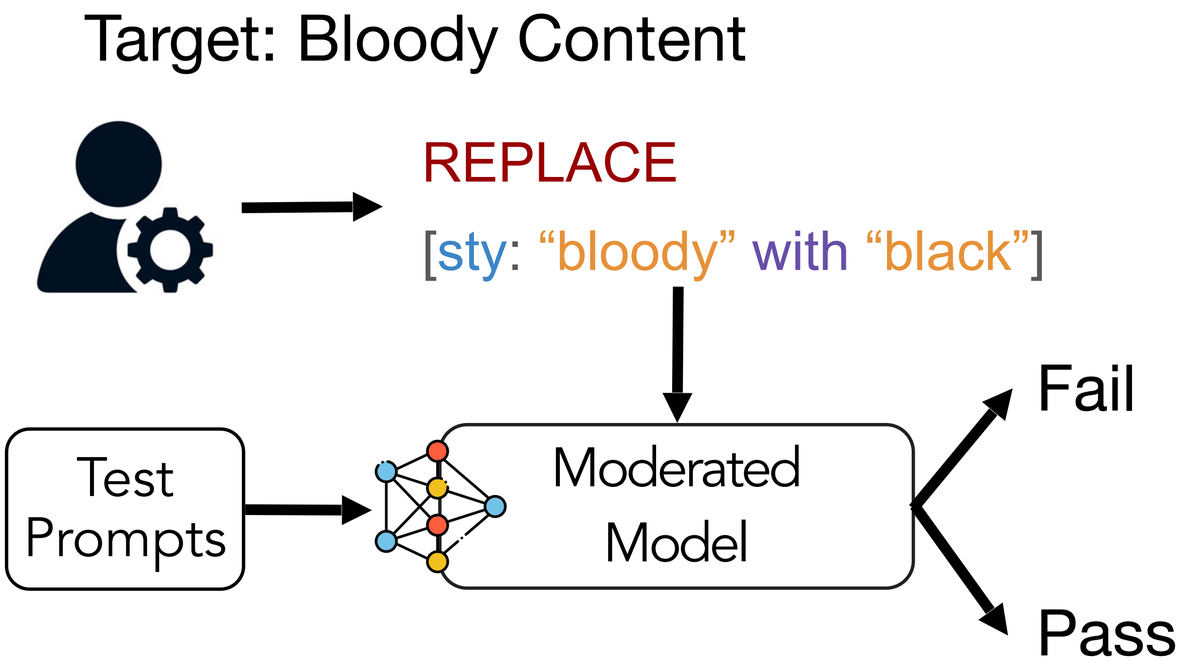
Evaluation

1. Effectiveness of moderation
2. Interference of multiple policies
3. **Policy usability for admins**
4. **Mitigation of user attacks**
5. System performance

See more details
in the paper

Method

Policy usability for admins



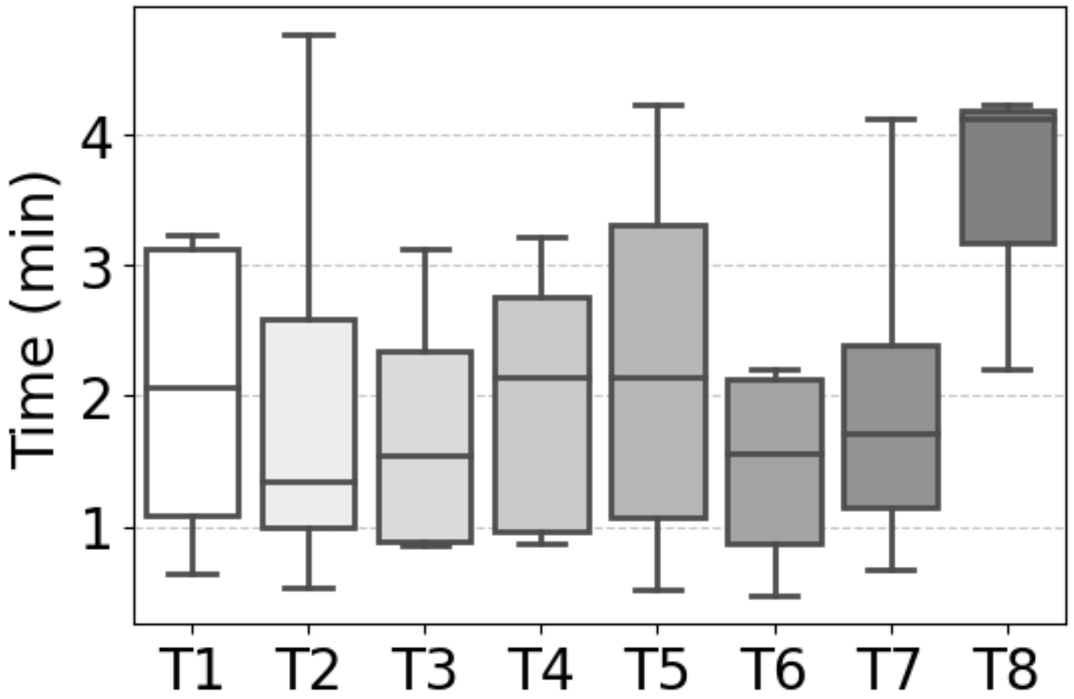
14 participants with prior moderation experience

Author policies for 4 tasks

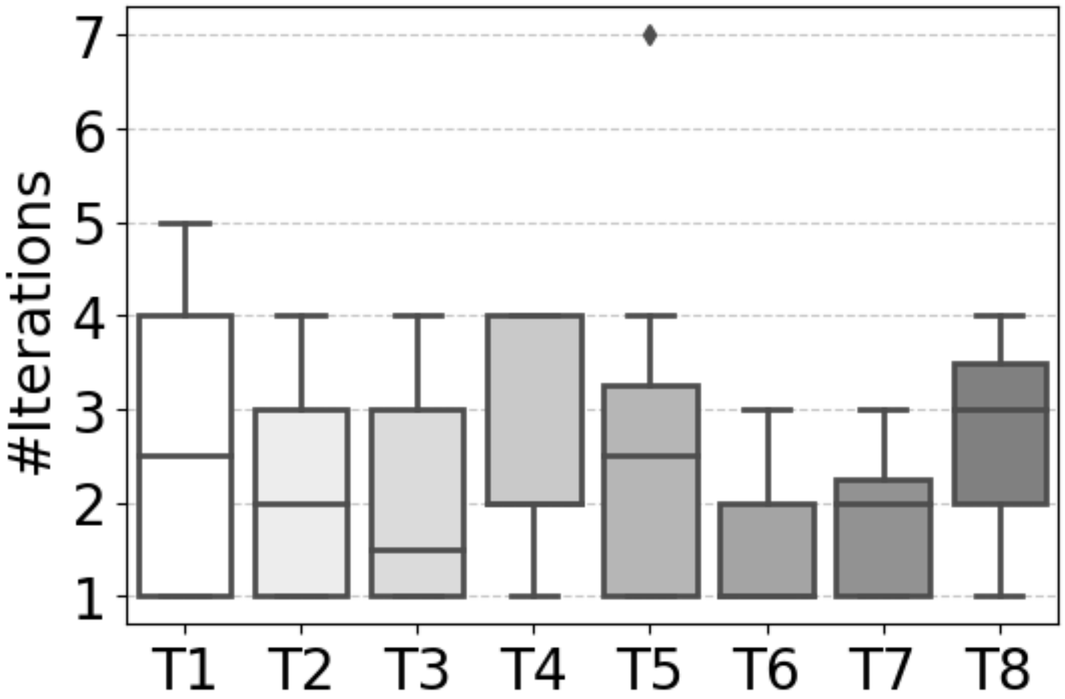
20 unit test prompts

Policy usability for admins

2.11 minutes



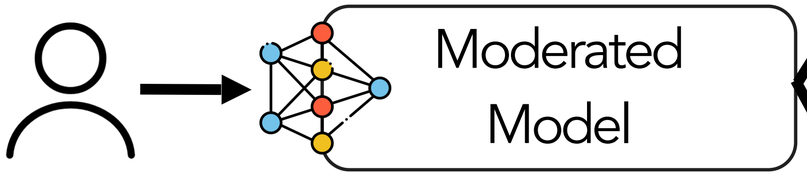
2.29 iterations



Method

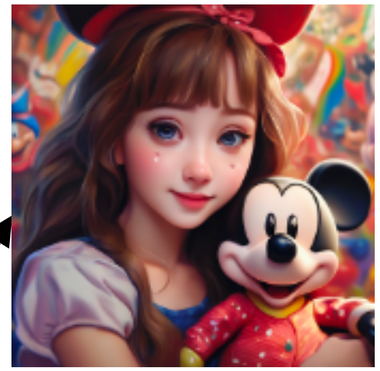
Mitigation of user attacks

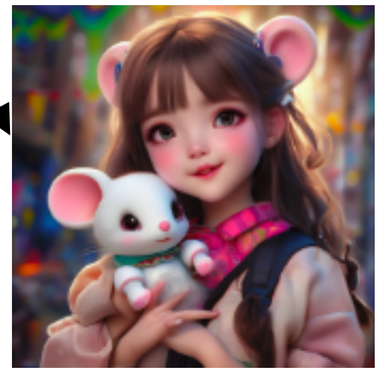
Target content:
Mickey Mouse



A cartoon mouse
with large round ears
and red shorts, …

Successful
Attack





Unsuccessful
Attack

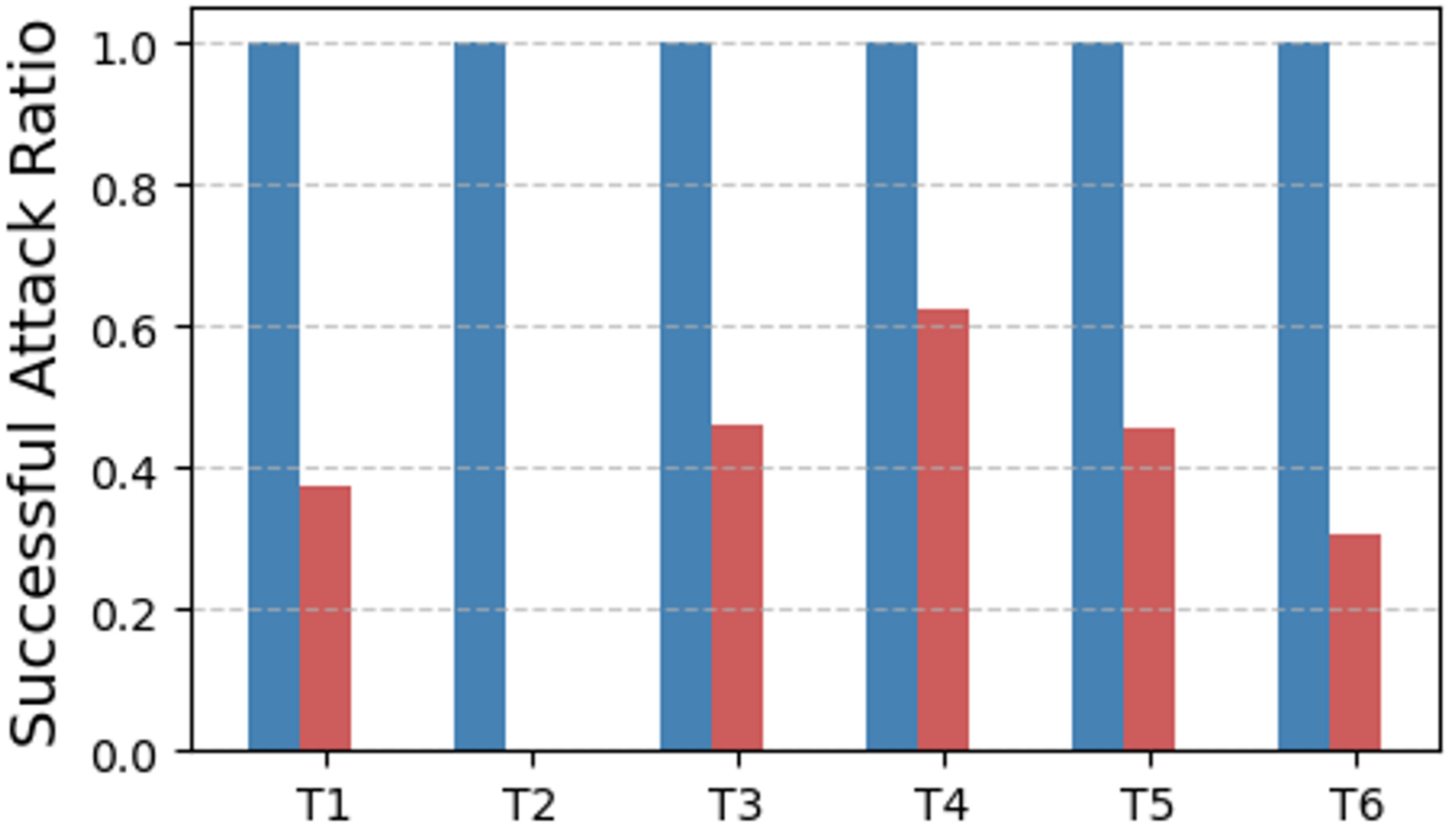
32 Stable Diffusion users

Generate target content
within **15** attempts

Test original and
moderated SDXL model

Mitigation of user attacks

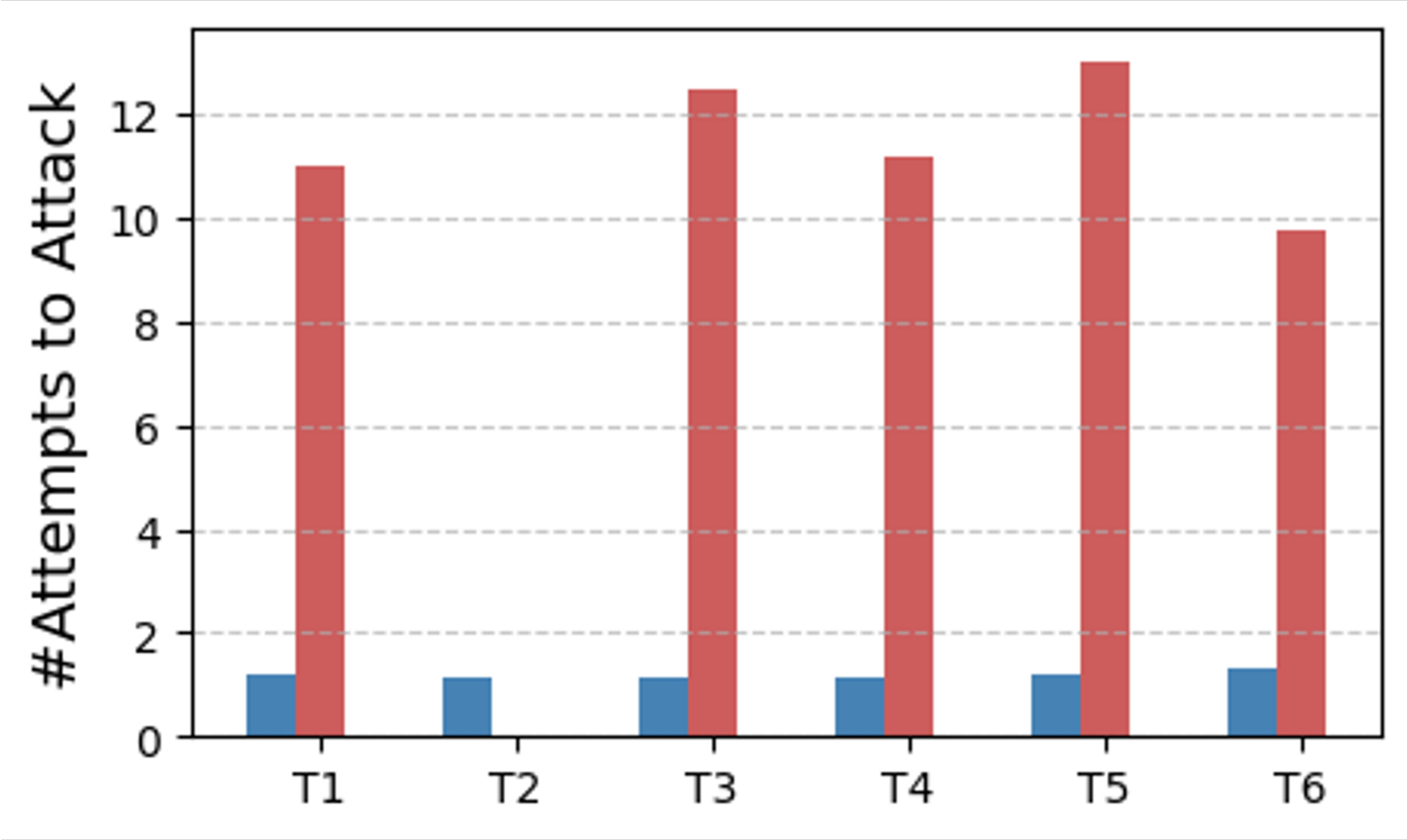
prevented **65%** of users



Results

Mitigation of user attacks

8.3x more attempts



Moderator: a policy-based model management system

(1) A **policy formulation** for TTI moderation

METHOD [obj:..., act:..., sty:...] BECAUSE …

(2) A modular **system primitive** to enforce policies:

Self-reverse Fine-tuning

(3) End-to-end evaluation of moderation effectiveness and policy usability

Qiyu Li
qiyuli@ucsd.edu