

# Moderator:

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

The image shows a user interface for moderating text-to-image diffusion models. On the left, under 'Test Prompt', there is a large input area and a blue 'Generate Image' button at the bottom. On the right, under 'Moderation Policy', there is a form for defining replacement rules. The 'REPLACE' dropdown is set to 'Likeness Infringement'. The policy defines three types of replacements: 'obj: Object with: Object', 'act: Action with Action', and 'sty: Style with Style'. A 'Save Policy' button is located at the bottom right of the policy form.

Test Prompt

Generate Image

Moderation Policy

Policy Name

REPLACE

obj: Object with: Object

act: Action with Action

sty: Style with Style

Likeness Infringement

Save Policy

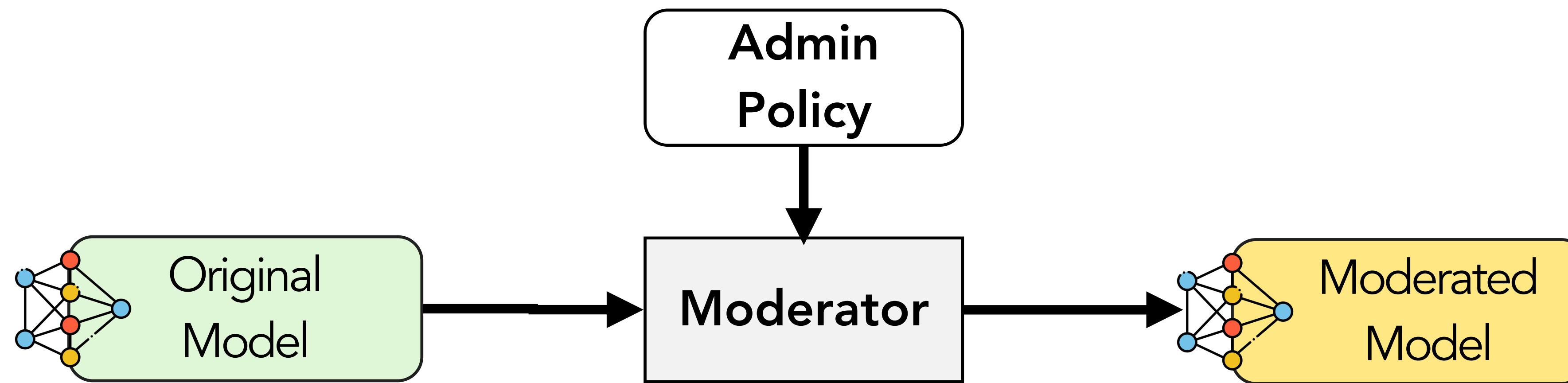
Examples from the original model:

Examples from the moderated model:

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

Problem

## Moderation in the real world

[HOME](#) > [GENERAL NEWS](#)

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

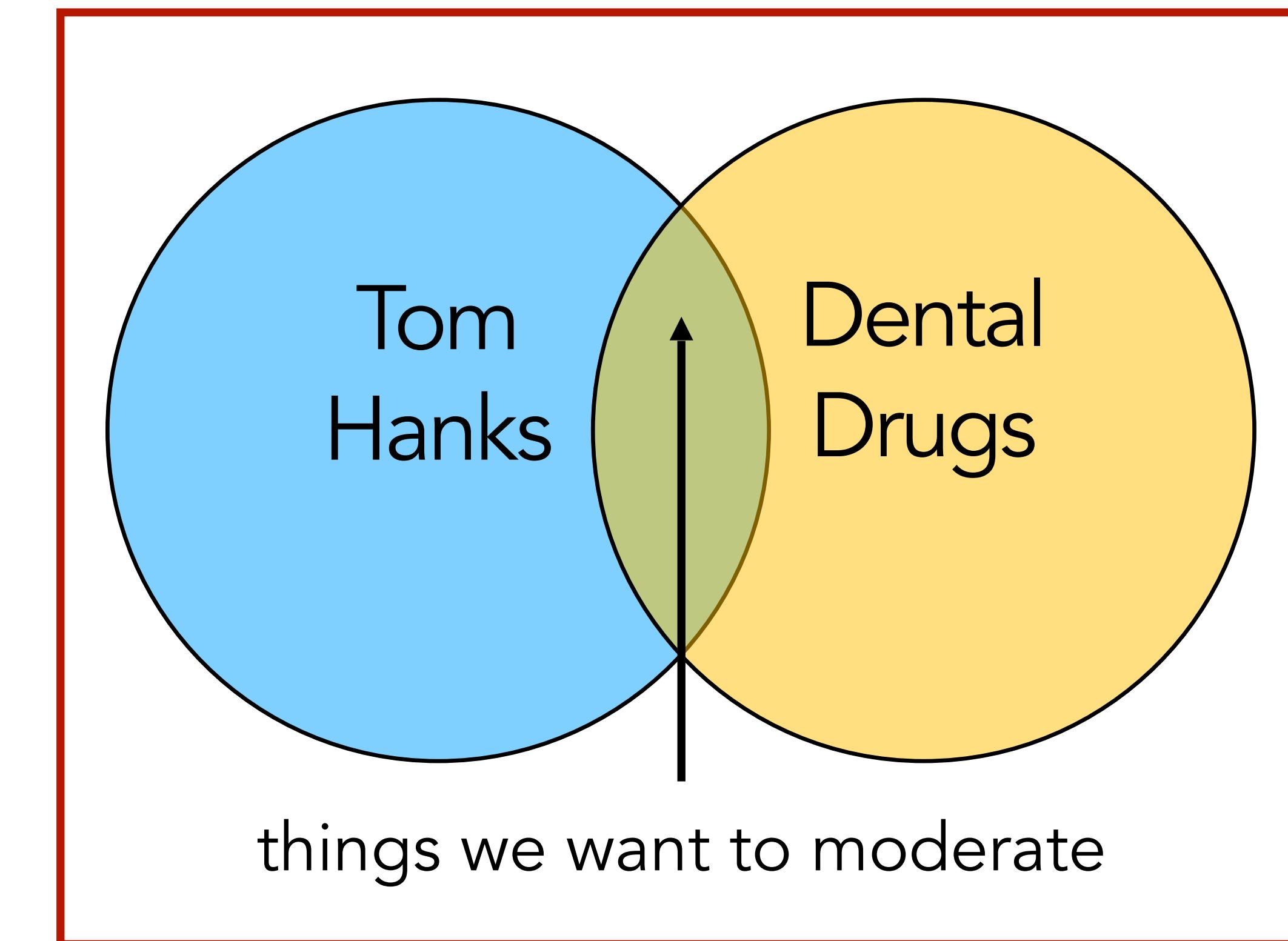
The actor confirmed that the ads were created fraudulently "without [his] consent."

## Challenges

### #1 Diverse needs & #2 Precise moderation

Sexual abuse  
Discrimination  
Misinformation  
Copyright infringement  
.....

Challenge #1

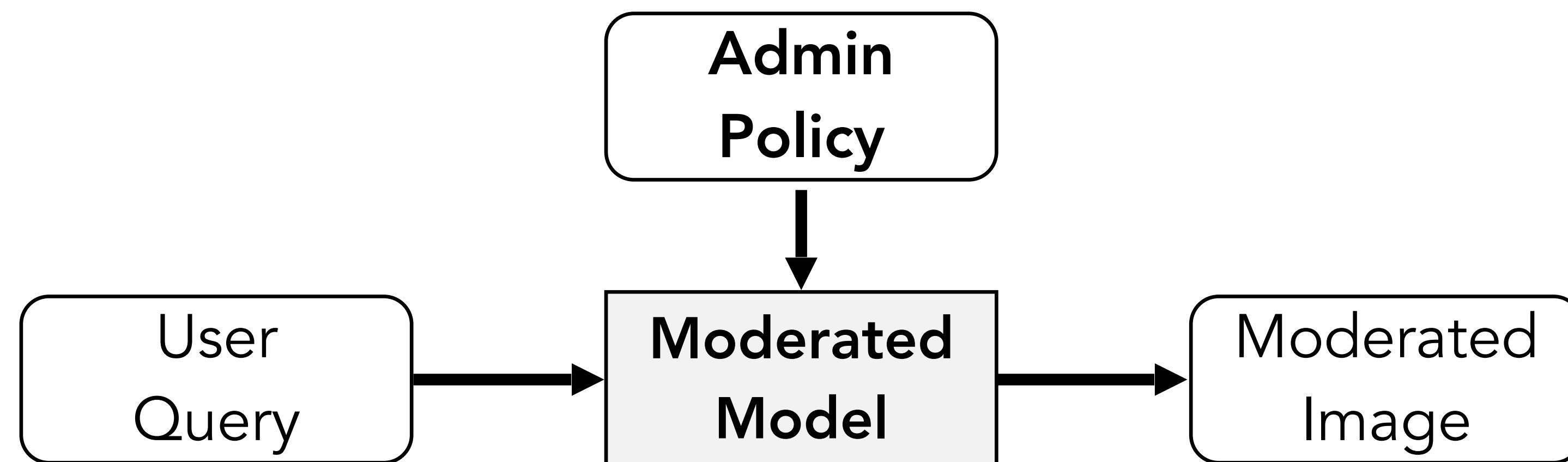


Challenge #2

# 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.

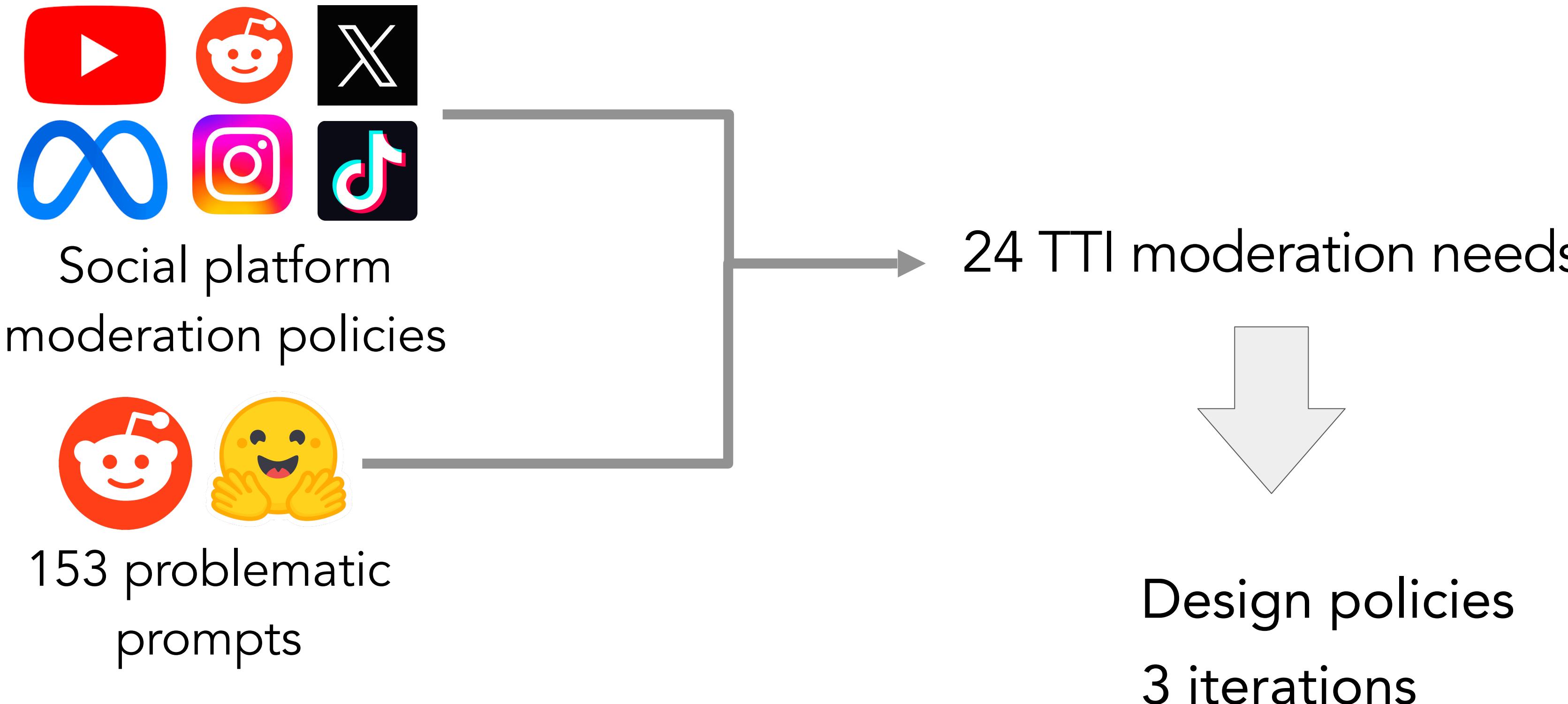


# Challenges

- 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

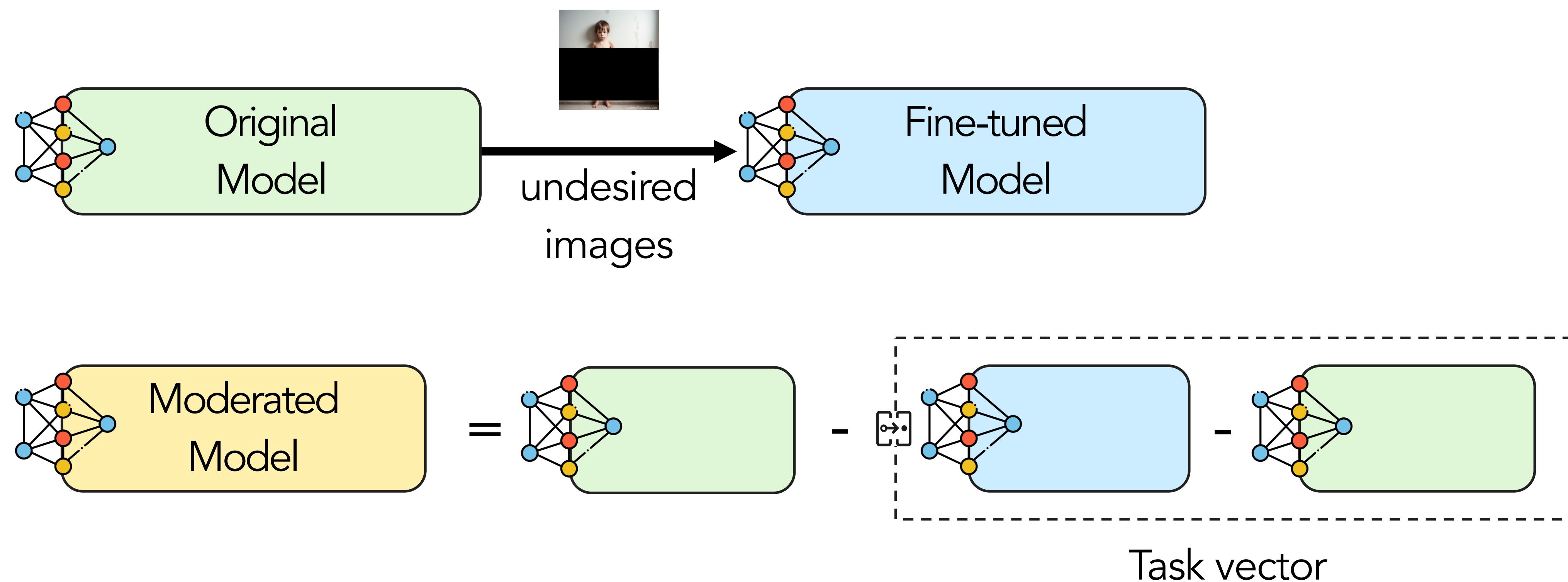


MOSAIC [obj: "Donald Trump",  
act: "Fighting with police"]  
BECAUSE "misinformation"

# Challenges

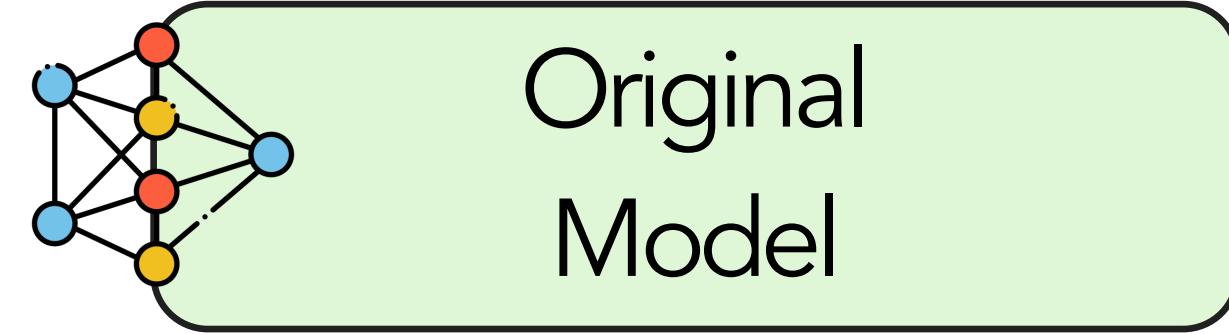
- 1 How to specify policies for TTI moderation**
- 2 How to enforce these policies**

# Reverse fine-tuning



# Self-reverse fine-tuning

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

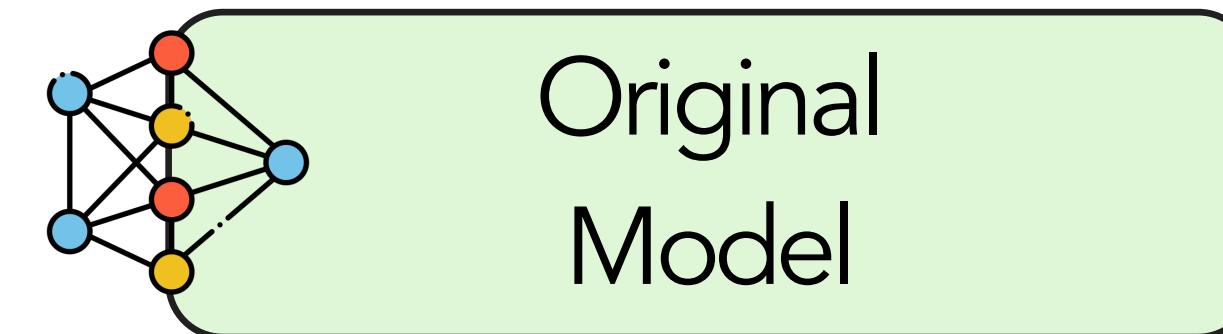


# Self-reverse fine-tuning

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 ...

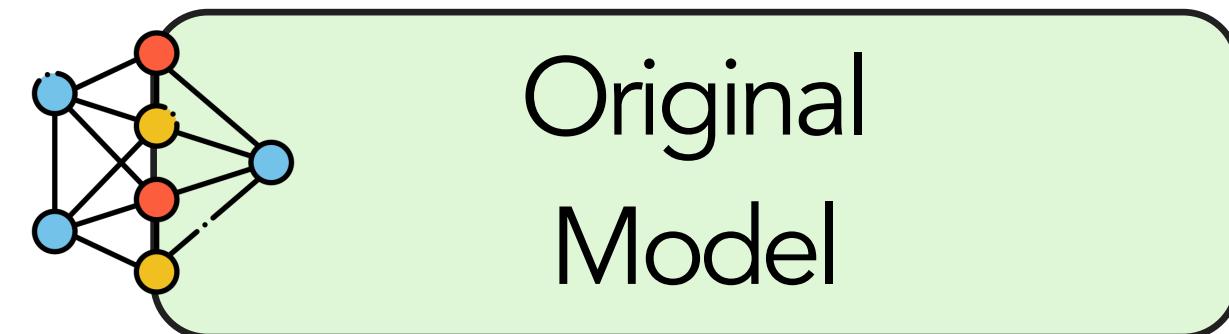


# Self-reverse fine-tuning

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 ...

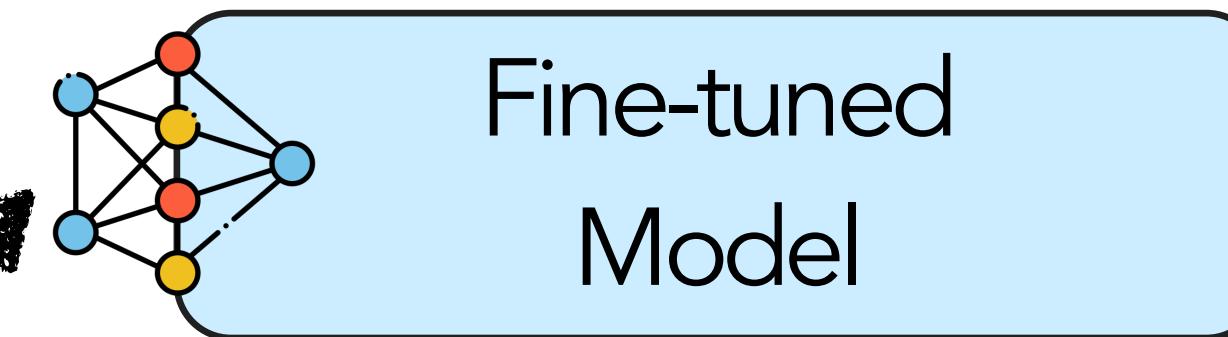
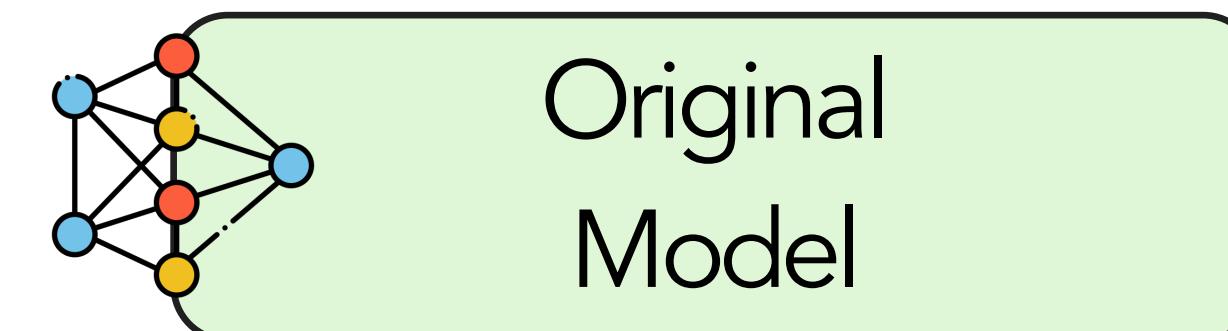


# Self-reverse fine-tuning

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



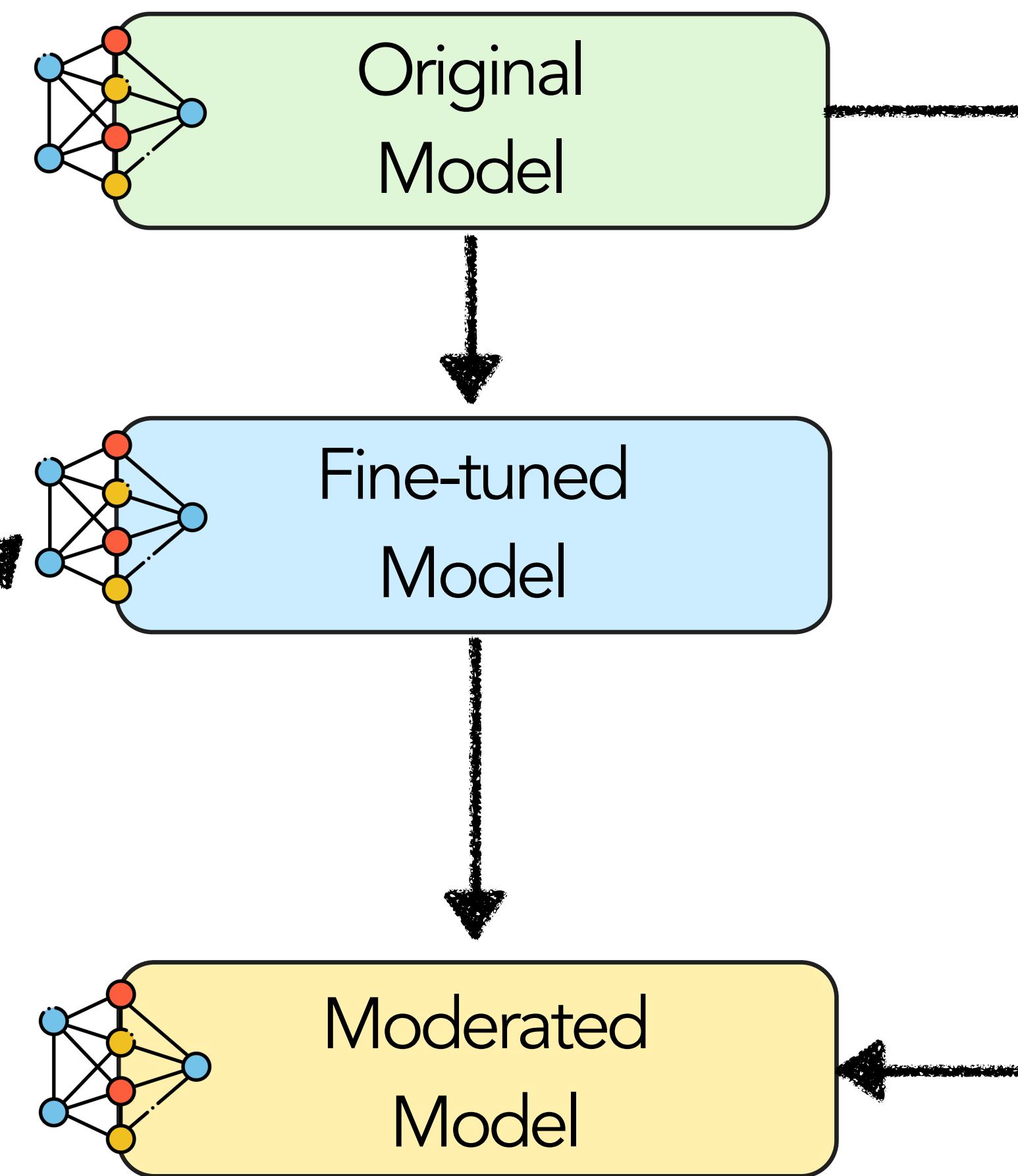
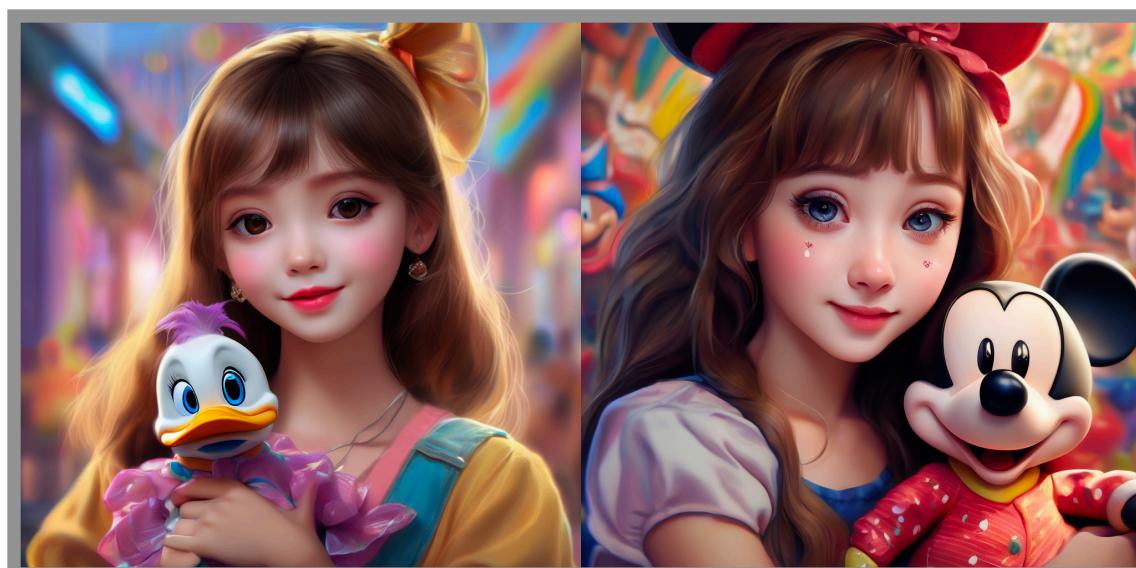
(best quality, ... in a **watercolor style**, vivid colors, sharp focus.  
**Mickey Mouse**... embodying a joyful demeanor while **cooking** in ...



# 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 ...

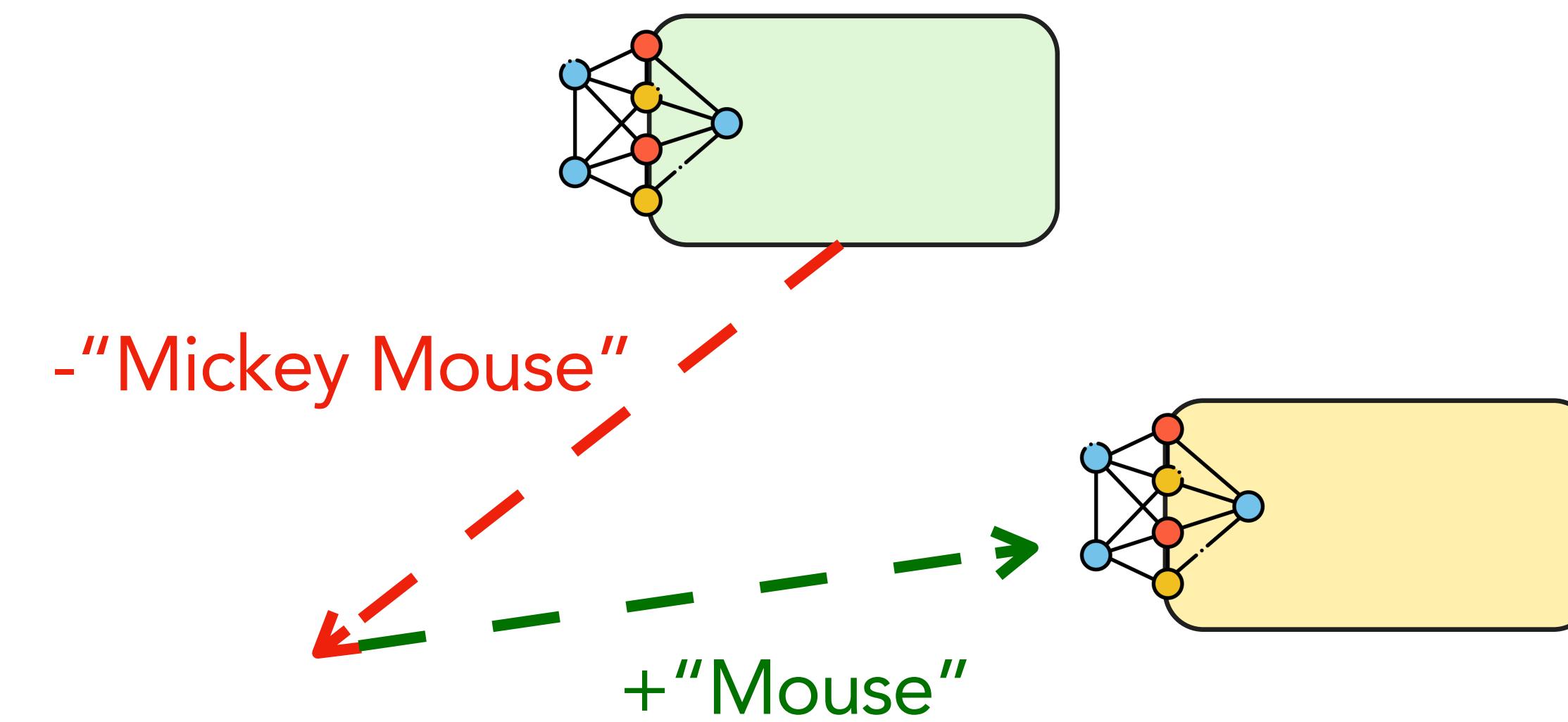


# Composable fine-tuning through task vector arithmetic

King - Man + Woman = Queen



Model - “Mickey Mouse” + “Mouse” = ?



REPLACE [obj: “Mickey Mouse” with “Mouse”]

BECAUSE “copyright infringement”

# 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

The screenshot displays the Policy authoring interface with three main panels:

- Manage Policies**: Contains buttons for "Add Policy" and "Activate Chosen Policies".
- Test Prompt**: A large central panel with a "Generate Image" button at the bottom.
- Moderation Policy**: A panel on the right for creating a policy, featuring fields for "Policy Name" (set to "REPLACE"), "obj: Object" (with dropdown for "with: Object"), "act: Action" (with dropdown for "with Action"), "sty: Style" (with dropdown for "with Style"), and a dropdown for "Likeness Infringement". It also includes a "Save Policy" button.

**Policy List** (Left Panel):

- Search filters: Keyword, Method (dropdown), Harm (dropdown).
- Buttons: "Search Policies" and "Clear".
- List of policies:
  - TomHanksMcDonald–Remove (checkbox, Edit, Delete)
  - Mosaic–Downey (checkbox, Edit, Delete)

**Examples from the original model:** (Below Test Prompt)

**Examples from the moderated model:** (Below Moderation Policy)

# Evaluation

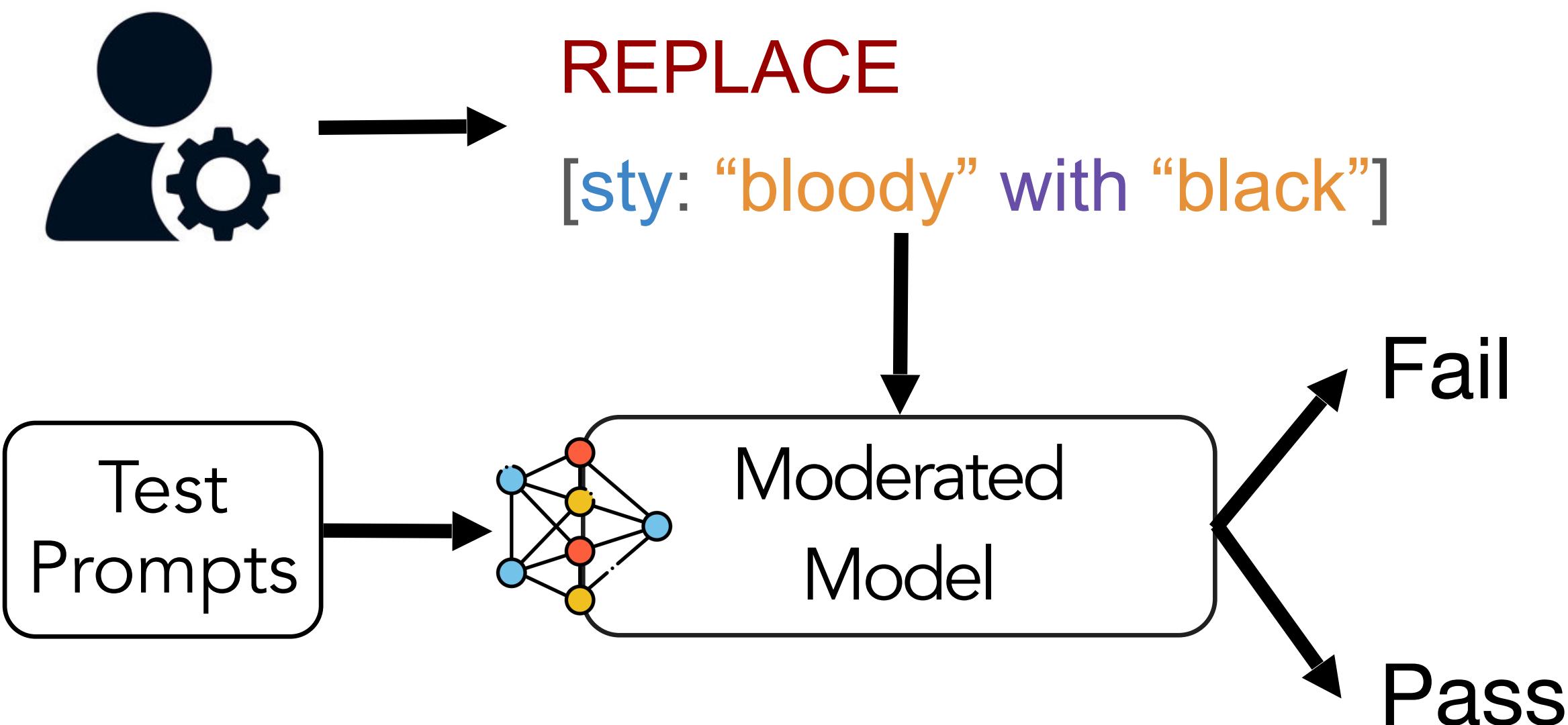
1. Effectiveness of moderation
2. Interference of multiple policies
- 3. Policy usability for admins**
- 4. Mitigation of user attacks**
5. System performance

See more details  
in the paper

## Method

# Policy usability for admins

Target: Bloody Content



**14** participants with prior moderation experience

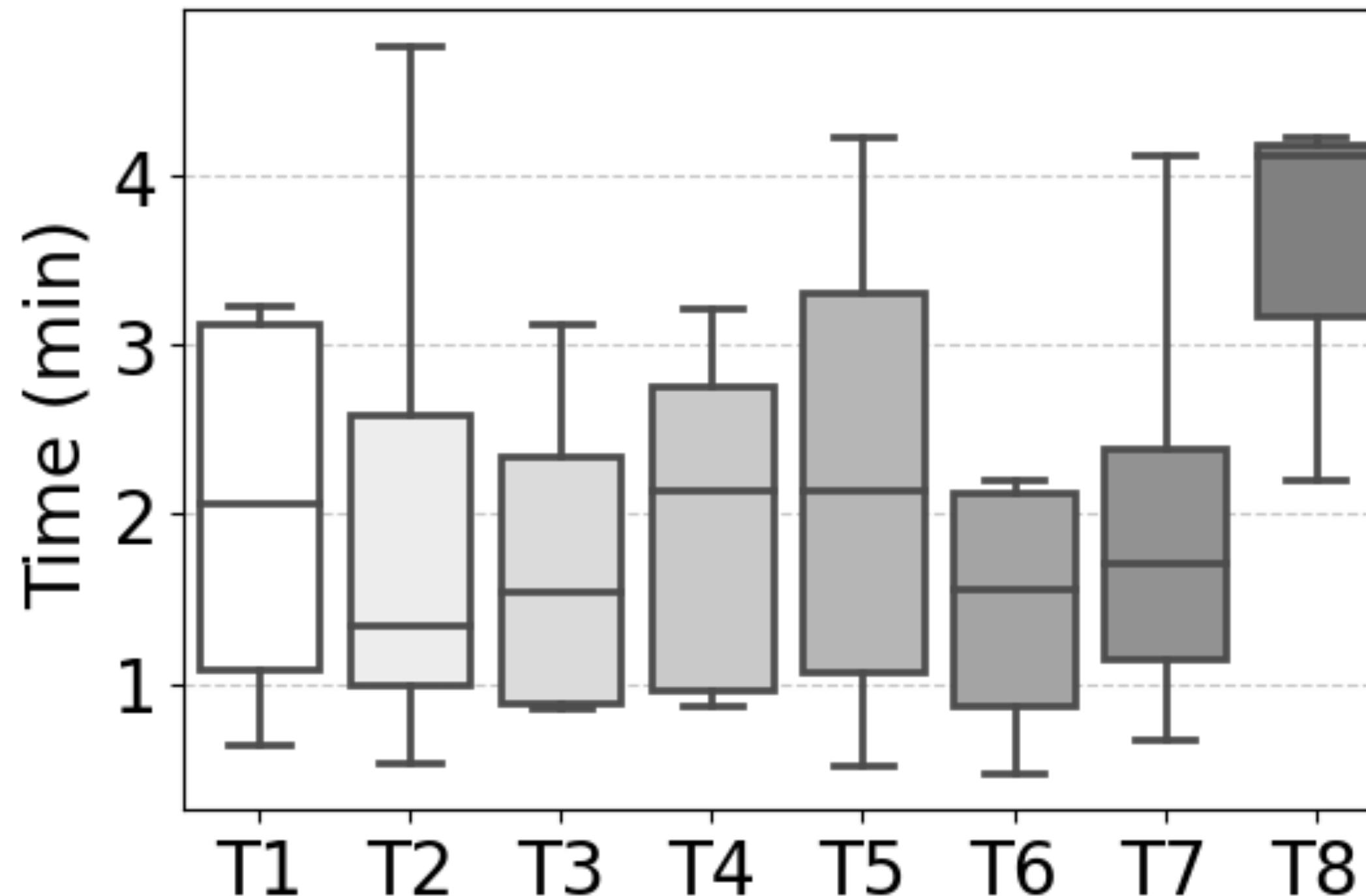
Author policies for 4 tasks

**20** unit test prompts

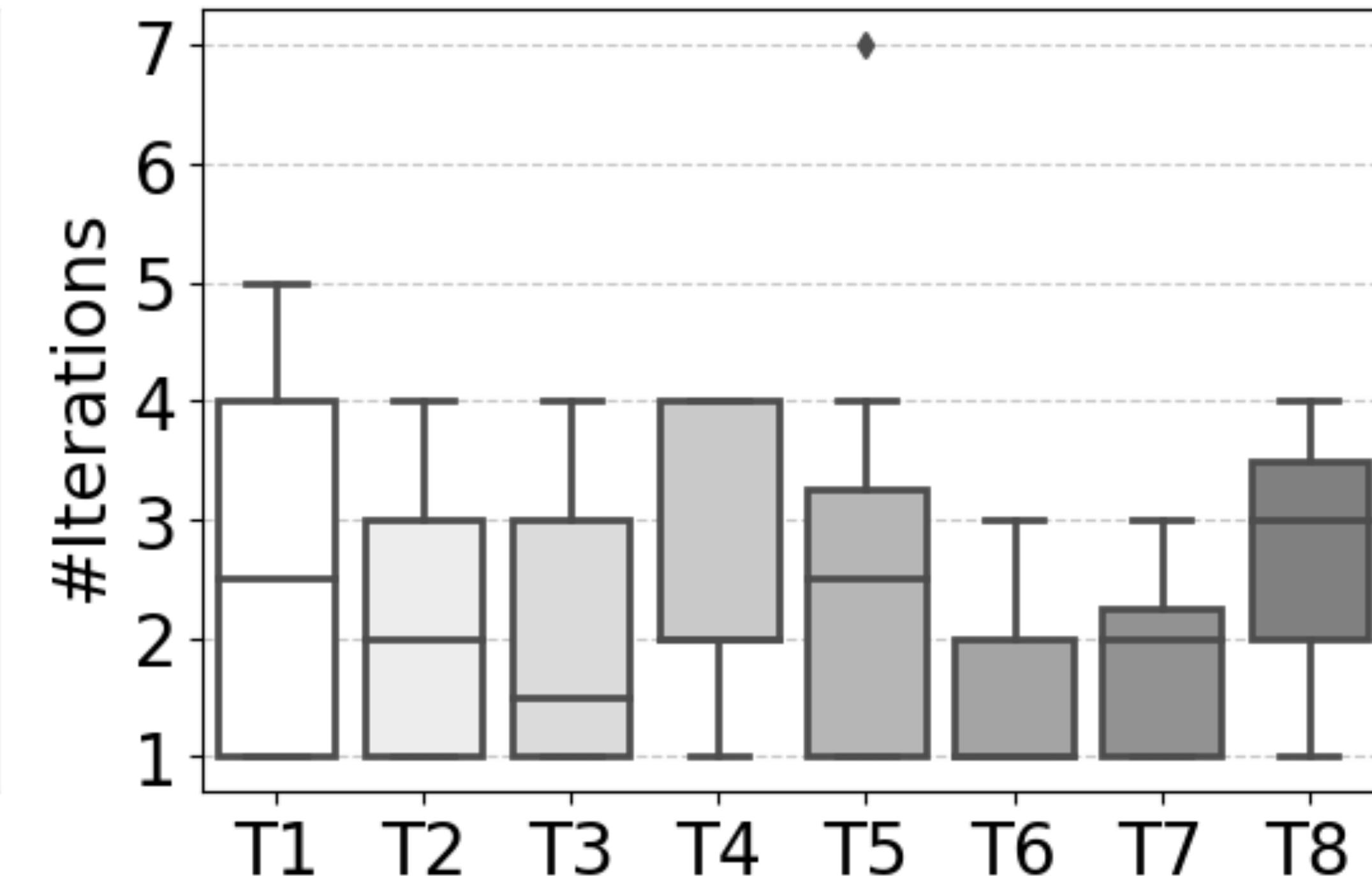
## Results

### Policy usability for admins

**2.11** minutes

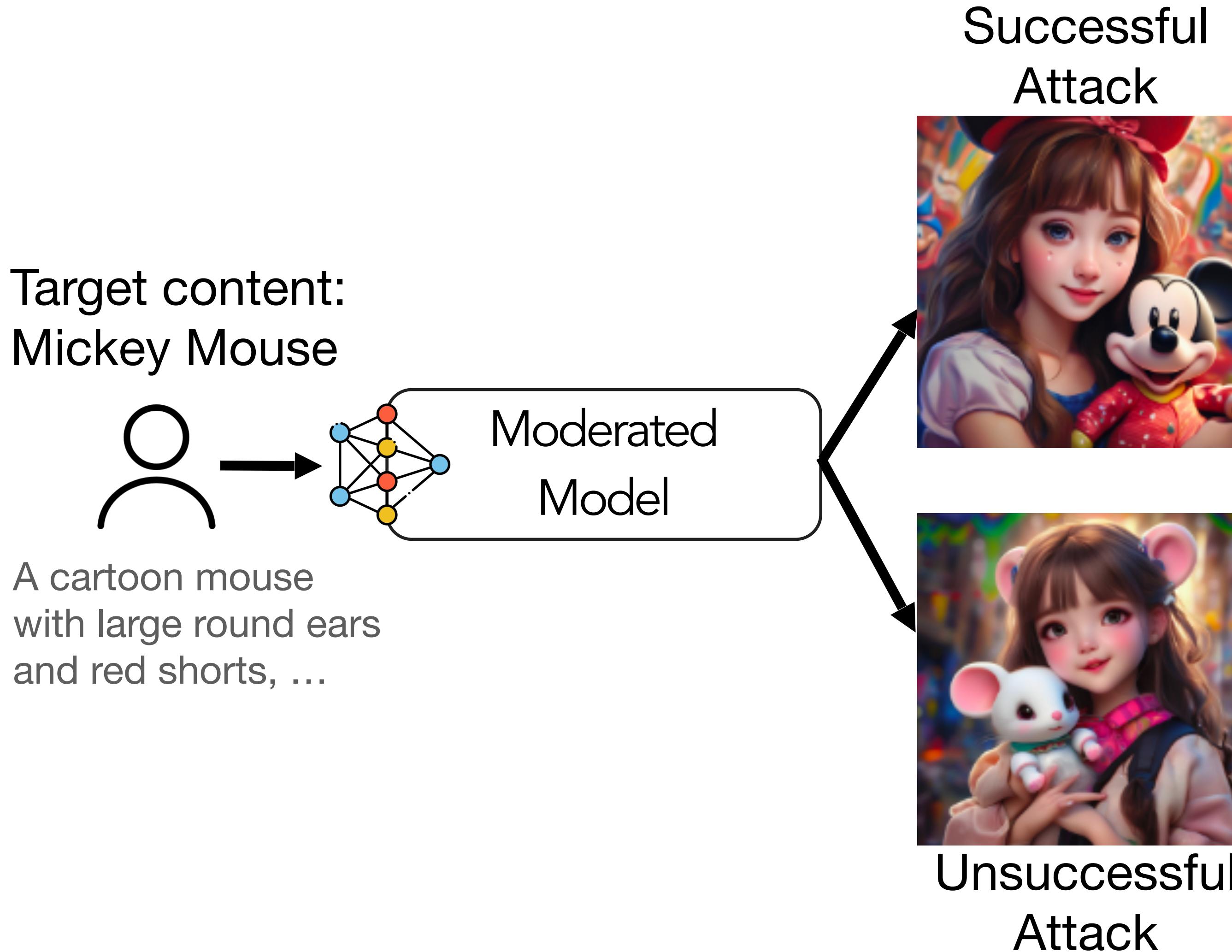


**2.29** iterations



## Method

# Mitigation of user attacks



**32** Stable Diffusion users

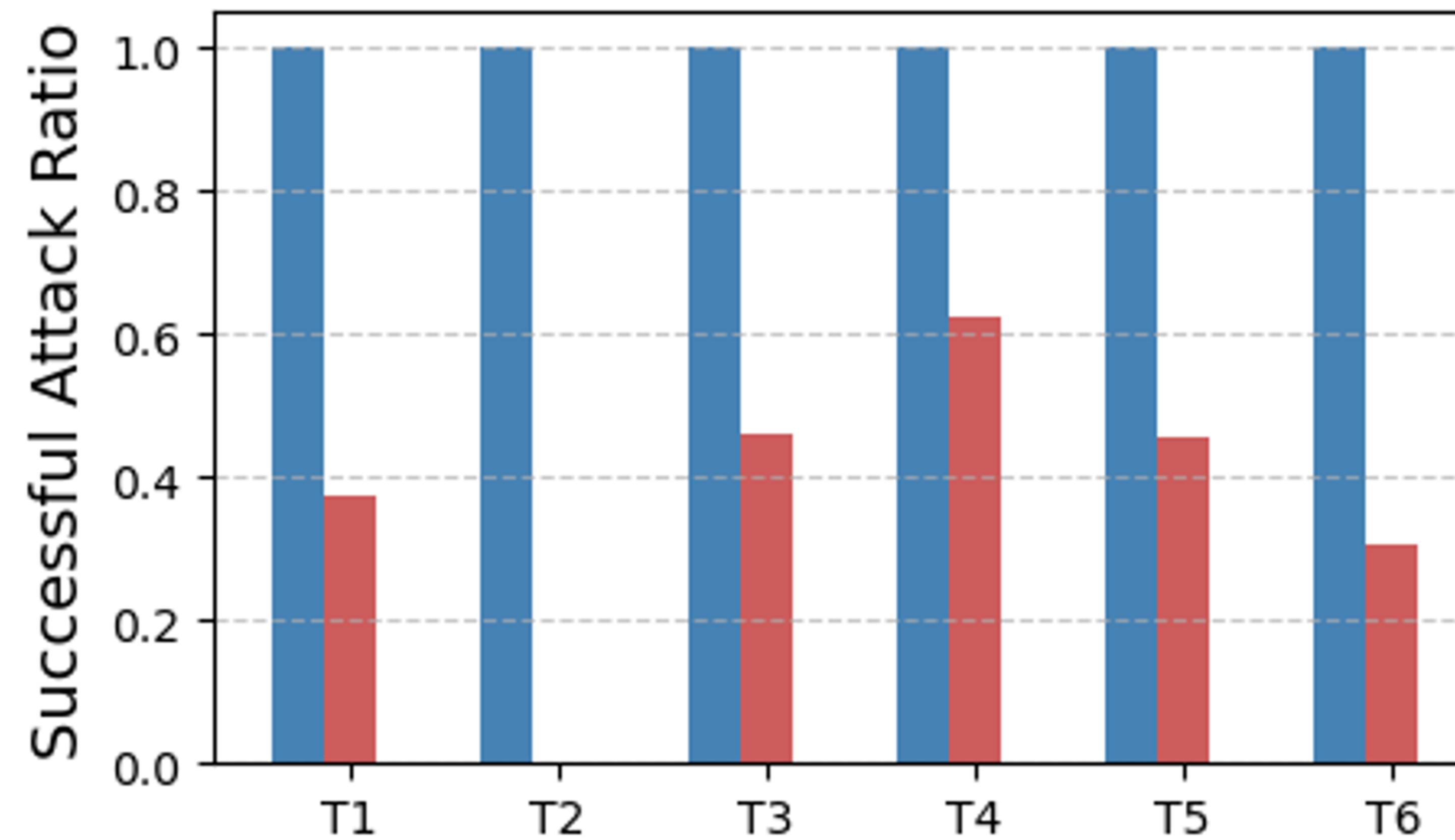
Generate target content  
within **15** attempts

Test original and  
moderated SDXL model

## Results

# Mitigation of user attacks

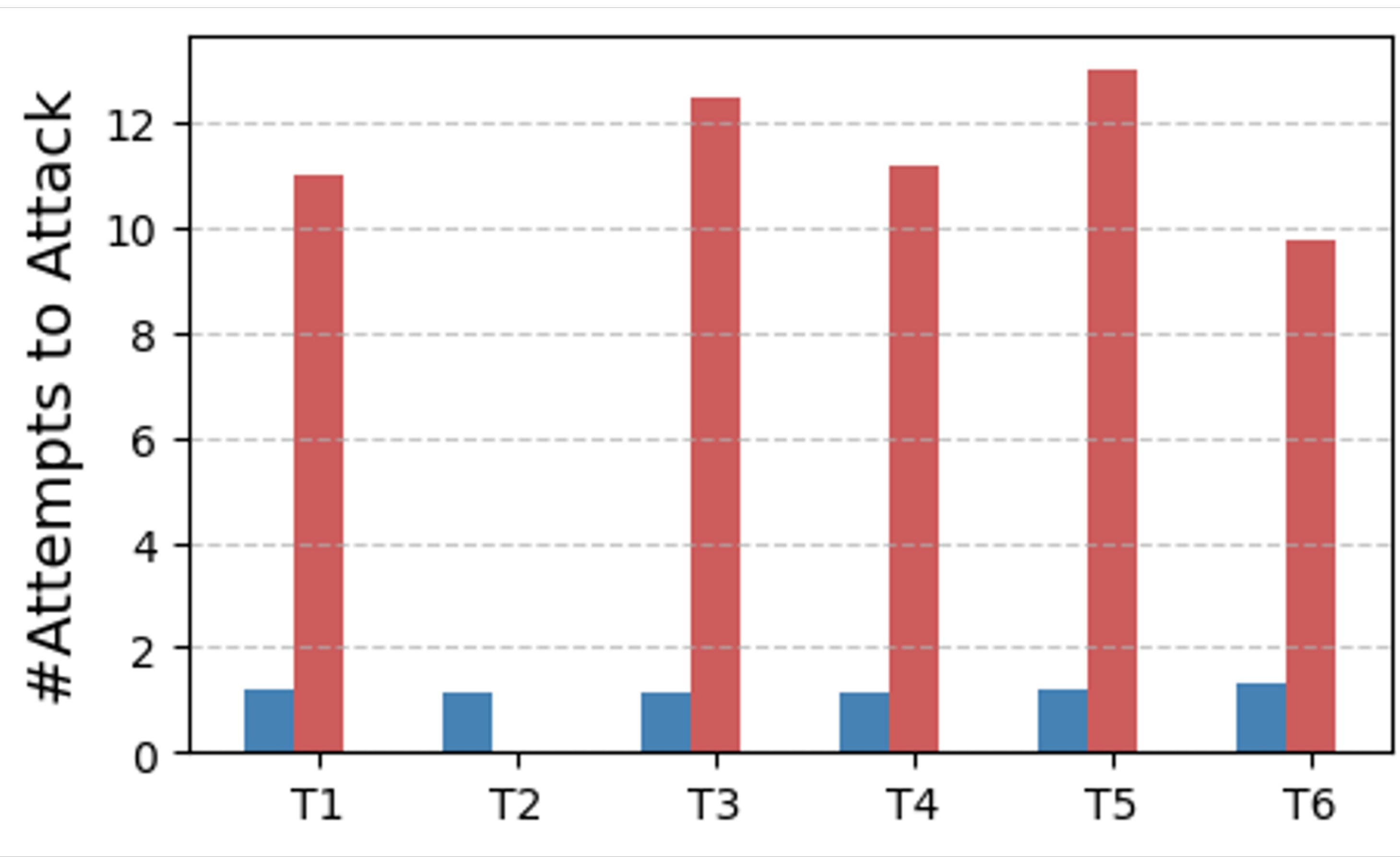
prevented **65%** of users



## 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