Policy-as-Prompt

Rethinking Content Moderation in the Age of LLMs

Konstantina Palla

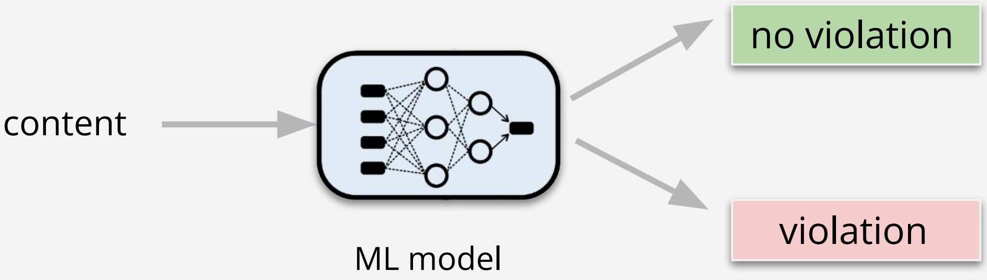
FAccT 2025, Athens

Work done with José Luis Garcia, Claudia Hauff, Francesco Fabbri, Henrik Lindström,
Daniel R.Taber, Andreas Damianou, Mounia Lalmas

Content Moderation*

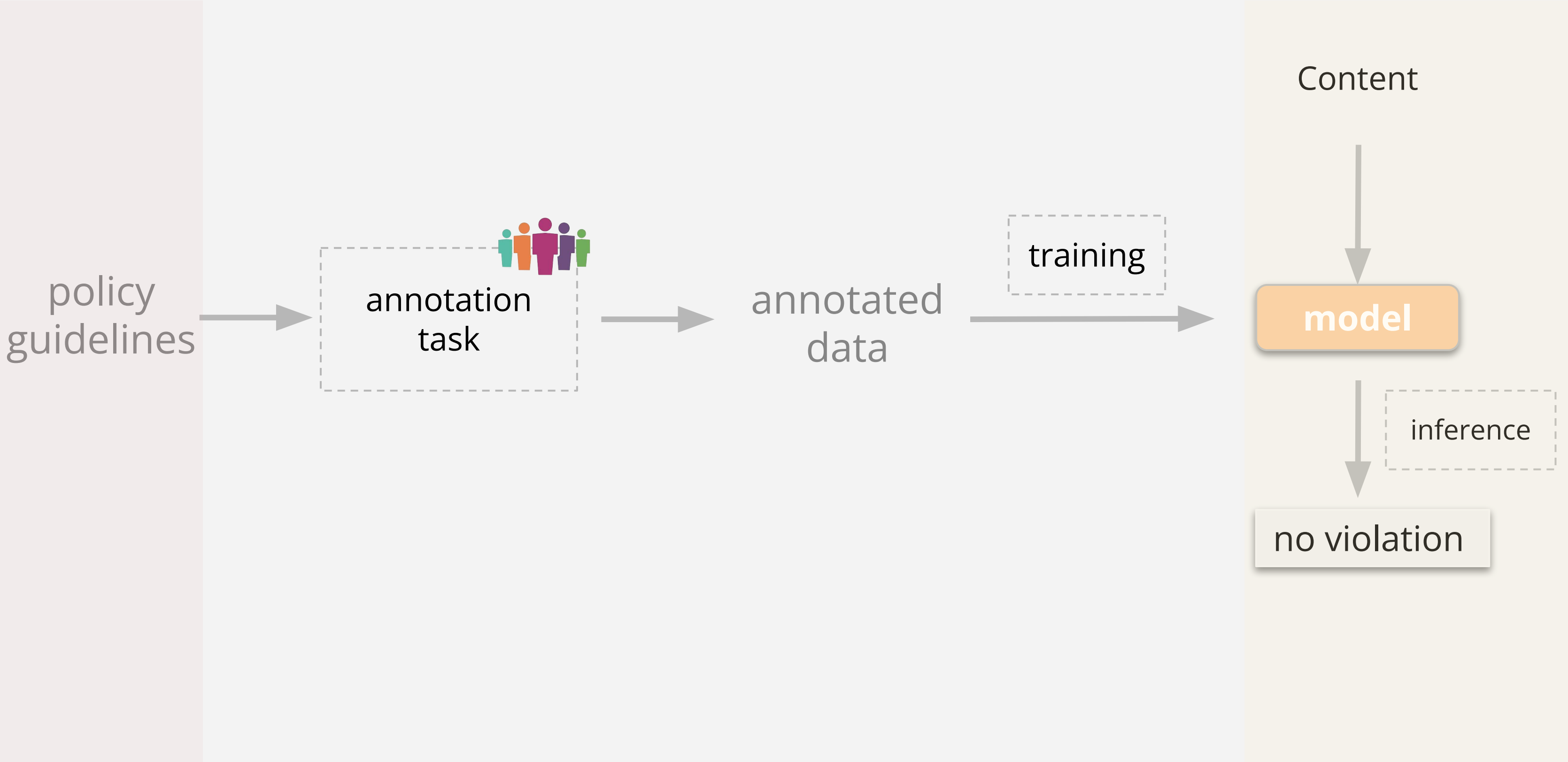
Ensure safe and inclusive online environments
Balance platform standards, user expectations, and regulations

Focus: on AI-assisted content moderation



* "content moderation" is used as a broad term encompassing both traditional moderation (decisions about allowing or disallowing content on a platform) and content sensitivity management.

Content Moderation: The ML (*traditionally*)



policy guidelines

annotation task

annotated data

training

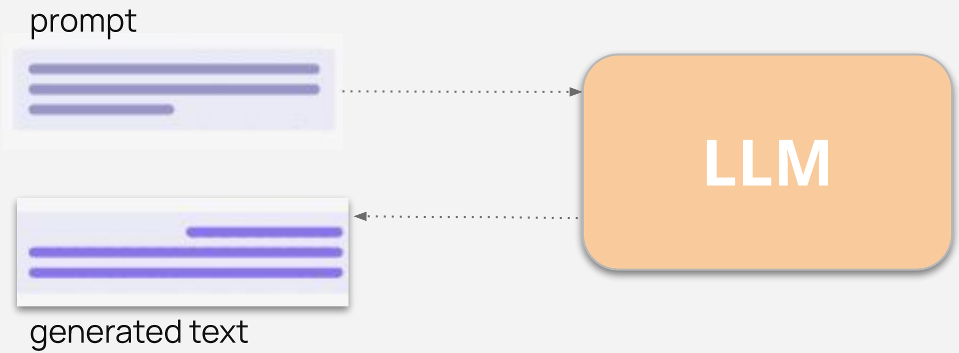
Content

model

inference

no violation

Prompting large language models

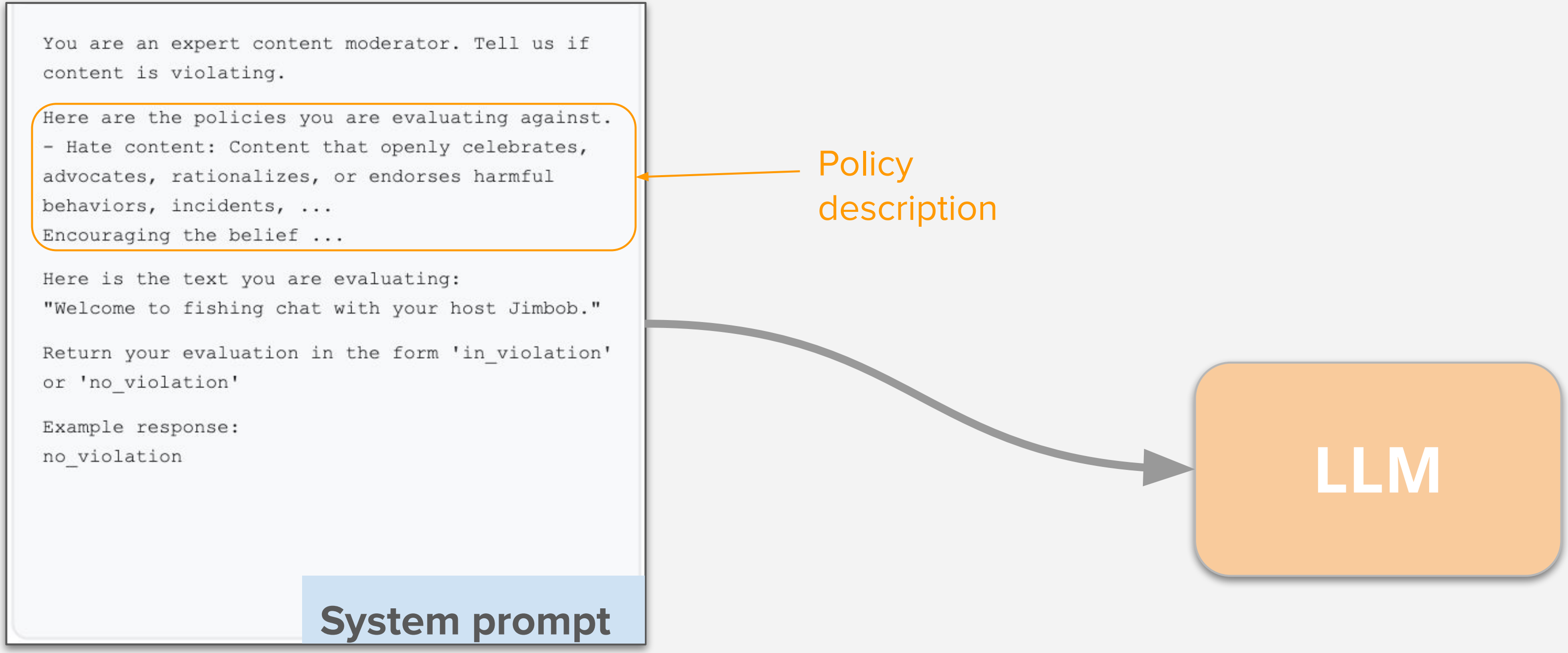


Prompt: input text that guides the model's response.

Allows for direct interaction with the model.

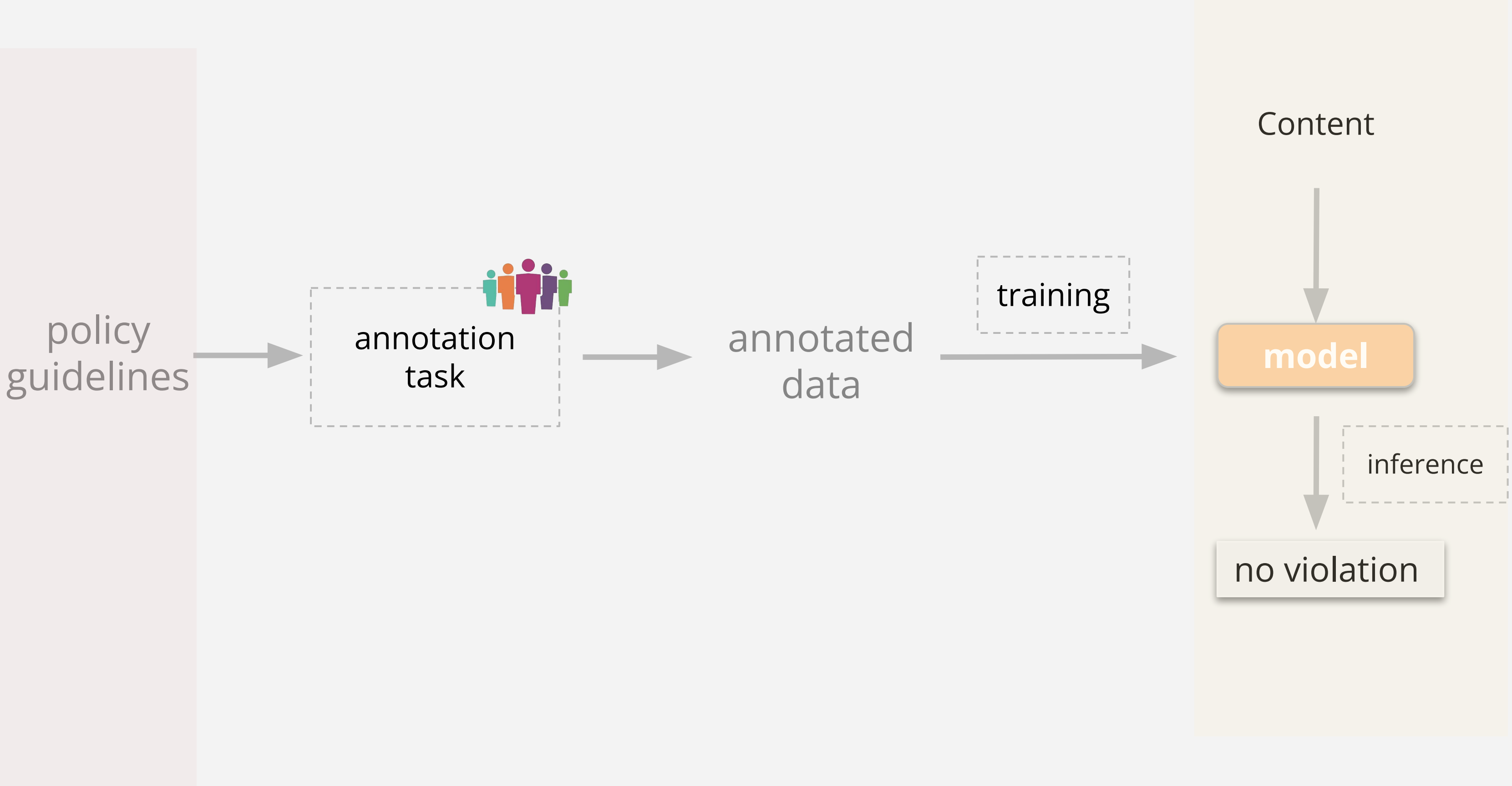
Prompts for *Safety alignment*

***Policy* Guidelines in prompt**



```
You are an expert content moderator. Tell us if
content is violating.

Here are the policies you are evaluating against.
- Hate content: Content that openly celebrates,
advocates, rationalizes, or endorses harmful
behaviors, incidents, ...
Encouraging the belief ...

Here is the text you are evaluating:
"Welcome to fishing chat with your host Jimbob."

Return your evaluation in the form 'in_violation'
or 'no_violation'

Example response:
no_violation
```

Content Moderation: The ML (*traditionally*)



policy guidelines

annotation task

annotated data

training

Content

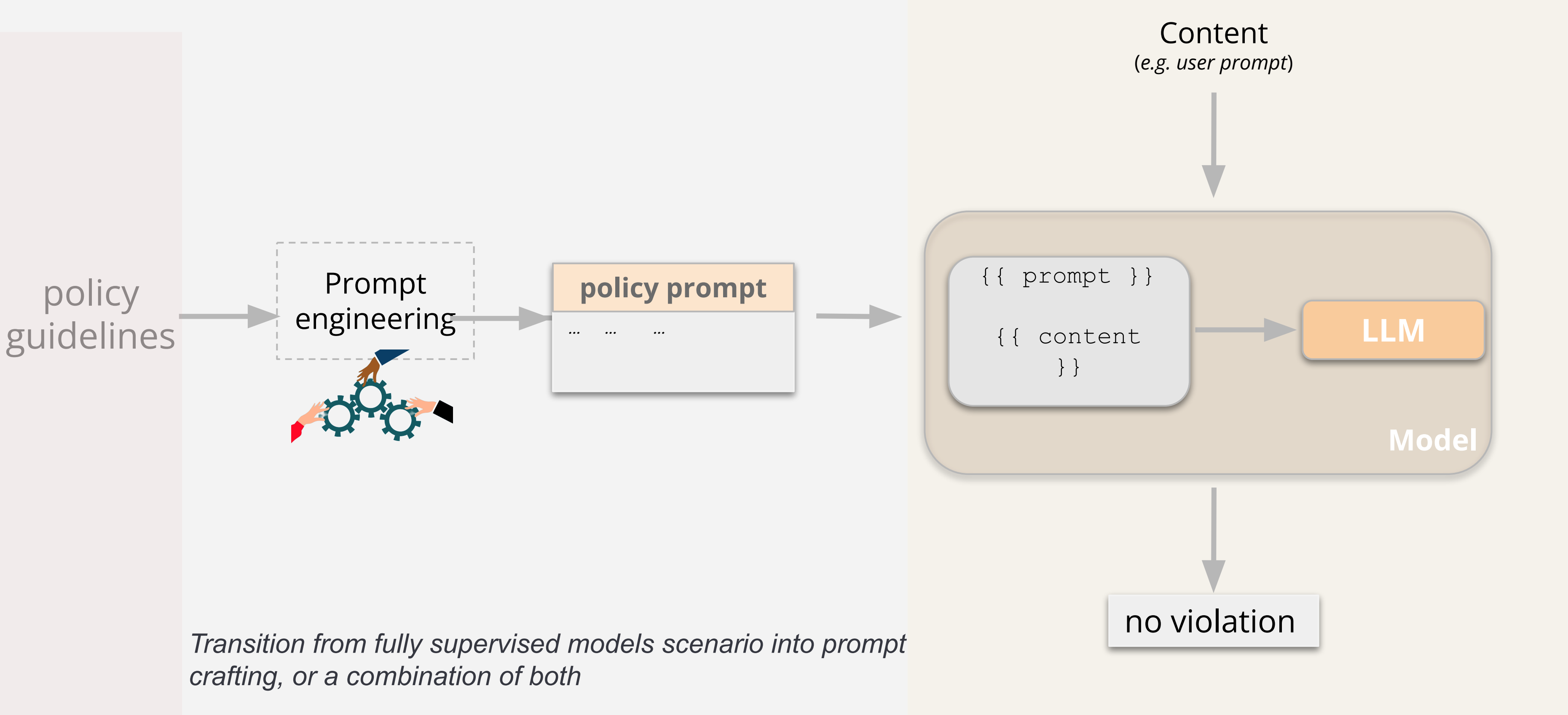
model

inference

no violation

Content Moderation

the new paradigm; *Policy-as-Prompt*

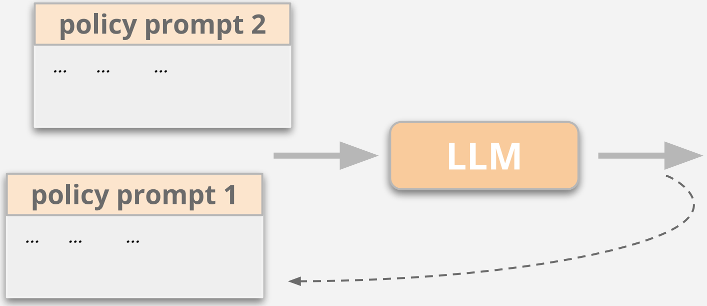


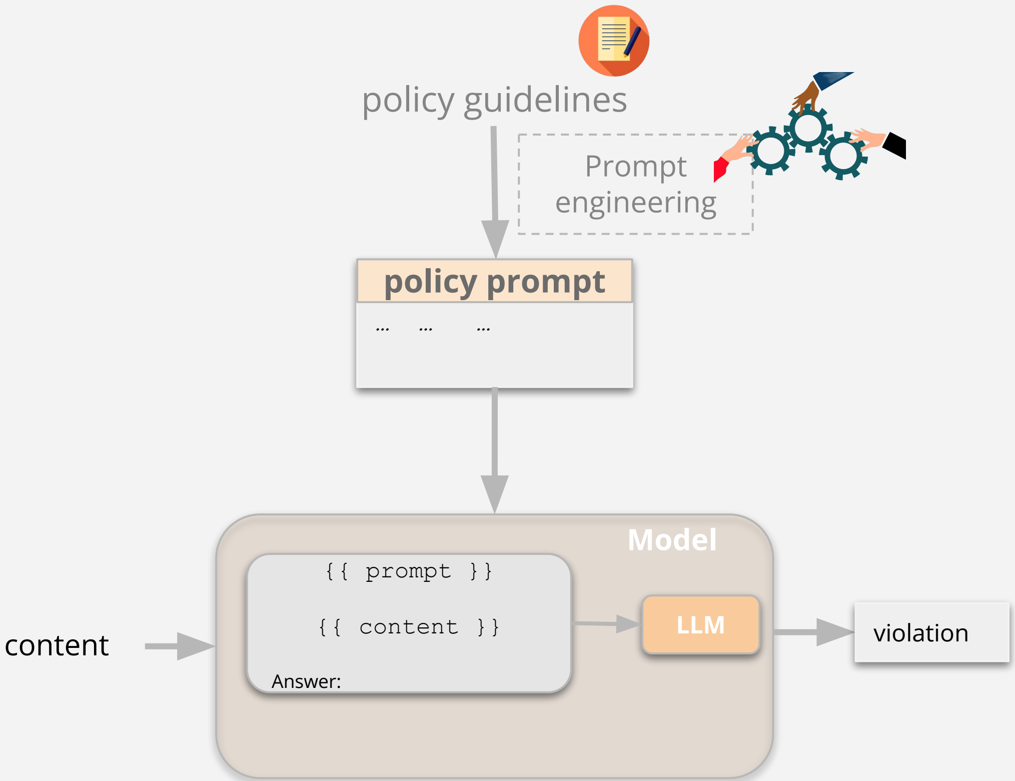
Transition from fully supervised models scenario into prompt crafting, or a combination of both

Policy-as-Prompt: the benefits

Ability to interpret policy *directly from text* - No (re)training required

Increased *flexibility* and *adaptability* in moderation (prompt modifications)





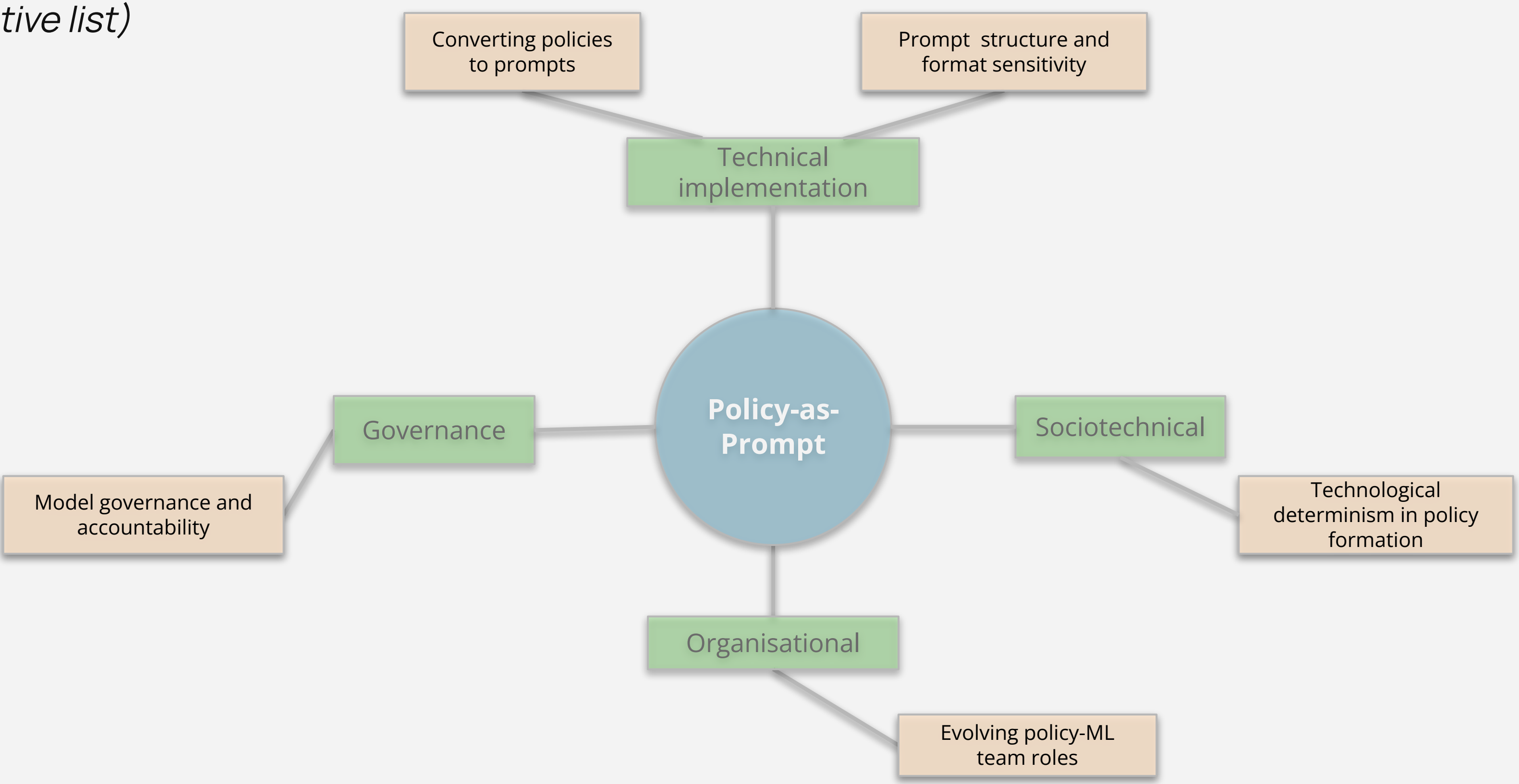
Challenges

Things we need to consider to effectively apply Policy-as-Prompt;

minimise risks and maximise benefits

Challenges in Transitioning to Policy-as-Prompt

(not an exhaustive list)



Policy-as-Prompt

Technical implementation

Converting policies to prompts

Prompt structure and format sensitivity

Governance

Model governance and accountability

Sociotechnical

Technological determinism in policy formation

Organisational

Evolving policy-ML team roles

Area

Technical Implementation

Challenge

Converting policies to prompts

How can we ensure that policy prompts accurately reflect moderation guidelines?

How can we ensure that policy prompts remain robust to formatting variations that significantly impact LLM behavior?

Traditional Supervised Approach

- Policies written exclusively for human interpretation
- Formalized across
  - Written policies, annotator guidelines, labeled training data

Policy-as-Prompt Approach

- Policies must be:
  - Human-readable
  - Machine-processable
- Full policy intent captured in a single prompt

Area

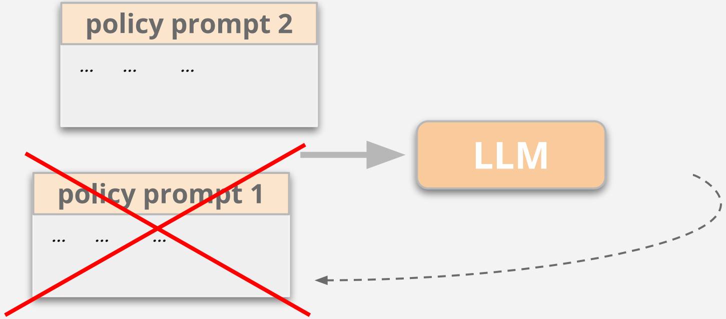
Technical Implementation

Challenge

Converting policies to prompts

How can we ensure that policy prompts accurately reflect moderation guidelines?

How can we ensure that policy prompts remain robust to formatting variations that significantly impact LLM behavior?



Verification Complexity

Prompt engineering relies on trial-and-error

LLMs struggle with nuanced contextual understanding

Subtle content detection is difficult
User request "***songs for a guilt-free feast***"

- Appears harmless
- Potential hidden reference to unhealthy eating habits

No human-annotated examples to learn from

Area

Technical Implementation

Challenge

Prompt structure and format sensitivity

How can we ensure that policy prompts accurately reflect moderation guidelines?

How can we ensure that policy prompts remain robust to formatting variations that significantly impact LLM behavior?

The critical role of prompt engineering

Text formatting is not just about appearance

Crucial in how LLMs interpret policy guidelines

LLM sensitivity

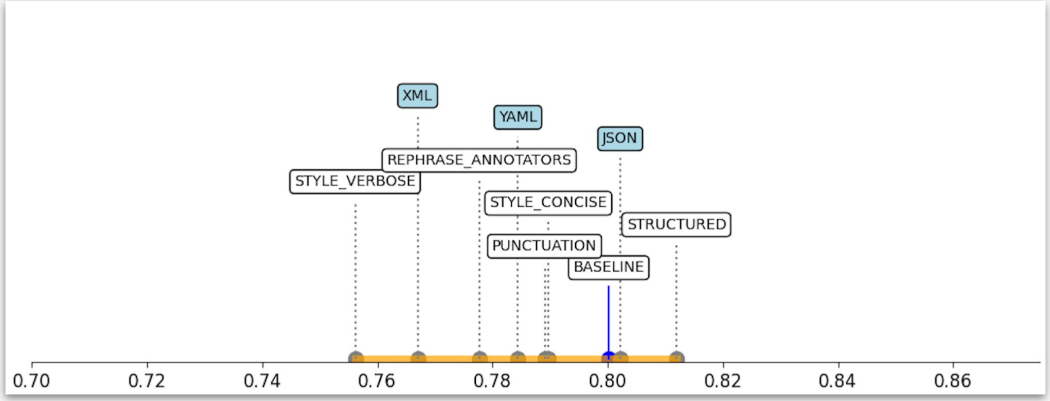
Performance varies based on:

- ○ Input length
- ○ Key information placement
- ○ Formatting details
  - ■ Whitespace
  - ■ Capitalization…

Challenge

Prompt structure and format sensitivity

Experiment: Sensitivity to Prompt variations



Accuracy varies significantly across different prompt types (from 75% to 81%)

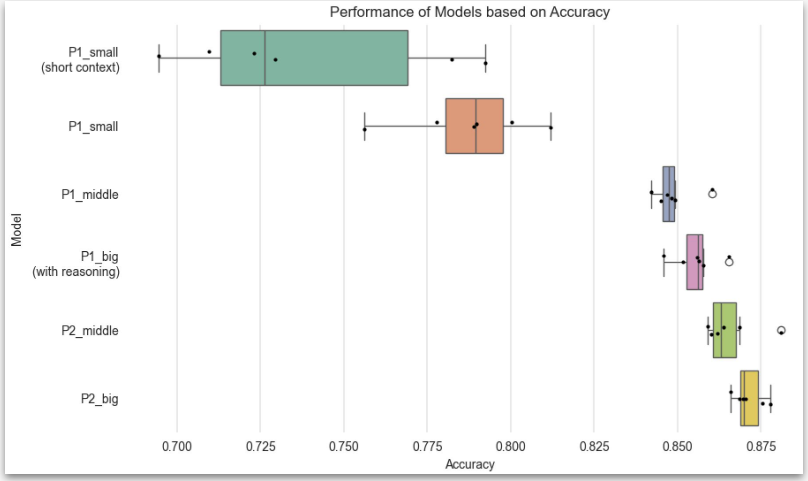
Structured prompts demonstrated the highest accuracy, highlighting model preference for organized information.

Verbose prompts had the lowest performance (bigger LLMs could potentially improve this)

Challenge

Prompt structure and format sensitivity

Experiment: Sensitivity to Prompt variations



Area

Sociotechnical

Challenge

Technological determinism in policy formation

Does embedding policies directly into LLM prompts risk oversimplifying complex societal and cultural nuances in content moderation?

Reversal of policy-technology relationship

Instead of technology serving policy goals, policies may be constrained by what LLMs can efficiently process. - ***technological determinism***

Pressure for machine-readable guidelines

Experts may be “forced” to prioritise structured, rigid rules over nuanced, context-dependent policies.

Oversimplification of complex social issues

Risk of reducing intricate moderation challenges into binary or overly simplistic rationales.

Homogenisation of policies

Cultural and contextual diversity may be lost in favour of uniform, one-size-fits-all approaches.

Area

Organisational

How does the shift to "Policy-as-Prompt" redefine the collaboration between policy teams and ML practitioners, and what new workflows are needed?

Challenge

Evolving policy-ML roles

Blurring of traditional roles

Policy authors need ML knowledge (e.g., prompt engineering).

ML practitioners engage in policy implementation.

Future Direction: "AI Policy Translators"

Area

Governance

Challenge

Model governance and accountability

How can we ensure traceability in "Policy-as-Prompt" moderation?

When moderation decisions lead to unintended outcomes, what is the process for identifying and addressing the issue?

Distributed responsibility

Trust & Safety defines policy intent, **ML engineers** structure prompts, **LLM providers** ensure contextual accuracy.

Challenges in issue resolution: *How to correct unintended moderation decisions*?

Requires cross-team collaboration: refining prompts, updating policies, or adjusting model behavior.

Complexity of documentation

Small prompt changes can impact enforcement.

Need to balance tracking modifications with operational efficiency.

Attribution & Documentation

Mitigation

Some strategies to mitigate the challenges

Mitigation Strategies
Enhanced Evaluation
Enhanced Prompt Engineering
Traceability of Prompts
Bridging Organisational Silos

Enhanced Evaluation

Addresses Technical & Sociotechnical Challenges

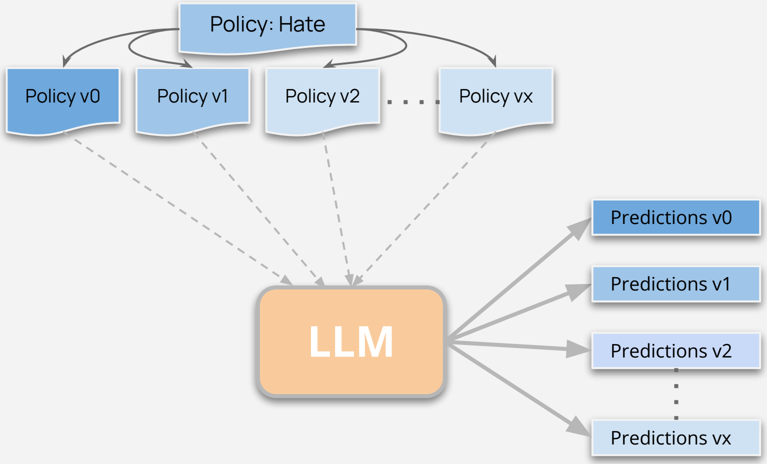
Technical sensitivity analysis - Stress test diverse prompts

Evaluate impact of formatting, phrasing and structure

Report performance across multiple prompt styles

Identify cases where similar policy phrasing leads to divergent model responses (predictive multiplicity)

E.g. use ***Rashomon sets*** to detect inconsistencies in edge cases



Sociotechnical evaluation

Beyond accuracy - assess societal readiness and adaptability

Use demographic fairness metrics to prevent disparities

Implement **case libraries** with real-world moderation edge cases to ensure nuanced, context aware decisions.

Mitigation Strategies
Enhanced Evaluation
Enhanced Prompt Engineering
Traceability of Prompts
Bridging Organisational Silos

Enhanced Prompt Engineering

Addresses Technical Challenges

Minimising machine misinterpretation → Craft prompts that capture multi-faceted content perspectives

Techniques:

Chain-of-thought reasoning, Meta-Prompting, Multi-Persona Reasoning …

Collaborative Feedback loop (- *future*)

Diverse LLMs to contribute to policy interpretation, mitigating biases and refining prompts.

Feedback loop - AI-assisted rewrites: LLMs suggest rewrites, identify gaps, loop for continuous improvement

Traceability of Prompts

Addresses Governance Challenges

Enhancing transparency and accountability

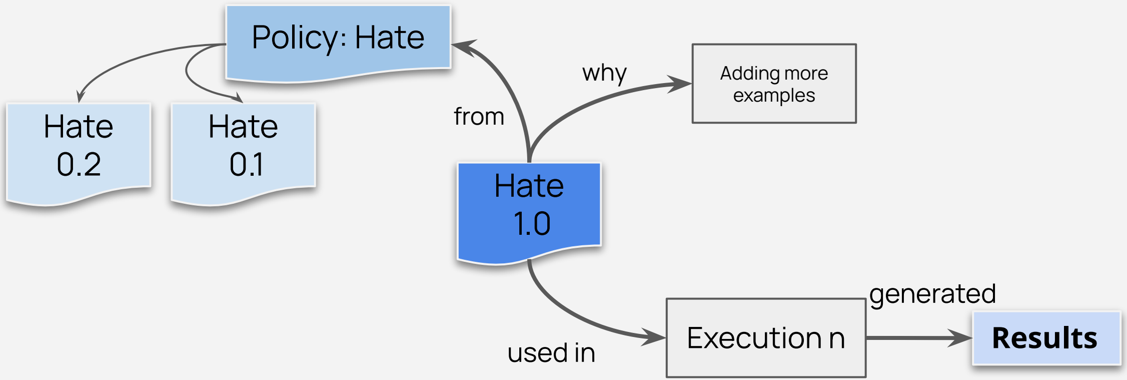
Implement a “prompt genealogy”; track changes in prompt structure, phrasting, rationale.

e.g. a version control system for prompts,similar to DVC and Pachyderm.

Key Features

Logs inputs, policy references and outputs for structured analysis

Support **audit trails, reproducibility**.



Bridging Organisational Silos

Addresses Organisational Challenges

ML practitioners vs **Policy Authors**

Joint working sessions, shared documentation practices, established feedback loops…

Long-Term vision: Develop **unified roles** that integrate policy and machine learning expertise

Conclusions

Hybrid systems today → Potential for autonomy tomorrow

LLMs currently assist moderation with human oversight and fine-tuned setups.

Transitioning to higher autonomy introduces ***new*** complexities & risks.

Path forward

Policy-as-Prompt enables dynamic, adaptable moderation frameworks.

Continued research needed to:

- address open challenges,
- improve model reliability, fairness and consistency

Read more about our work

Arxiv preprint (arXiv:2502.18695): *"Policy-as-Prompt: Rethinking Content Moderation in the Age of Large Language Models"*

Come and chat with me

konstantinap@spotify.com

The team

José Luis Garcia

Claudia hauff

Francesco Fabbri

Henrik Lindström

Daniel R. Taber

Andreas Damianou

Mounia Lalmas