AI Safety Guardrails

Tattle

Structure of the Workshop

Ways of Thinking About Guardrails

Content focused guardrails

Privacy and Copyright

RAI as Domain of Concerns (Discussed Yesterday)

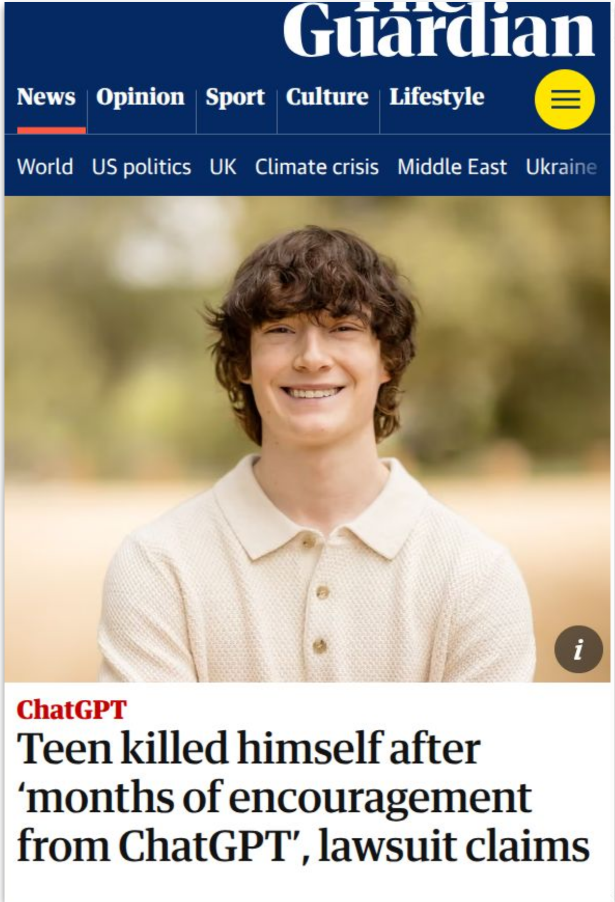
- Privacy
- Explainability
- Security against adversarial attacks such as poisoning and evasion
- Fairness/Bias
- AI Safety:
    - Toxicity
    - Hallucinations
    - Endorsing activities that lead to self harm or harm to others

The ABC Model of Platform Safety

- **Content**: analyze content for malicious or accidental exposure to harm
- **Actors**: identifying influential or malicious users
- **Behaviors**: analyzing user behaviour for unsafe engagement

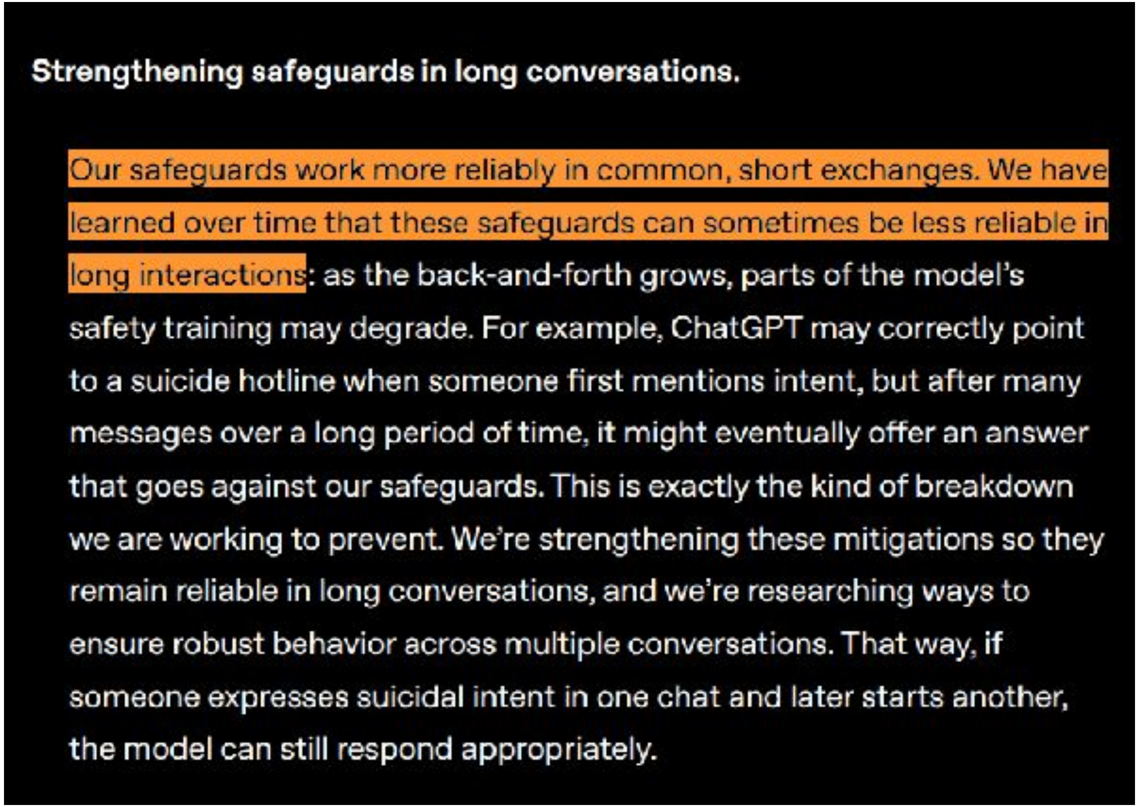
Adapted *very* loosely from Camille Francoise Framework for Disinformation Campaigns

Actor/ Unsafe Behaviour in the Context of AI



ChatGPT

Teen killed himself after 'months of encouragement from ChatGPT', lawsuit claims



Strengthening safeguards in long conversations.

Our safeguards work more reliably in common, short exchanges. We have learned over time that these safeguards can sometimes be less reliable in long interactions: as the back-and-forth grows, parts of the model's safety training may degrade. For example, ChatGPT may correctly point to a suicide hotline when someone first mentions intent, but after many messages over a long period of time, it might eventually offer an answer that goes against our safeguards. This is exactly the kind of breakdown we are working to prevent. We're strengthening these mitigations so they remain reliable in long conversations, and we're researching ways to ensure robust behavior across multiple conversations. That way, if someone expresses suicidal intent in one chat and later starts another, the model can still respond appropriately.

Technical Implementation

Lexical

Focuses on exact matching of words.

Cons: Doesn’t grasp the meaning and context of words. Limited because

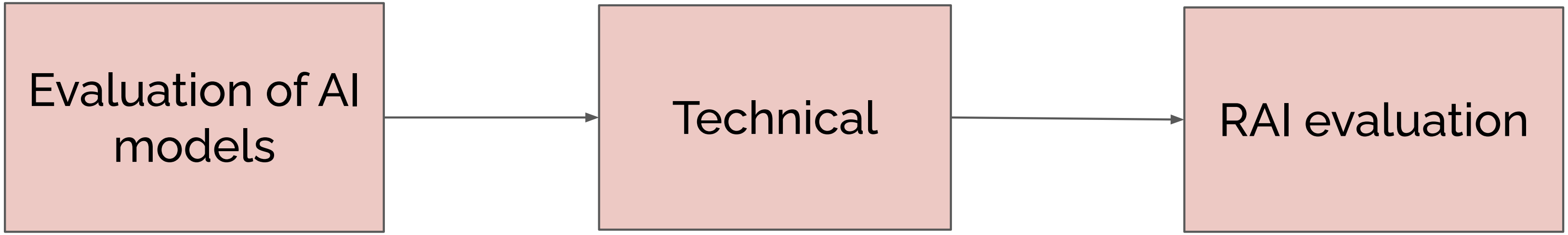
Semantic

Tries to understand the meaning and context of the query, even if exact keywords aren’t used.

E.g., machine learning models

Cons: Slower performance and complex in nature

Scope of the discussion



Evaluation of AI models

Technical

RAI evaluation

Tattle Slur List

The slur list was created from 2021-22 by 30 researchers and activists in Indian languages in the process of building a robust dataset for the plugin's features.

In 2023, gender and feminist rights organisations partnered with us to add slur words and contribute to metadata on the slurs by participating in online annotation sessions. We are continuing to expand our slur list in this iteration.

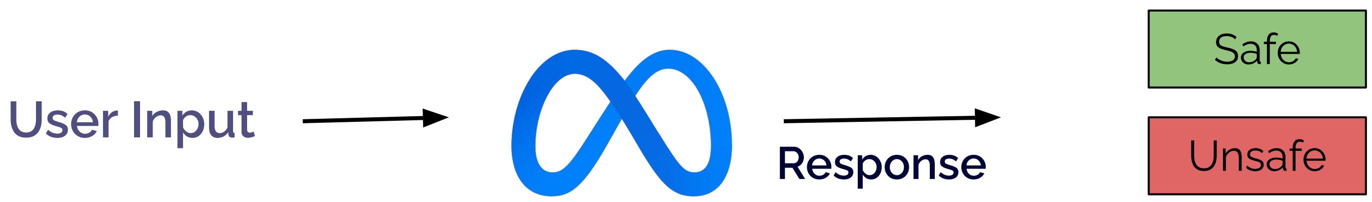
At present the slur list consists of **630** words in **4** languages: Indian-English, Hindi, Tamil and Malayalam.

https://github.com/tattle-made/Uli/blob/main/browser-extension/plugin/scripts/2023-12-21-slur-metadata/data.csv

Custom Flag List

- A list of words and phrases tailored for your use-case
- Captures local, cultural, or use-case specific terms

Using Llama Guard



User Input

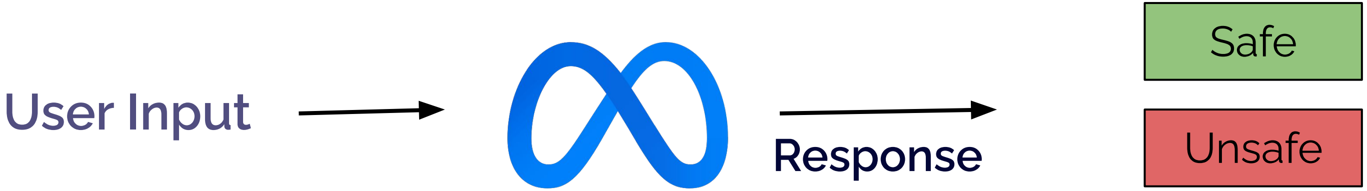
Response

Safe

Unsafe

Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations

Using Llama Guard



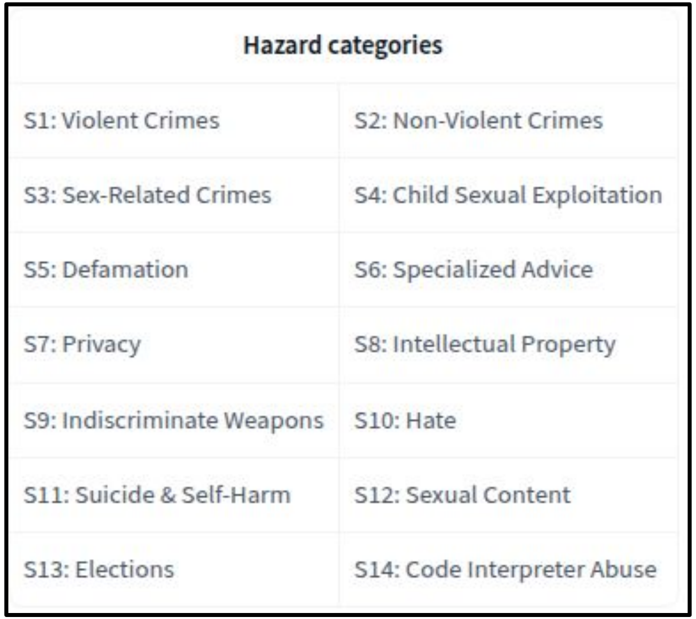
User Input

Response

Safe

Unsafe

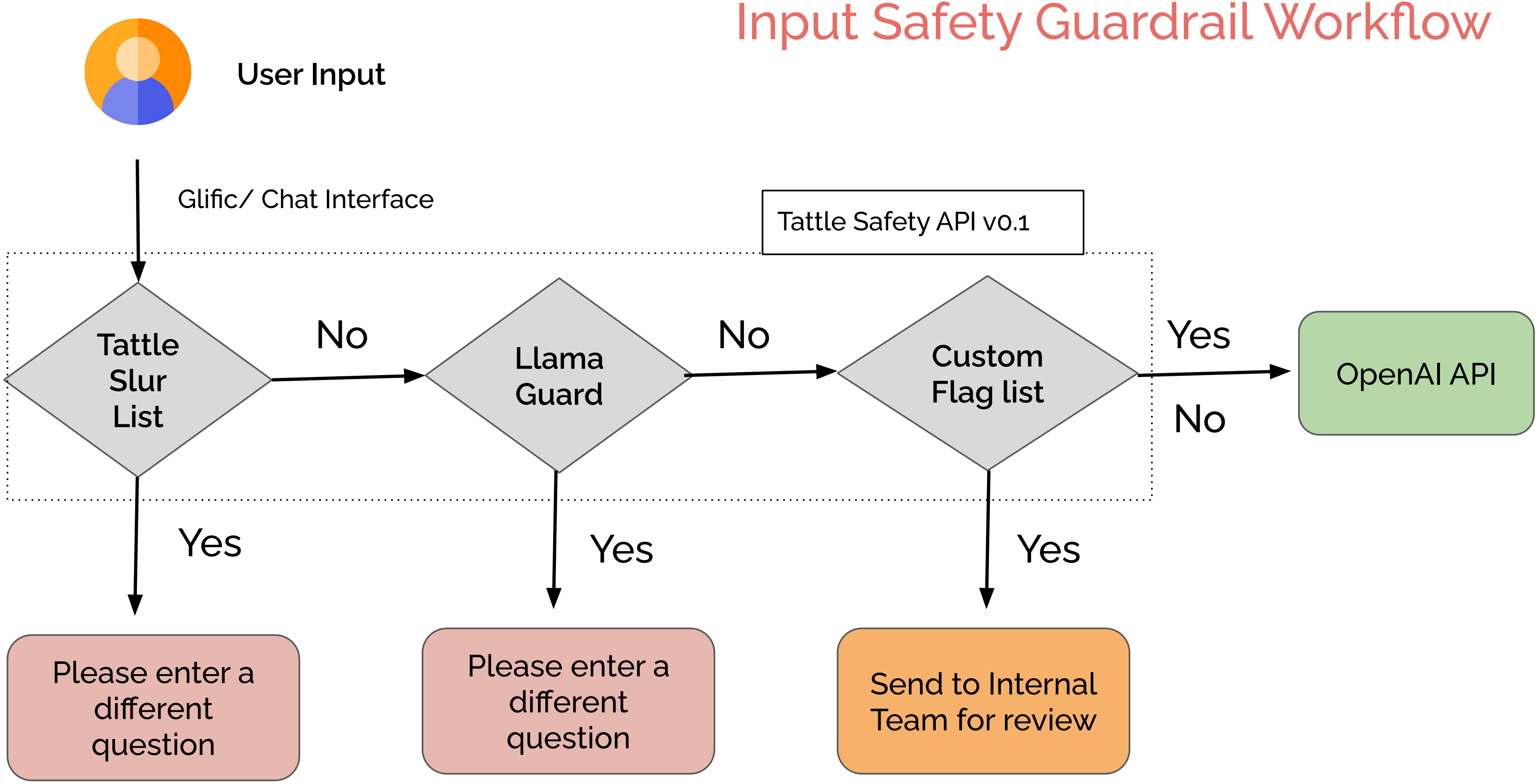
Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations



Hazard categories	
S1: Violent Crimes	S2: Non-Violent Crimes
S3: Sex-Related Crimes	S4: Child Sexual Exploitation
S5: Defamation	S6: Specialized Advice
S7: Privacy	S8: Intellectual Property
S9: Indiscriminate Weapons	S10: Hate
S11: Suicide & Self-Harm	S12: Sexual Content
S13: Elections	S14: Code Interpreter Abuse

https://huggingface.co/meta-llama/Llama-Guard-3-8B

Input Safety Guardrail Workflow



User Input

Glific/ Chat Interface

Tattle Safety API v0.1

Tattle Slur List

No

Llama Guard

No

Custom Flag list

Yes

OpenAI API

No

Yes

Yes

Yes

Please enter a different question

Please enter a different question

Send to Internal Team for review

Demo

⚠ Content Warning - It is possible that the demo might lead to conversation on sensitive topics or use of derogatory terms.

Tattle Safety API Response

Input

- “text”: a string of text content

https://github.com/tattle-made/llamaguard-uli

Output (Response)

- “should_moderate”: a boolean flag
- “reason”: "safe" | "tattle_slur_list" | "llama_guard" | "flag_list"

- Other misc metadata like response time, status code

Privacy and Copyright Considerations

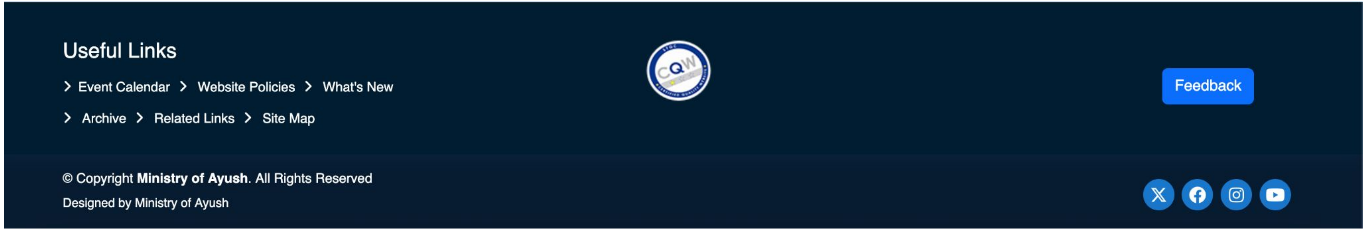
- Safety considerations when building chatbots/prediction models also include privacy and copyright.

- In this context, there are two broad types of data that you may be collecting: private data (from users) and scraped data (from datasets, internet)

- Private data refers to data not available in the public domain, and includes personal data that users may share with you.
  - Data protection is an important facet for all of us
  - How do you sustain as an organisation?

Privacy and Copyright Considerations

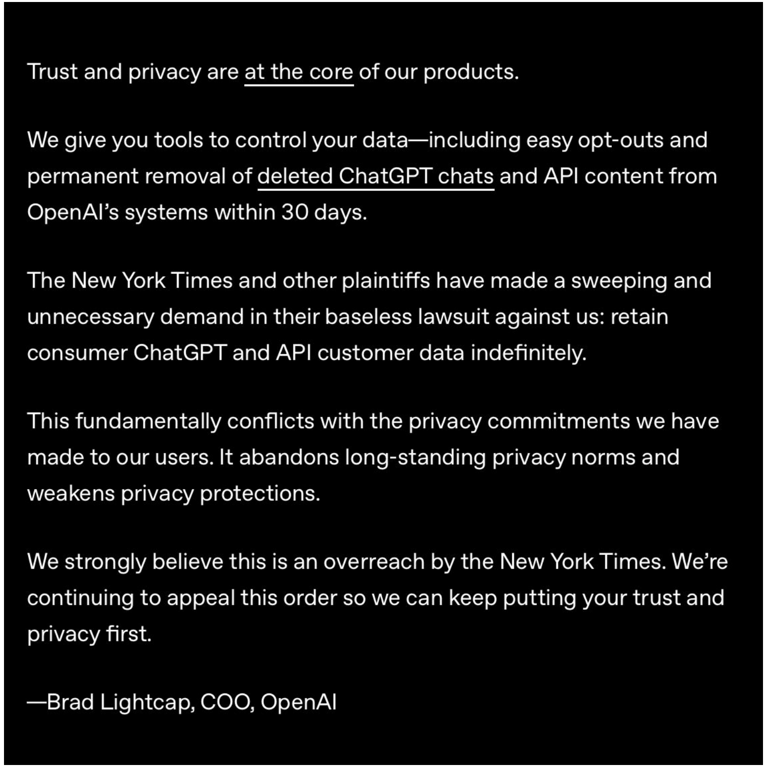
- Scraped data: you may collect data from the internet that is available publicly, but there are distinctions in how you may use them:
  - Public domain: The information is not owned by anybody, is not protected under the IP regime, and is free to use
  - Proprietary information: Information is copyrighted, and must be used subject to the license restrictions stated by the owner (©, all rights reserved)
  - Open license: Information is copyrighted, but under the open license you have the right to modify/copy as the case may be.



© Copyright **Ministry of Ayush**. All Rights Reserved
Designed by Ministry of Ayush

Contentions around Privacy and Copyright

- Scraping raw data public documents v. scraping from a dataset of public documents
- Mosaic effect: aggregating multiple datasets may reveal sensitive information that individually may not be discernible.
  - Netflix Prize Dataset study: de-anonymising movie viewing records
- Enterprise API v Consumer API



Trust and privacy are at the core of our products.

We give you tools to control your data—including easy opt-outs and permanent removal of deleted ChatGPT chats and API content from OpenAI's systems within 30 days.

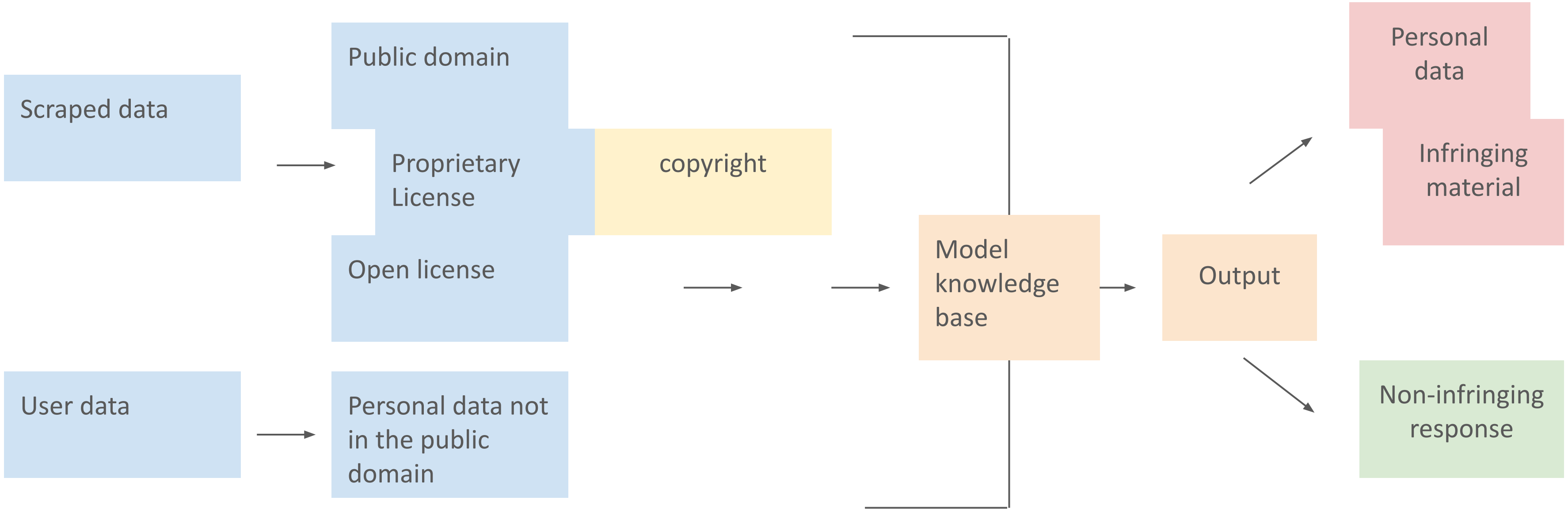
The New York Times and other plaintiffs have made a sweeping and unnecessary demand in their baseless lawsuit against us: retain consumer ChatGPT and API customer data indefinitely.

This fundamentally conflicts with the privacy commitments we have made to our users. It abandons long-standing privacy norms and weakens privacy protections.

We strongly believe this is an overreach by the New York Times. We're continuing to appeal this order so we can keep putting your trust and privacy first.

—Brad Lightcap, COO, OpenAI

Data Mapping: Privacy and Copyright



Preliminary Checklist

There are two types of data you can collect: personal data and public data
With personal data, your checklist might look something like this:

- ❑ What personal data are you collecting?
- ❑ What is the purpose of the collection?
- ❑ How long do you plan on storing/retaining this data?
- ❑ Are you complying with industry standards for securing personal data?
- ❑ If you are collecting personal data from users,
  - ❑ Are you taking their consent?
  - ❑ How can they send requests for data deletion? What data deletion policy do you have in place?
- ❑ Are you using an enterprise account or a consumer account?

*not to be constituted as legal advice

Public data

With public data, your checklist looks slightly different:

- ❑ Check if there are any licensing requirements attached to it
  - ❑ Proprietary data should not be scraped, and should be used subject to a licensing agreement with the provider.
  - ❑ Data available under open data licenses have attribution requirements or specify how this data can be used, ensure you comply with that.

*not to be constituted as legal advice

THANK YOU!

Tattle

www.tattle.co.in