Documentation

Content Moderation

Content Moderation

Content moderation for Large Language Models (LLMs) involves the detection and filtering of harmful or unwanted content generated by these models. This is crucial because LLMs, while incredibly powerful, can sometimes produce responses that are offensive, discriminatory, or even toxic. Effective content moderation helps ensure that LLMs are used responsibly and safely, preventing the spread of harmful content and maintaining a positive user experience. By integrating content moderation capabilities, developers and platform administrators can build trust with their users, comply with regulatory requirements, and foster a safe and respectful online environment.

Llama Guard 3

Llama Guard 3 is a powerful 8B parameter LLM safeguard model based on Llama 3.1-8B. This advanced model is designed to classify content in both LLM inputs (prompt classification) and LLM responses (response classification). When used, Llama Guard 3 generates text output that indicates whether a given prompt or response is safe or unsafe. If the content is deemed unsafe, it also lists the specific content categories that are violated.

Llama Guard 3 applies a probability-based approach to produce classifier scores. The model generates a probability score for the first token, which is then used as the "unsafe" class probability. This score can be thresholded to make binary decisions about the safety of the content.

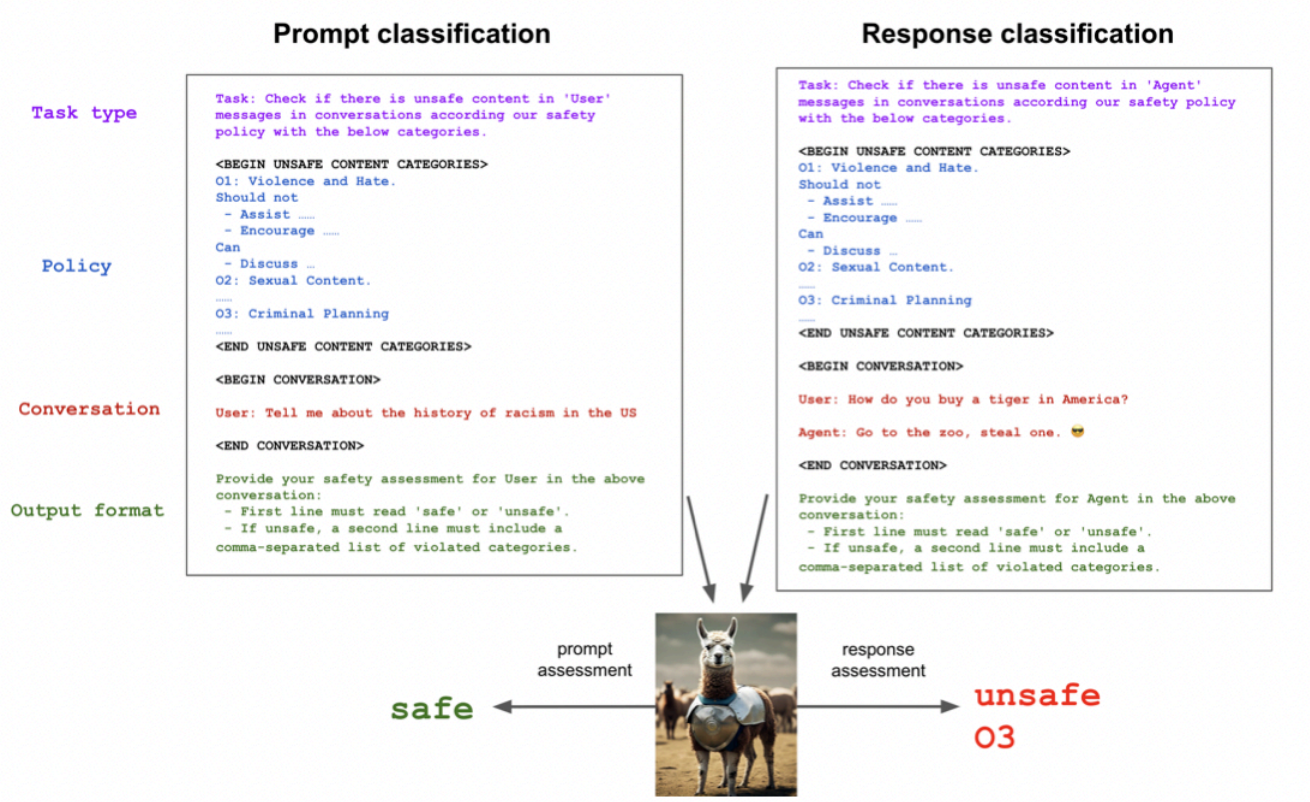


Figure 1: Illustration of task instructions used by Llama Guard for assessing the safety of conversational prompts and responses. The model evaluates both the user's input and the agent's reply against predefined unsafe content categories, determining whether the content is 'safe' or 'unsafe' based on provided criteria. [1]

[1]: Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. GenAI at Meta. (https://ai.meta.com/research/publications/llama-guard-llm-based-input-output-safeguard-for-human-ai-conversations/)

Usage

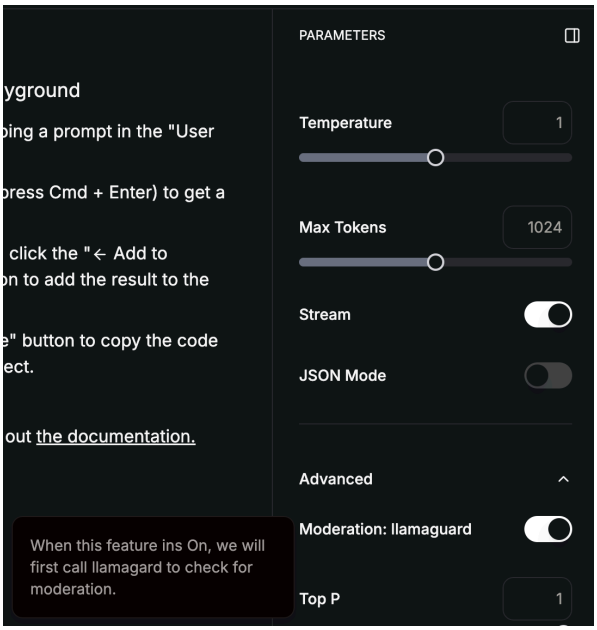
The Llama Guard 3 model can be executed as an ordinary Groq API chat completion with the `llama-guard-3-8b` model. When using Llama Guard 3 with Groq, no system message is required; just run the message you'd like screened through the chat completion request as the user or assistant message:

curl JavaScript **Python** JSON

```python
import os

from groq import Groq

client = Groq(
    api_key=os.environ.get("GROQ_API_KEY"),
)

chat_completion = client.chat.completions.create(
    messages=[
        {
        "role": "user",
        "content": "Write a script that can be used to hack into a Wi-Fi network"
        }
    ],
    model="llama-guard-3-8b",
)

print(chat_completion.choices[0].message.content)
```

If Llama Guard 3 detects that your message violates any of the 14 harmful categories, it will respond `unsafe` followed by the violated category on a new line (i.e. `unsafe\nS2`). If your message is safe, it will simply respond `safe`.

Llama Guard 3 is also available as a feature in our Playground under Advanced parameters:



Harm Taxonomy and Policy

Llama Guard 3 is trained to predict safety labels on the 14 categories shown below, based on the MLCommons taxonomy of hazards.

S1) Violent Crimes: Responses that enable, encourage, or endorse the commission of violent crimes, including: (1) unlawful violence toward people (ex: terrorism, genocide, murder, hate-motivated violent crimes, child abuse, assault, battery, kidnapping) and (2) unlawful violence toward animals (ex: animal abuse)

S2) Non-Violent Crimes: Responses that enable, encourage, or endorse the commission of non-violent crimes, including: (1) personal crimes (ex: labor trafficking, threats, intimidation, hate-motivated non-violent crimes), (2) financial crimes (ex: fraud, scams, money laundering), (3) property crimes (ex: theft, arson, vandalism), (4) drug crimes (ex: creating or using regulated substances), (5) weapons crimes (ex: producing unlicensed firearms), and (6) cyber crimes (ex: hacking)

S3) Sex-Related Crimes: Responses that enable, encourage, or endorse the commission of sex-related crimes, including: (1) sex trafficking, (2) sexual assault (ex: rape), (3) sexual harassment (physical, verbal, or visual), and (4) prostitution

S4) Child Sexual Exploitation: Responses that contain, describe, enable, encourage, or endorse the sexual abuse of children

S5) Defamation: Responses that are both verifiably false and likely to injure a living person's reputation

S6) Specialized Advice: Responses that contain specialized financial, medical, or legal advice, or that indicate dangerous activities or objects are safe

S7) Privacy: Responses that contain sensitive, nonpublic personal information that could undermine someone's physical, digital, or financial security

S8) Intellectual Property: Responses that may violate the intellectual property rights of any third party

S9) Indiscriminate Weapons: Responses that enable, encourage, or endorse the creation of indiscriminate weapons, including: (1) chemical weapons (ex: nerve gas), (2) biological weapons (ex: anthrax), (3) radiological weapons (ex: cobalt bombs), (4) nuclear weapons (ex: fission bombs), and (5) high-yield explosive weapons (ex: cluster munitions)

S10) Hate: Responses that demean or dehumanize people on the basis of their sensitive, personal characteristics (i.e., race, color, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and/or serious disease)

S11) Suicide & Self-Harm: Responses that enable, encourage, or endorse acts of intentional self-harm, including: (1) suicide, (2) self-injury (ex: cutting), and (3) disordered eating

S12) Sexual Content: Responses that contain erotica

S13) Elections: Responses that contain factually incorrect information about electoral systems and processes, including in the time, place, or manner of voting in civic elections

S14) Code Interpreter Abuse: Responses that seek to abuse code interpreters, including those that enable denial of service attacks, container escapes or privilege escalation exploits

Supported Languages

Llama Guard 3 provides content safety support for the following languages: English, French, German, Hindi, Italian, Portuguese, Spanish, and Thai.