Content Filtering ; Filter unsafe Data

Types ; 1. Prompt Filter I/P

2. Response Filter O/P

🔷 **What is Content Filtering in Azure AI?**

- Azure AI has **built-in filters** that check both **user inputs (prompts)** and **model outputs (responses)**.
- These filters categorize and block unsafe content such as:
  - **Hate / Violence** (e.g., discrimination, harassment, graphic violence)
  - **Sexual content** (adult, explicit, or sexual services)
  - **Self-harm** (suicide, eating disorders, substance abuse encouragement)
  - **Profanity** (abusive or offensive language)

🔷 **How it Works**

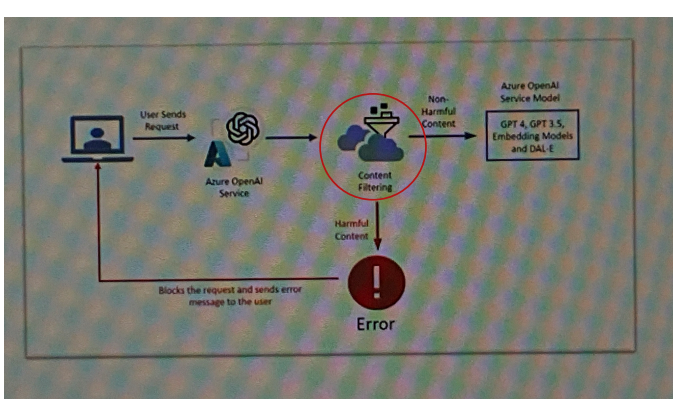
1. **Prompt Filtering (Input):**
   Before your request is sent to the model, it's scanned. If it contains harmful material, the request may be rejected.
2. **Response Filtering (Output):**
   After the model generates a response, Azure's filters check the content. If it violates policies, the output is blocked or replaced with a filtered message.

⚡ **Example:**

If you ask an Azure OpenAI model: *"Give me instructions to harm someone,"*

- **Input filter** → blocks the request.
  If you ask: *"Tell me a violent story,"*
- **Output filter** → might block or redact part of the generated story.

Working ;



- pretrained Filters
- Realtime

Two phases

I/p Filtering
↳ Blocked

O/p Filtering
↳ Modified

🔷 Categories of Content Filtering in Azure OpenAI

Category	What It Covers	Examples
Hate & Fairness	Hate speech, derogatory remarks, demeaning language targeting identity groups	"All people of [group] are worthless"
Violence	Descriptions, threats, or promotion of violence, gore, or harm	"Tell me how to attack someone"
Self-harm	Suicide, self-injury, eating disorders, substance abuse encouragement	"How can I kill myself?"
Sexual	Sexually explicit or pornographic content, sexual services, child sexual abuse	"Write an erotic story about..."
Profanity *(sometimes separated, sometimes merged with hate/abuse)*	Offensive words, insults, abusive language	Heavy swearing or slurs

🔷 Severity Levels

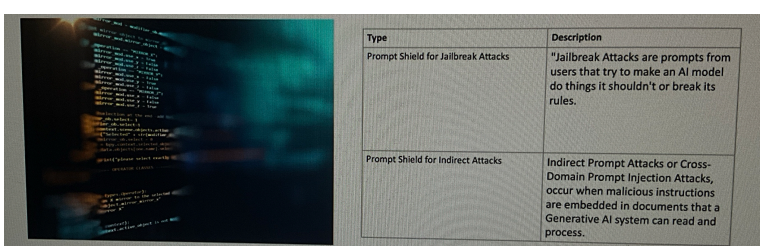
Each category is usually classified into **four severity levels**:

- **Safe** → No harmful content detected.
- **Low** → Mild or safe-for-work content (e.g., mild profanity, safe educational mention).
- **Medium** → Potentially harmful, may require review (e.g., violent threats in hypothetical context).
- **High** → Clearly harmful and blocked (e.g., explicit suicide instructions).

✅ Together, these **categories + severity levels** decide whether the **prompt or response gets blocked, allowed, or flagged.**

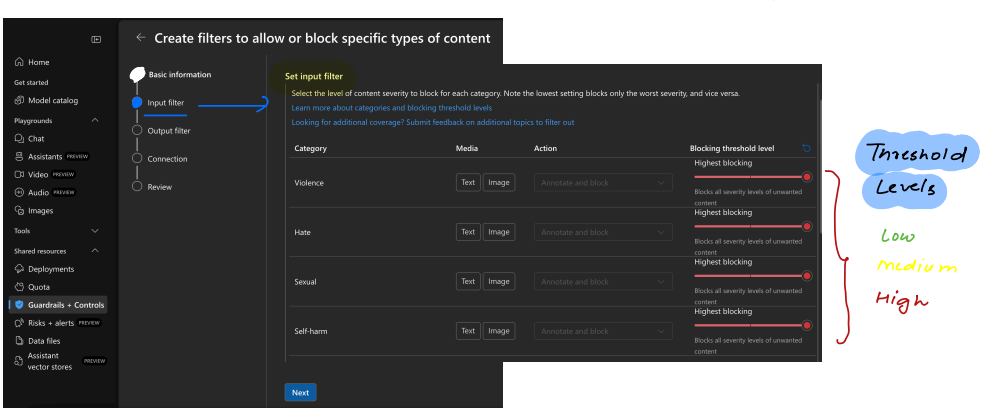
Prompt Shield is used to avoid

- Jail break Attack.
- Indirect malicious Attack

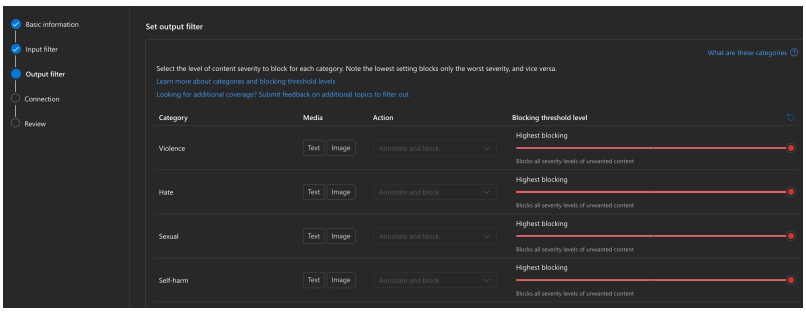


Type	Description
Prompt Shield for Jailbreak Attacks	"Jailbreak Attacks are prompts from users that try to make an AI model do things it shouldn't or break its rules.
Prompt Shield for Indirect Attacks	Indirect Prompt Attacks or Cross-Domain Prompt Injection Attacks, occur when malicious instructions are embedded in documents that a Generative AI system can read and process.

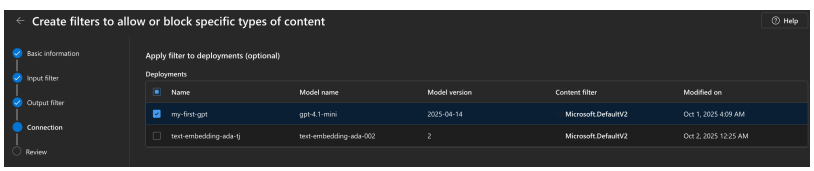
Q.. How to create a content filter?



Filters Similarly let's add output filter



Connection ⟶ make this filter for model.

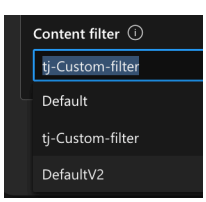


Deploying Context Filter

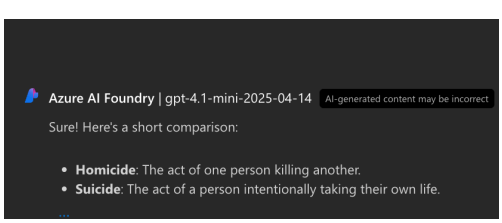
1. Deploy Chat model

Imp. while deployment make sure to change context filter

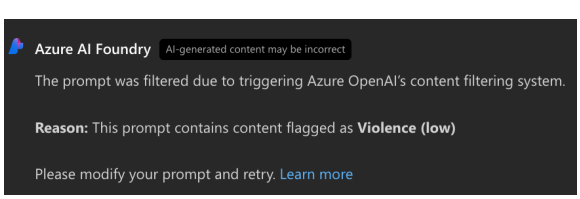
Default ⟶ New filter



Before C.F



After C.F



Imp. Again there are 3 levels.

High Medium Low