

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

- Prompt Filtering (Input):**
Before your request is sent to the model, it's scanned. If it contains harmful material, the request may be rejected.
- 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:

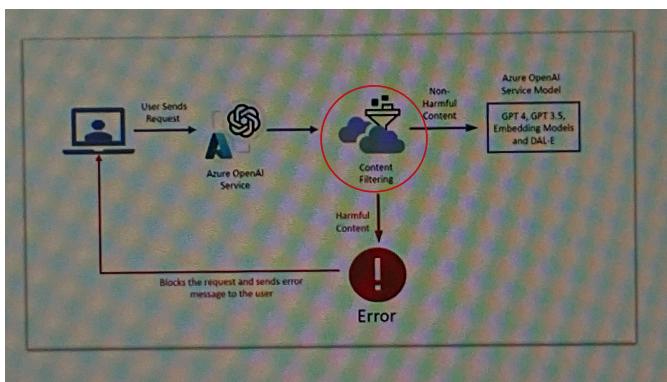
If you ask an Azure OpenAI model: "Give me instructions to harm someone,"

- 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?

The interface shows a sidebar with 'Guardrails + Controls' selected. The main area has a title 'Create filters to allow or block specific types of content'. On the left, there's a tree view with 'Input filter' selected. The right panel shows a table for 'Set input filter' with rows for Violence, Hate, Sexual, and Self-harm. Each row has columns for Category, Media (Text, Image), Action (Annotate and block), and Blocking threshold level (with a slider). A blue callout bubble labeled 'Threshold Levels' points to the sliders, with 'Low' (green), 'Medium' (yellow), and 'High' (red) color-coded arrows pointing to the respective slider positions.

Filters Similarly let's add output filter

Category	Media	Action	Blocking threshold level
Violence	Text Image	Annotate and block	Highest blocking Blocks all severity levels of unwanted content
Hate	Text Image	Annotate and block	Highest blocking Blocks all severity levels of unwanted content
Sexual	Text Image	Annotate and block	Highest blocking Blocks all severity levels of unwanted content
Self-harm	Text Image	Annotate and block	Highest blocking Blocks all severity levels of unwanted content

Connection → make this filter for model.

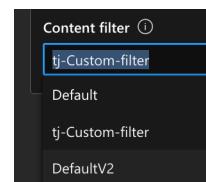
Name	Model name	Model version	Content filter	Modified on
my-first-gpt	gpt-4.1-mini	2025-04-14	Microsoft.DefaultV2	Oct 1, 2025 4:09 AM
test-embedding-ada-tj	test-embedding-ada-002	2	Microsoft.DefaultV2	Oct 2, 2025 12:25 AM

Deploying Context Filter

1. Deploy Chat mode!

Imp. while deployment make sure to change context filter

Default → New filter



Before C.F.

After C.F.

Azure AI Foundry | gpt-4.1-mini-2025-04-14

Sure! Here's a short comparison:

- **Homicide:** The act of one person killing another.
- **Suicide:** The act of a person intentionally taking their own life.

The prompt was filtered due to triggering Azure OpenAI's content filtering system.
Reason: This prompt contains content flagged as **Violence (low)**.
Please modify your prompt and retry. [Learn more](#)

Azure AI Foundry | gpt-4.1-mini-2025-04-14

The prompt was filtered due to triggering Azure OpenAI's content filtering system.
Reason: This prompt contains content flagged as **Violence (low)**.
Please modify your prompt and retry. [Learn more](#)

Imp. Again there are 3 levels.

High medium low