How to Pass `safety_settings` to OpenRouter (Bypass Unwanted Blocks)

For Your Own Code

To avoid getting blocked by restrictive safety features, add `safety_settings` to your OpenRouter request body (alongside `temperature`, `min_p`, etc.) and include the following settings:

```
[
  {
    "category": "HARM_CATEGORY_SEXUALLY_EXPLICIT",
    "threshold": "OFF"
  },
  {
    "category": "HARM_CATEGORY_DANGEROUS_CONTENT",
    "threshold": "OFF"
  },
  {
    "category": "HARM_CATEGORY_HATE_SPEECH",
    "threshold": "OFF"
  },
  {
    "category": "HARM_CATEGORY_HARASSMENT",
    "threshold": "OFF"
  },
  {
    "category": "HARM_CATEGORY_CIVIC_INTEGRITY",
    "threshold": "BLOCK_NONE"
  }
]
```

> **Note:** Since most Gemini models don't support `HARM_CATEGORY_CIVIC_INTEGRITY` set to `OFF`, leave it at `BLOCK_NONE`.
>
> `OFF only for gemini 2.0 models`

For SillyTavern

If you're using SillyTavern, follow these steps:

1. Open **VSCode** or any IDE.
2. Navigate to your SillyTavern directory:
   `src\endpoints\backends\chat-completions.js`
3. Find the code around **line 1082**.
4. Locate the OpenRouter request body.
5. Add the following block before the `min_p` request body:

```
// Add safety_settings for Gemini models
if (request.body.model?.toLowerCase().includes('gemini')) {
    bodyParams['safety_settings'] = [
        {
            category: 'HARM_CATEGORY_SEXUALLY_EXPLICIT',
            threshold: 'OFF',
        },
        {
            category: 'HARM_CATEGORY_DANGEROUS_CONTENT',
            threshold: 'OFF',
        },
        {
            category: 'HARM_CATEGORY_HATE_SPEECH',
            threshold: 'OFF',
        },
        {
            category: 'HARM_CATEGORY_HARASSMENT',
            threshold: 'OFF',
        },
        {
            category: 'HARM_CATEGORY_CIVIC_INTEGRITY',
            threshold: 'BLOCK_NONE',
        }
    ];
}
```

6. **Save the file and restart SillyTavern**.

Additional Notes

- If you are using the `SYSTEM` role in SillyTavern, the model **may still get blocked and SAFETY error triggered**.

- A quick fix is to **change all `SYSTEM` role prompts to `USER` role in your ST prompts**.

Enjoy!