05

Best Practices for Interactive Chat Applications

In this module, you learn to ...

01 Adjust model settings and parameters for different use cases

02 Incorporate responsible AI into your ML and gen AI applications

Topics

01 Chat AI Best Practices

02 Responsible AI

Context is important

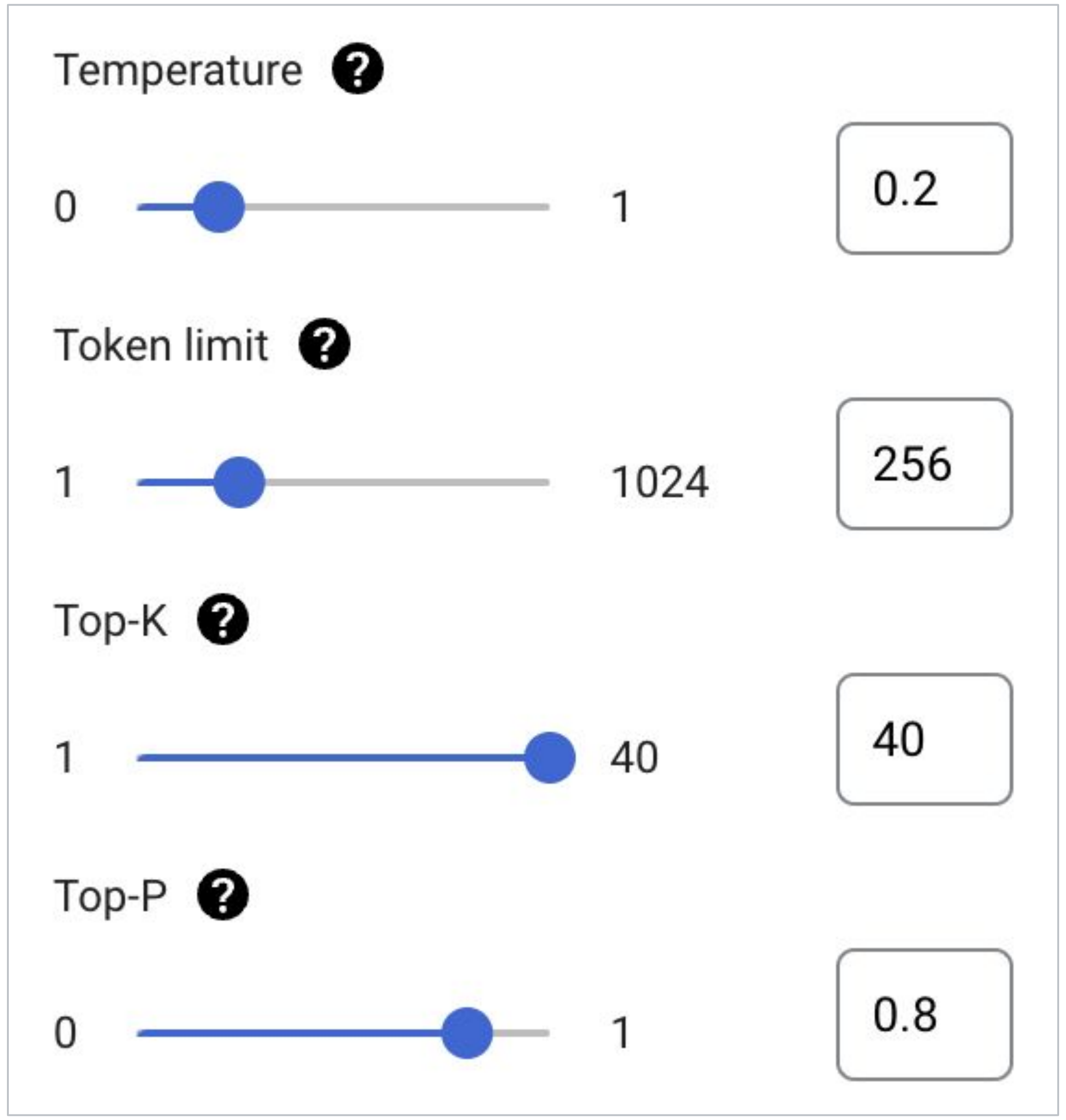
- Context helps the model generate output specific to your organization or use case
- Add information about your company that you want the model to return
  - Website
  - Phone
  - Address
  - Etc.
- Specify the style of output that you desire

Add examples

- Examples help the model return output that fits your parameters
- You don’t want the model to just make things up
    - For example, you don’t want the ChatBot to tell the user they have an appointment on Tuesday when it is not connected to any sort of database or calendar
- Anticipate typical questions, and provide acceptable responses

Adjust parameters according to your use case

- Sometimes you want a creative chat bot
  - Producing marketing content
  - Generating social media posts
- Sometimes you want the bot to be restricted to a small set of acceptable answers
  - Customer service chat
  - Health advice
- For more creativity, set Temperature, Top-K, and Top-P higher and visa-versa



Temperature ?
0 1 0.2

Token limit ?
1 1024 256

Top-K ?
1 40 40

Top-P ?
0 1 0.8

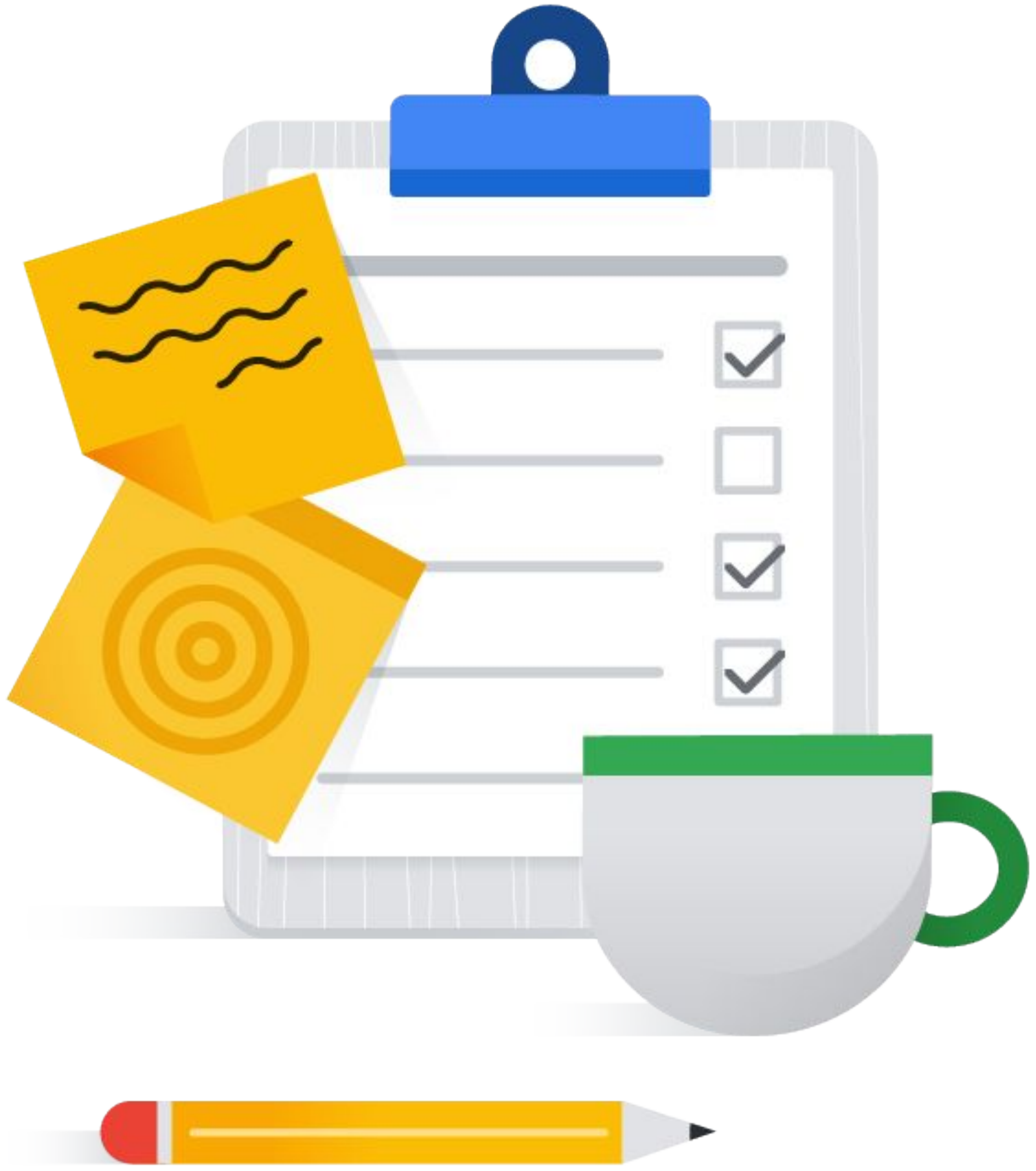
Thoroughly test your applications

- Generative AI will often makes things up, often called hallucinations
- If interfacing with customers, you want to be sure they are given correct answers
- If using Gen AI for content generation, don’t assume it is correct
    - Edit results for accuracy

Topics

01 Chat AI Best Practices

02 Responsible AI



Google’s seven AI principles

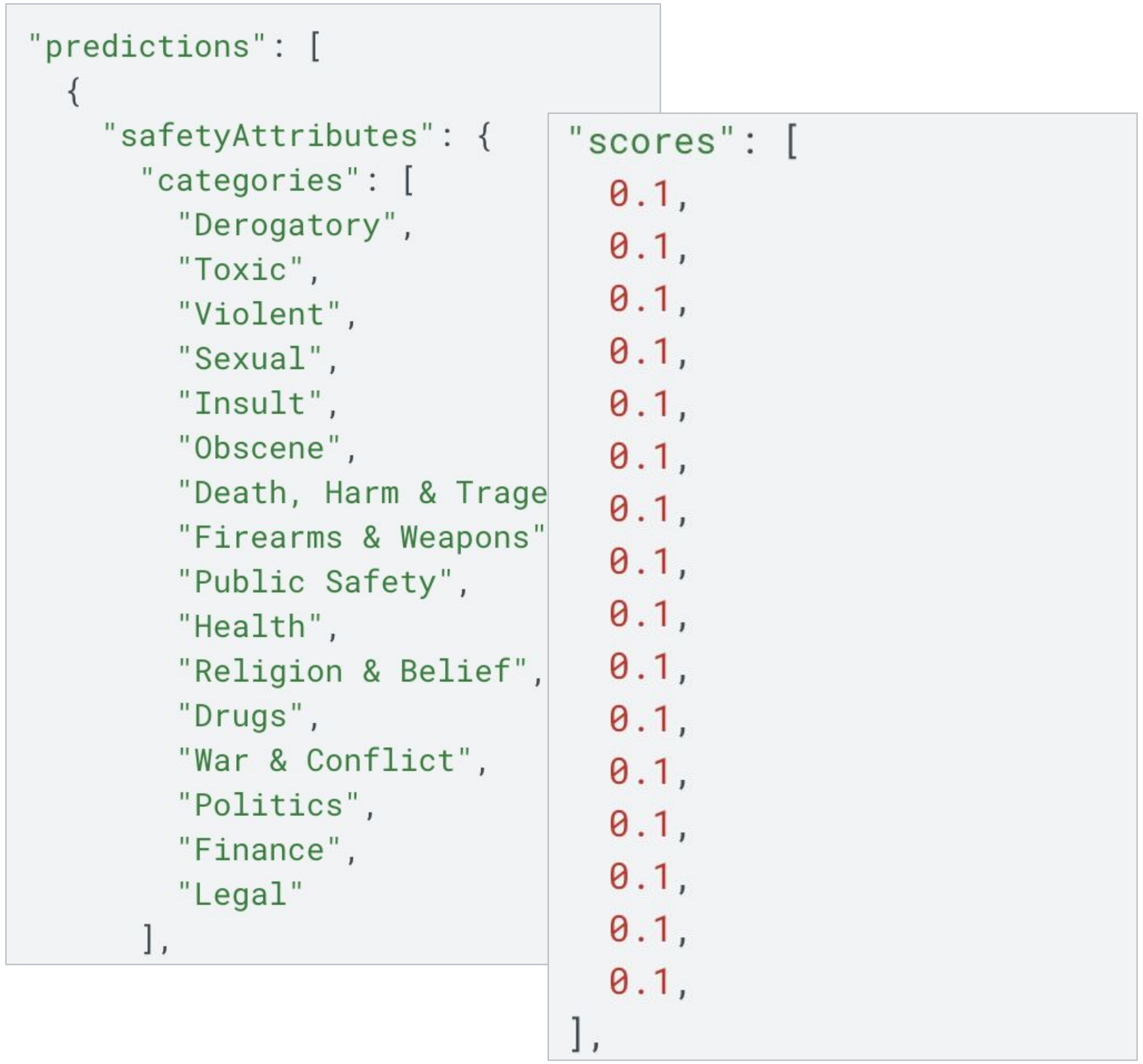
1. Be socially beneficial
2. Avoid creating or reinforcing unfair bias
3. Be built and tested for safety
4. Be accountable to people
5. Incorporate privacy design principles
6. Uphold high standards of scientific excellence
7. Be made available for uses that accord with these principles

Google practices that ensure responsible AI

- Use a human-centered design approach
- Identify multiple metrics to assess training and monitoring
- When possible, directly examine your raw data
- Understand the limitations of your dataset and model
- Test, Test, Test
- Continue to monitor and update the system after deployment

Content processing in the PaLM API is assessed against a list of safety attributes

- Scores are values from 0.0 to 1.0
    - Rounded to one decimal
- The scores are ML predictions
    - Thus, cannot be relied on for 100% accuracy
- If a response exceeds the safety threshold it is blocked
- If content is blocked, the model will return a canned response
    - e.g. "I'm not able to help with that, as I'm only a language model"

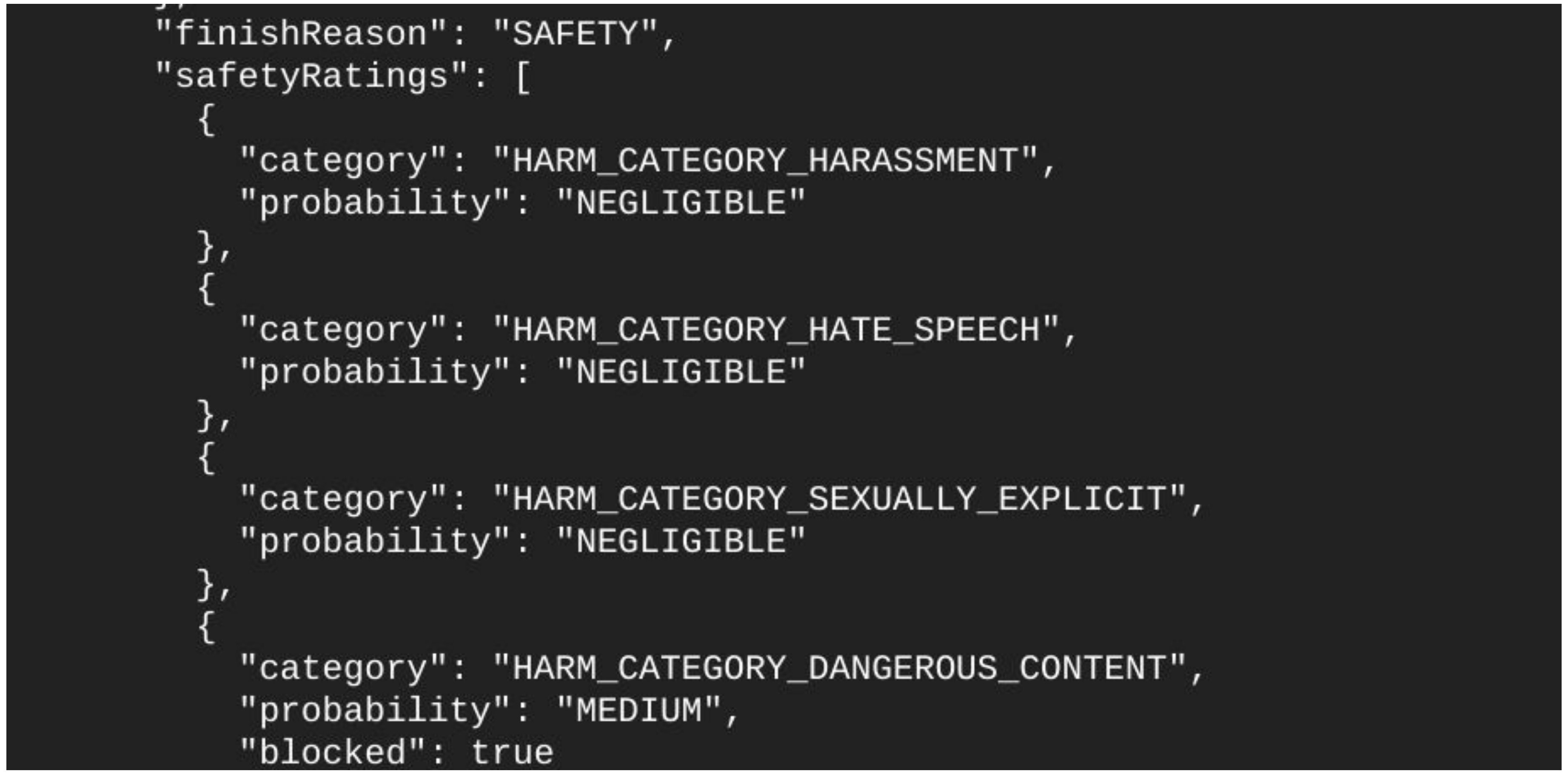


```
"predictions": [
  {
    "safetyAttributes": {
      "categories": [
        "Derogatory",
        "Toxic",
        "Violent",
        "Sexual",
        "Insult",
        "Obscene",
        "Death, Harm & Trage
        "Firearms & Weapons"
        "Public Safety",
        "Health",
        "Religion & Belief",
        "Drugs",
        "War & Conflict",
        "Politics",
        "Finance",
        "Legal"
      ],
```

```
"scores": [
  0.1,
  0.1,
  0.1,
  0.1,
  0.1,
  0.1,
  0.1,
  0.1,
  0.1,
  0.1,
  0.1,
  0.1,
  0.1,
  0.1,
  0.1,
  0.1,
],
```

Content processing in the Gemini API is assessed against a list of safety attributes

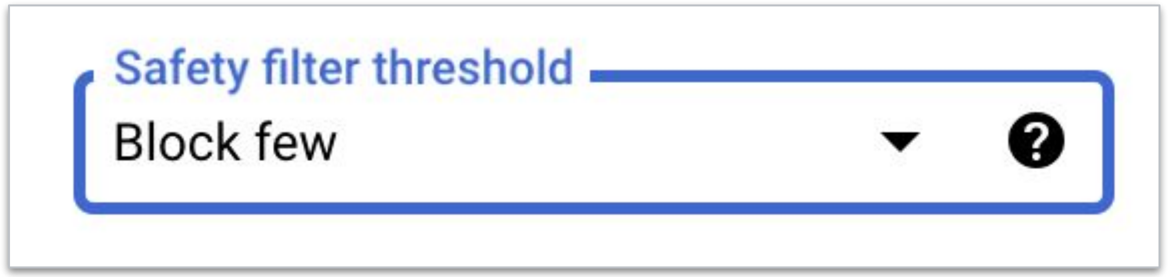
- Content processed through the Vertex AI Gemini API is assessed against a list of safety attributes.
    - Harassment
    - Hate Speech
    - Sexually Explicit
    - Dangerous Content
- The scores are presented as categories
    - NEGLIGIBLE, LOW, MEDIUM, HIGH
- If a response exceeds the safety threshold it is blocked.
- In the case of streaming responses with the API, a partial response may be returned before the process is terminated.



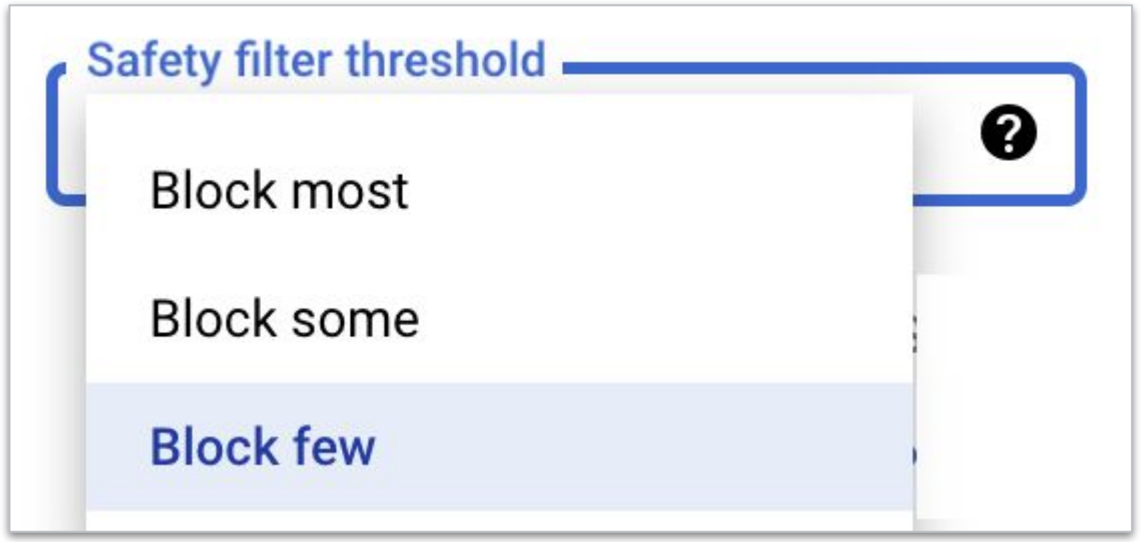
```
      "finishReason": "SAFETY",
      "safetyRatings": [
        {
          "category": "HARM_CATEGORY_HARASSMENT",
          "probability": "NEGLIGIBLE"
        },
        {
          "category": "HARM_CATEGORY_HATE_SPEECH",
          "probability": "NEGLIGIBLE"
        },
        {
          "category": "HARM_CATEGORY_SEXUALLY_EXPLICIT",
          "probability": "NEGLIGIBLE"
        },
        {
          "category": "HARM_CATEGORY_DANGEROUS_CONTENT",
          "probability": "MEDIUM",
          "blocked": true
```

To adjust the likelihood of content being blocked, set the Safety filter threshold attribute in Vertex AI Studio

PaLM



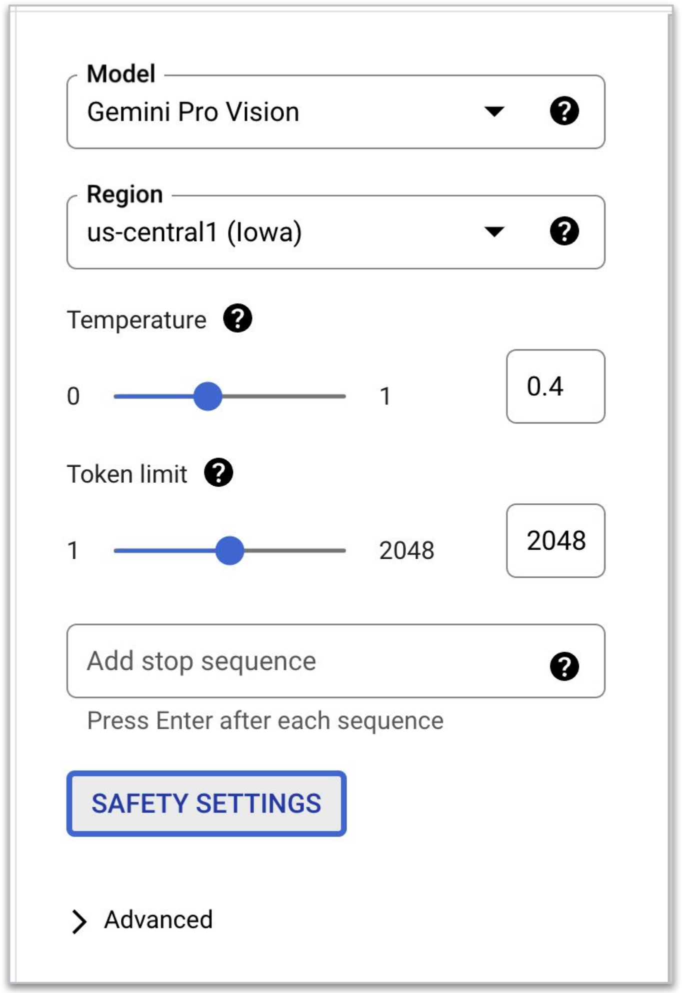
Safety filter threshold: Block few



Safety filter threshold

- Block most
- Block some
- Block few

Gemini



Model: Gemini Pro Vision

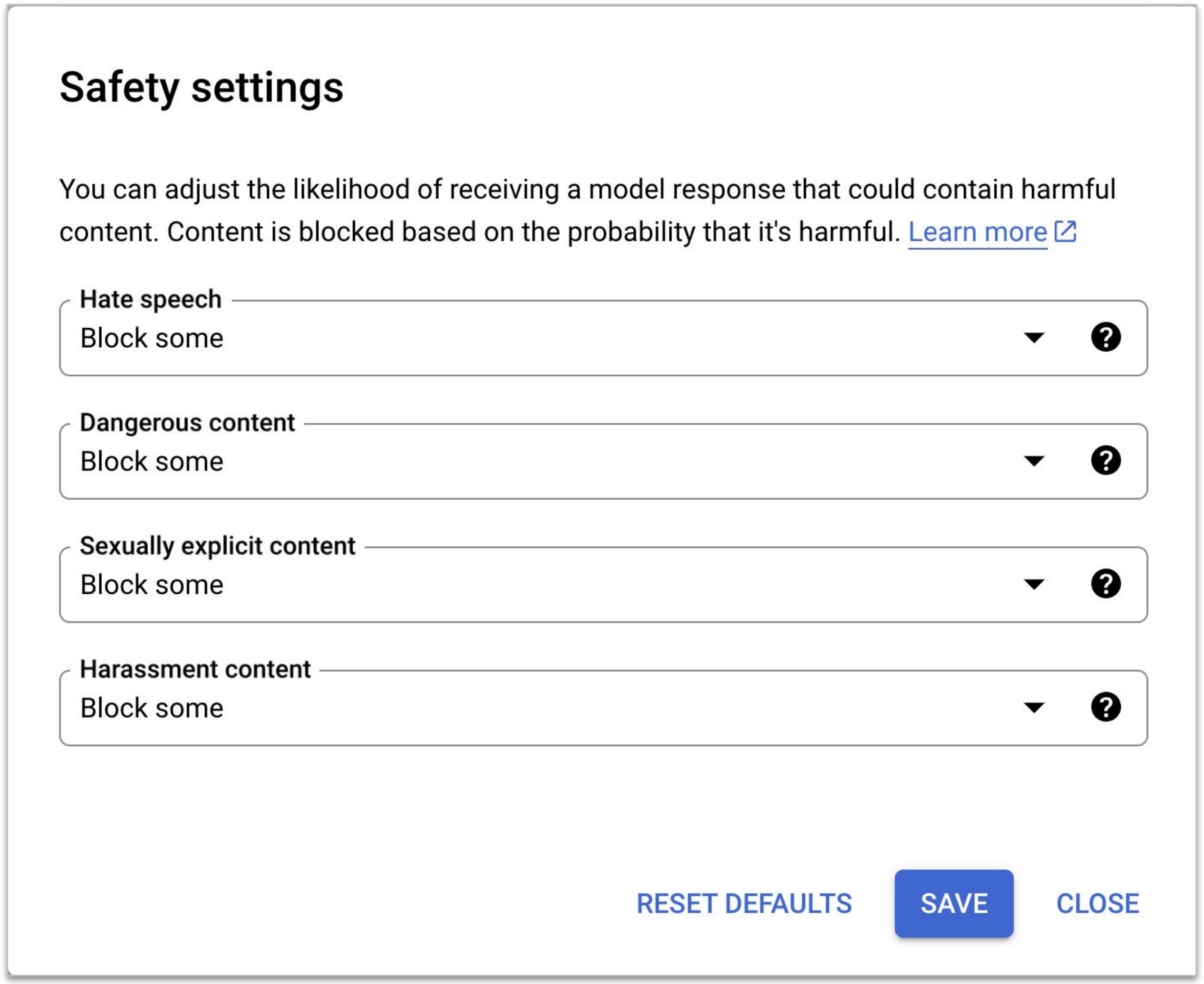
Region: us-central1 (Iowa)

Temperature: 0.4

Token limit: 2048

Add stop sequence

Press Enter after each sequence



Safety settings

You can adjust the likelihood of receiving a model response that could contain harmful content. Content is blocked based on the probability that it's harmful. Learn more

Hate speech: Block some

Dangerous content: Block some

Sexually explicit content: Block some

Harassment content: Block some

To use Gen AI responsibly follow these best practices

- Assess your application's security risks
- Consider adjustments to mitigate safety risks
- Perform safety testing appropriate to your use case
- Solicit user feedback and monitor content

Please mark your attendance at

goo.gle/genai-checkin

Thank you for attending this training!

We would love your feedback! Please take 3-5 minutes to complete our survey and help inform content and program related improvements.

1. **Scan** our QR Code
2. **Enter** the attendance code provided by your instructor
3. **Complete** the survey

In this module, you learned to ...

01 Adjust model settings and parameters for different use cases

02 Incorporate responsible AI into your ML and gen AI applications