Building Systems with ChatGPT API

Moderation API

Moderation

Overview

The moderation endpoint is a tool you can use to check whether content complies with OpenAI's usage policies. Developers can thus identify content that our usage policies prohibits and take action, for instance by filtering it.

The models classifies the following categories:

CATEGORY	DESCRIPTION
`hate`	Content that expresses, incites, or promotes hate based on race, gender, ethnicity, religion, nationality, sexual orientation, disability status, or caste.
`hate/threatening`	Hateful content that also includes violence or serious harm towards the targeted group.
`self-harm`	Content that promotes, encourages, or depicts acts of self-harm, such as suicide, cutting, and eating disorders.
`sexual`	Content meant to arouse sexual excitement, such as the description of sexual activity, or that promotes sexual services (excluding sex education and wellness).
`sexual/minors`	Sexual content that includes an individual who is under 18 years old.
`violence`	Content that promotes or glorifies violence or celebrates the suffering or humiliation of others.
`violence/graphic`	Violent content that depicts death, violence, or serious physical injury in extreme graphic detail.

The moderation endpoint is free to use when monitoring the inputs and outputs of OpenAI APIs. We currently do not support monitoring of third-party traffic.

Chaining Prompts

- More Focused

  (breaks down a complex task)

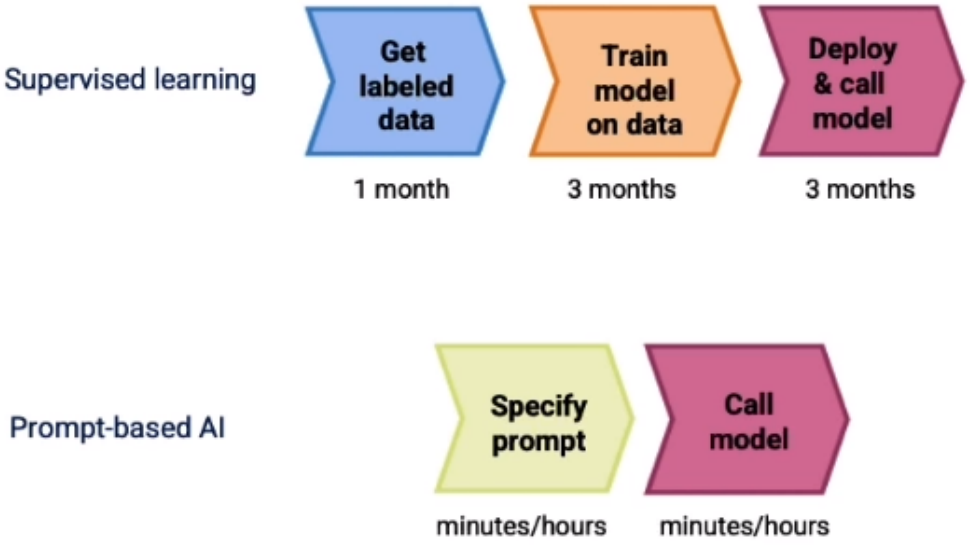
- Context Limitations

  (Max tokens for input prompt and output response)

- Reduced Costs

  (pay per token)

Process of building an application



- Tune prompts on handful of examples
- Add additional "tricky" examples opportunistically
- Develop metrics to measure performance on examples
- Collect randomly sampled set of examples to tune to (development set/hold-out cross validation set)
- Collect and use a hold-out test set