# Optimizing API Usage

**Esteban Herrera**
Author

@eh3rrera | eherrera.net

# Rate limiting

**Technique for limiting the number of requests that an API will accept within a certain time frame.**

# OpenAI Rate Limiting Policy

Requests Per Minute (RPM)

Requests Per Day (RPD)

Tokens Per Minute (TPM)

Tokens Per Day (TPD)

Images Per Minute (IPM)

# The Moderations API

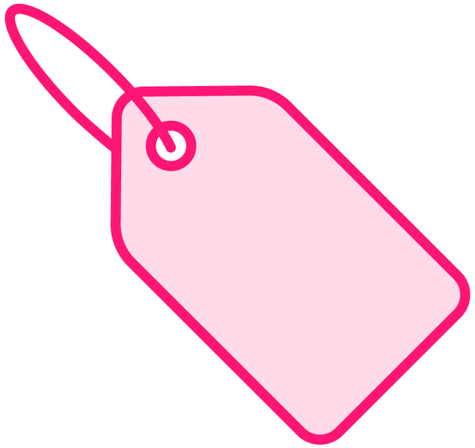
# Counting Tokens

# Benefits of Counting Tokens before Sending the Request

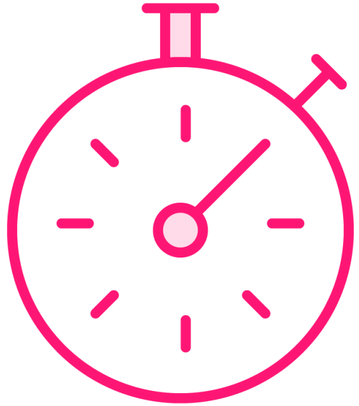
- Anticipate costs
- Effective text partitioning
- Improve efficiency
- Prevent errors

# Choosing a Model

# Two Most Important Factors

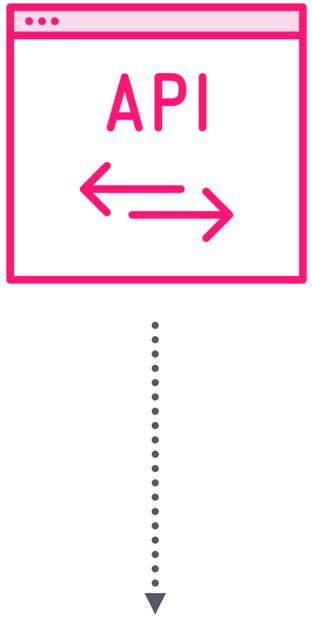


Cost



Processing time

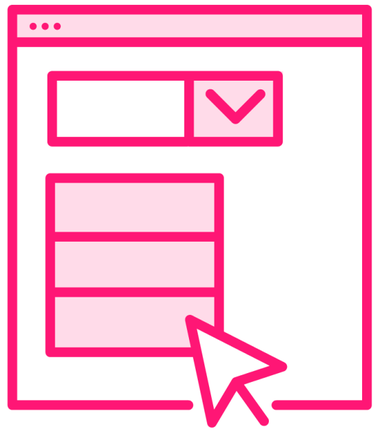
# Streaming a Response



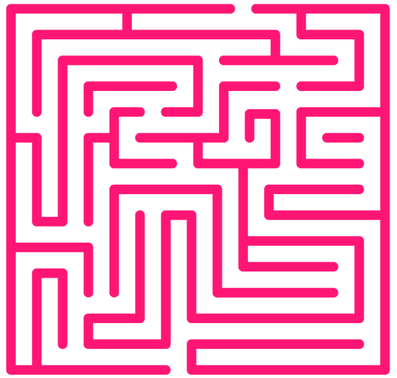

How can I help you?

# Streaming Trade-offs

Coherence and quality



User experience

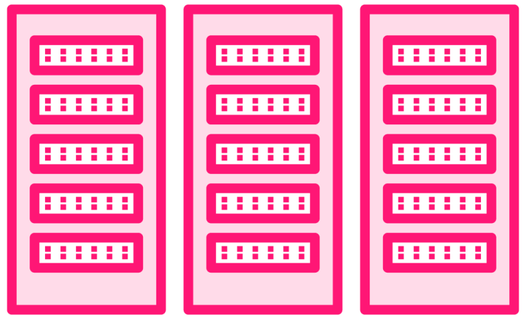


Implementation complexity

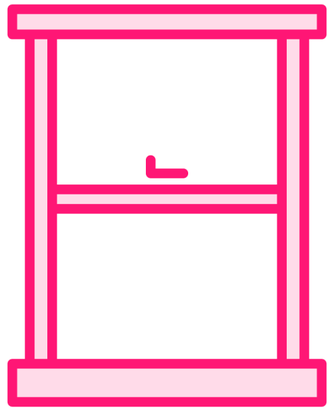
# Selecting the right GPT model is a balancing act.

# More Factors to Consider

Accuracy

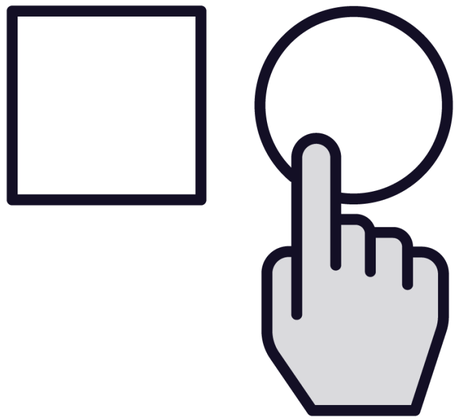


Capacity and scalability



Context window length

# Tips for Choosing a Model



Define clear objectives

Understand your user base

Consider the data

Leverage existing research

Explore different vendors

Begin with a less expensive, faster model

# Summary

## Concepts

- Artificial Intelligence (AI)
- Machine Learning (ML)
- Deep Learning (DL)
- Generative Pretrained Transformers (GPT)
- Prompt
- Token
- Context window

# Summary

## OpenAI

- Offers access to its GPT models through a REST API
  - HTTP for communication
  - JSON for data exchange
  - API key for authorization

# Summary

## OpenAI API endpoints

- Text completions
- Image generation
- Speech-to-text
- Moderation models
- And others...

# Summary

**OpenAI's official Python library**

- Generating images
- Transcribing audio
- Chat Completions API

# Summary

**Job interview demo**

- Using the Playground for experimentation
- Multi-step prompts
- Function calling
- Error handling
- Rate limiting
- Moderations API
- Counting tokens

# Summary

**Factors for selecting a model**

- Cost
- Processing time
- Accuracy
- User needs
- Data availability

**An iterative, strategic approach can help in identifying the most suitable model for your application**

# Thank you.