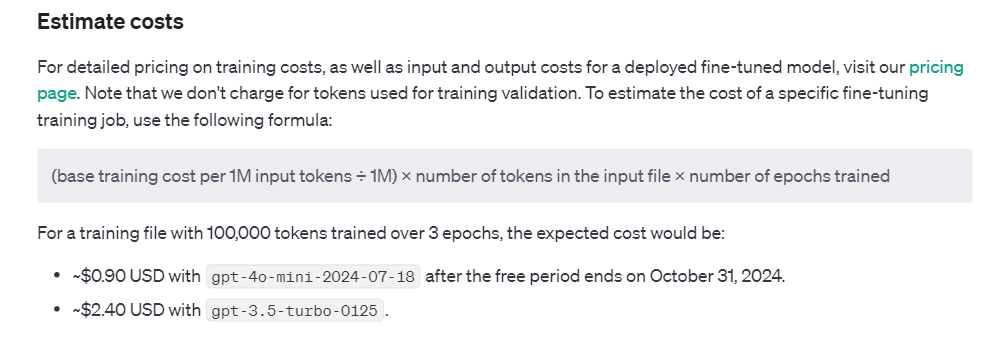
**COMPLETE GUIDE ON GPT-3.5**

GPT VERSION:

1. GPT-4o-mini
2. GPT-4o-mini-2024-07-18
3. GPT-3.5-TURBO
4. GPT-3.5-TURBO-16K
5. GPT=3.5-TURBO-1106
6. GPT-3.5-TURBO-0.125

Fine-Tuning:



Fine-tuning is available

gpt-4o-2024-08-06

gpt-4o-mini-2024-07-18

gpt-4-0613

gpt-3.5-turbo-0125

gpt-3.5-turbo-1106

gpt-3.5-turbo-0613

babbage-002

davinci-002

Fine-tuning guide link

<https://platform.openai.com/docs/guides/fine-tuning>

Parameters:

* Fine-tuning stage
* inference stage (generation stage)

1. Temperature (inference stage):

You can think of temperature like randomness, with 0 being least random (or most deterministic) and 2 being most random (least deterministic). When using low values for temperature (e.g. 0.2) the model responses will tend to be more consistent but may feel more robotic. Values higher than 1.0, especially values close to 2.0, can lead to erratic model outputs. If your goal is creative outputs, a combination of a slightly higher than normal temperature (e.g. 1.2) combined with a prompt specifically asking the model to be creative may be your best bet, but we encourage experimentation.

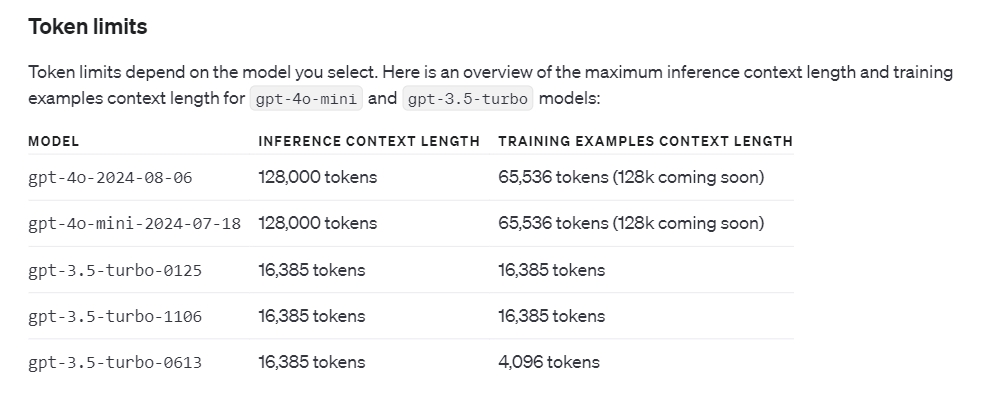
0 🡺 least randomness ( or most deterministic)

2 🡺 most randomness (least deterministic )

* 1. 🡺 normal temperature with prompt

0.2 🡺 more consistent but response feel robotic

1. Token Length (inference stage):



1. frequency penalty (inference stage):

The frequency penalty parameter in GPT is used to control how often the model repeats the same tokens or words during generation.

Purpose: The frequency penalty helps reduce repetition in the model’s output. Without it, GPT may sometimes repeat words, phrases, or ideas, especially in longer outputs.

How It Works: The frequency penalty adjusts the probability of repeating a token. If the frequency penalty is applied, tokens that have appeared more frequently in the generated text are penalized, making them less likely to be chosen again.

Frequency penalty range (0-9)

Where **0** means allows to more repetition and

Where **1** means aggressively avoid the choose duplicates

1. presence penalty

**Purpose**: The presence penalty encourages the model to introduce new ideas or words into the generated text. It makes the model less likely to repeat words that have already been used, even if they’ve only been used once.

 If a word (token) has already been used in the text, applying a presence penalty makes it less likely to be reused, encouraging the model to explore new words and concepts.

 It doesn't focus on how many times a token has been used (like the frequency penalty), but on whether it has been used at all.

**Parameter Range**:

* A **higher presence penalty** (e.g., 1 or more) discourages the model from repeating words and pushes it to generate more novel content.
* A **lower presence penalty** (e.g., 0) has no effect, allowing words or tokens to be repeated without discouragement.
*  **High presence penalty**: Leads to more diverse and creative text generation, but it may cause the model to avoid repeating important words, which can make the output less cohesive.
*  **Low presence penalty**: Allows repetition, which can be helpful in contexts where maintaining certain themes or key terms is important, but can also lead to repetitive or redundant outputs.

Training Dataset Example Format

* If you're working with **GPT-4o-mini or GPT-3.5-turbo**, stick with the conversational chat format.
* If you're fine-tuning **Babbage-002** or **Davinci-002**, use the prompt-completion pair format.