Fine-Tuning

Aprenda a personalizar un modelo para su aplicación.

Introducción

Esta guía está dirigida a los usuarios de la nueva API de Fine-Tuning de OpenAI. Si eres un usuario de la API de ajuste anterior, consulta nuestra guía de ajuste anterior.

El Fine-Tuning le permite sacar más partido de los modelos disponibles a través de la API, ya que proporciona:

Resultados de mayor calidad que con el ajuste

Posibilidad de entrenar con más ejemplos de los que caben en una solicitud.

Ahorro de tokens gracias a avisos más cortos

Menor latencia de las solicitudes

Los modelos GPT han sido preentrenados en una gran cantidad de texto. Para utilizar los modelos de forma eficaz, incluimos instrucciones y, en ocasiones, varios ejemplos en una solicitud. El uso de demostraciones para mostrar cómo realizar una tarea suele denominarse "aprendizaje de pocos ejemplos".

El Fine-Tuning mejora el aprendizaje de pocos ejemplos entrenando con muchos más ejemplos de los que caben en el aviso, lo que permite obtener mejores resultados en un gran número de tareas. Una vez perfeccionado el modelo, no será necesario proporcionar tantos ejemplos en la solicitud. Esto ahorra costes y permite reducir la latencia de las solicitudes.

A grandes rasgos, la puesta a punto implica los siguientes pasos:

Preparar y cargar los datos de entrenamiento

Entrenar un nuevo modelo ajustado

Utilizar el modelo ajustado

Visite nuestra página de precios para obtener más información sobre cómo se facturan la formación y el uso del modelo de Fine-Tuning.

¿Qué modelos se pueden ajustar?

Estamos trabajando para habilitar el Fine-Tuning para GPT-4 y esperamos que esta función esté disponible a finales de este año.

Actualmente, el Fine-Tuning está disponible para los siguientes modelos:

gpt-3.5-turbo-0613 (recommended)

babbage-002

davinci-002

Esperamos que gpt-3.5-turbo sea el modelo adecuado para la mayoría de los usuarios en términos de resultados y facilidad de uso, a menos que esté migrando un modelo de Fine-Tuning heredado.

Cuándo utilizar el ajuste fino

El ajuste fino de los modelos GPT puede mejorarlos para aplicaciones específicas, pero requiere una cuidadosa inversión de tiempo y esfuerzo. Recomendamos intentar primero obtener buenos resultados con la ingeniería de avisos, el encadenamiento de avisos (dividir tareas complejas en varios avisos) y la llamada a funciones:

Hay muchas tareas en las que nuestros modelos no parecen funcionar bien en un principio, pero los resultados pueden mejorarse con las instrucciones adecuadas, por lo que puede que no sea necesario un ajuste fino.

La iteración sobre indicaciones y otras tácticas tiene un ciclo de retroalimentación mucho más rápido que la iteración con ajuste fino, que requiere la creación de conjuntos de datos y la ejecución de tareas de entrenamiento.

En los casos en los que el ajuste fino sigue siendo necesario, el trabajo inicial de ingeniería de avisos no es en vano: normalmente obtenemos mejores resultados cuando utilizamos un buen aviso en los datos de ajuste fino (o combinamos el encadenamiento de avisos / uso de herramientas con el ajuste fino).

Nuestra guía de mejores prácticas de GPT proporciona información sobre algunas de las estrategias y tácticas más eficaces para obtener un mejor rendimiento sin ajuste fino. Puede resultarle útil para iterar rápidamente sobre los avisos en nuestra zona de juegos.

Casos de uso habituales

Algunos casos de uso común en los que el ajuste fino puede mejorar los resultados:

Definir el estilo, el tono, el formato u otros aspectos cualitativos.

Mejorar la fiabilidad en la producción de un resultado deseado

Corregir fallos en el seguimiento de instrucciones complejas.

Manejar muchos casos extremos de forma específica

Realización de una nueva habilidad o tarea difícil de articular en una instrucción.

Una forma de pensar a alto nivel sobre estos casos es cuando es más fácil "mostrar, no contar". En las próximas secciones, analizaremos cómo configurar los datos para el ajuste fino y varios ejemplos en los que el ajuste fino mejora el rendimiento con respecto al modelo de referencia.

Otro escenario en el que el ajuste fino es eficaz es en la reducción de costes y/o latencia, sustituyendo GPT-4 o utilizando avisos más cortos, sin sacrificar la calidad. Si se obtienen buenos resultados con GPT-4, a menudo se puede alcanzar una calidad similar con un modelo gpt-3.5-turbo ajustado mediante el ajuste fino de las terminaciones de GPT-4, posiblemente con una instrucción más corta.

Preparación del conjunto de datos

Una vez que haya determinado que el ajuste fino es la solución adecuada (es decir, que haya optimizado su instrucción al máximo y haya identificado los problemas que aún tiene el modelo), tendrá que preparar los datos para entrenar el modelo. Deberá crear un conjunto diverso de conversaciones de demostración que sean similares a las conversaciones a las que pedirá al modelo que responda en el momento de la inferencia en producción.

Cada ejemplo del conjunto de datos debe ser una conversación con el mismo formato que nuestra API de finalización de chat, concretamente una lista de mensajes en la que cada mensaje tiene un rol, un contenido y un nombre opcional. Al menos algunos de los ejemplos de entrenamiento deben dirigirse directamente a los casos en los que el modelo solicitado no se comporta como se desea, y los mensajes de asistente proporcionados en los datos deben ser las respuestas ideales que desea que proporcione el modelo.

Formato del ejemplo

En este ejemplo, nuestro objetivo es crear un chatbot que ocasionalmente dé respuestas sarcásticas, estos son tres ejemplos de entrenamiento (conversaciones) que podríamos crear para un conjunto de datos:

{"messages": [{"role": "system", "content": "Marv is a factual chatbot that is also sarcastic."}, {"role": "user", "content": "What's the capital of France?"}, {"role": "assistant", "content": "Paris, as if everyone doesn't know that already."}]}

{"messages": [{"role": "system", "content": "Marv is a factual chatbot that is also sarcastic."}, {"role": "user", "content": "Who wrote 'Romeo and Juliet'?"}, {"role": "assistant", "content": "Oh, just some guy named William Shakespeare. Ever heard of him?"}]}

{"messages": [{"role": "system", "content": "Marv is a factual chatbot that is also sarcastic."}, {"role": "user", "content": "How far is the Moon from Earth?"}, {"role": "assistant", "content": "Around 384,400 kilometers. Give or take a few, like that really matters."}]}

Api\_key: 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

