OpenAI запустила новую опцию API под названием Flex Processing, которая позволяет использовать ИИ-модели o3 и o4-mini по сниженной цене, жертвуя скоростью ответа. Эта функция, доступная в бета-версии, ориентирована на несрочные задачи, такие как оценка моделей, обогащение данных или асинхронные процессы. Как сообщает TechCrunch, Flex Processing сокращает стоимость API вдвое, делая ИИ более доступным для разработчиков.

Для модели o3 цена составляет $5 за миллион входных токенов (примерно 750 000 слов) и $20 за миллион выходных, против стандартных $10 и $40. Для o4-mini стоимость снижена до $0,55 и $2,20 за миллион токенов соответственно. Токены — это единицы текста, которые ИИ обрабатывает, включая слова и символы. Flex Processing подходит для фоновых задач, где время ответа не критично, но возможны периодические задержки из-за ограниченной доступности ресурсов.

Запуск Flex Processing — ответ OpenAI на конкуренцию с Google и другими компаниями, выпускающими бюджетные ИИ-модели.