Google запустила Gemini 2.5 Flash, новую модель искусственного интеллекта, доступную в бета-версии через приложение Gemini, API, Google AI Studio и Vertex AI. Эта модель, отличающаяся скоростью и эффективностью, позволяет разработчикам управлять процессом «мышления» ИИ.

Gemini 2.5 Flash — это компактная версия, превосходящая Gemini 2.0 Flash по производительности, но сохраняющая низкую стоимость. Разработчики могут задавать «бюджет мышления» — количество токенов (единиц текста, вроде слов или символов), выделяемых на обдумывание ответа. Например, цена составляет $0,15 за миллион входных токенов, а выходные — $0,60 без мышления или $3,50 с ним.

Это позволяет балансировать между скоростью, качеством и затратами, что идеально для приложений, требующих обработки больших объемов данных, например, чат-ботов.

Модель поддерживает функцию Canvas для работы с текстом и кодом, а в будущем добавит глубокое изучение (Deep Research). В приложении Gemini она заменила экспериментальную версию 2.0 Thinking, но пока тоже помечена как «экспериментальная». По словам представителя Google Тулси Доши, запуск в бета-версии поможет собрать отзывы для улучшения модели.