

Google tung Gemini 3.8 Flash, AI mạnh hơn nhưng vẫn giữ giá thấp

14:02 03/09/2026

Google vừa ra mắt Gemini 3.8 Flash và phiên bản chuyên an ninh mạng Gemini 3.8 Flash Cyber, hướng tới các tác vụ suy luận phức tạp, lập trình dài hạn và vận hành AI agent. Đáng chú ý, Gemini 3.8 Flash được Google định giá từ 0,75 USD/1 triệu token đầu vào và 3,75 USD/1 triệu token đầu ra, tương đương mức giá giới thiệu của thế hệ trước.

Không chỉ nhanh, Gemini 3.8 Flash còn được thiết kế để “làm việc” nhiều hơn

Google vừa giới thiệu Gemini 3.8 Flash, mẫu AI mới nhất trong dòng Flash, đồng thời ra mắt Gemini 3.8 Flash Cyber dành riêng cho lĩnh vực an ninh mạng. Theo Google, Gemini 3.8 là mẫu reasoning và coding mạnh nhất của hãng tại thời điểm ra mắt, nhưng vẫn duy trì tốc độ và mức chi phí thấp của dòng Flash.

Điểm đáng chú ý nhất nằm ở chiến lược giá. Gemini 3.8 Flash hiện có mức giá giới thiệu 0,75 USD cho mỗi 1 triệu token đầu vào và 3,75 USD cho mỗi 1 triệu token đầu ra. Mức giá này được áp dụng đến hết ngày 31/12/2026; từ ngày 1/1/2027, giá dự kiến tăng lên lần lượt 1,5 USD và 7,5 USD cho mỗi 1 triệu token.

Với các doanh nghiệp và nhà phát triển sử dụng API ở quy mô lớn, mức giá thấp có thể tạo ra lợi thế đáng kể khi triển khai những hệ thống AI phải xử lý khối lượng yêu cầu liên tục.

Google cho biết Gemini 3.8 Flash được xây dựng cho những tác vụ đòi hỏi nhiều bước suy luận, lập trình dài hạn và AI agent có khả năng tự thực hiện chuỗi nhiệm vụ.

Trên benchmark DeepSWE v1.1 dành cho kỹ thuật phần mềm dài hạn, Google cho biết Gemini 3.8 Flash có thể tự giải quyết các vấn đề kỹ thuật phức tạp theo quy trình từ đầu đến cuối và vượt qua phần lớn các mô hình frontier có quy mô lớn hơn, trong khi chi phí thấp hơn đáng kể.



Gemini 3.8 Flash ra mắt (ảnh: nguồn Internet)

Một điểm khác biệt trong cách vận hành của Gemini 3.8 Flash là mô hình không nhất thiết phải tìm cách đưa ra câu trả lời nhanh nhất. Với những nhiệm vụ phức tạp, model có thể thực hiện thêm các bước suy luận, gọi công cụ nhiều lần và sử dụng thêm token nhằm cải thiện chất lượng kết quả.

Google gọi đây là cách tiếp cận để Gemini 3.8 Flash “làm việc chăm chỉ hơn”. Người phát triển cũng có thể hạ mức độ suy luận trong những trường hợp ưu tiên tốc độ và tiết kiệm chi phí.

Theo Google, Gemini 3.8 Flash đạt 54,9% trên HLE-Verified, một bài đánh giá khả năng xử lý các vấn đề cần nhiều bước suy luận trong nhiều lĩnh vực, đồng thời cải thiện kết quả so với Gemini 3.7 Flash ở các bài đánh giá dành cho tác vụ tài chính, pháp lý và AI agent.

Google đưa AI vào những công việc phức tạp hơn

Sự thay đổi đáng chú ý của Gemini 3.8 Flash nằm ở việc Google không còn định vị Flash đơn thuần là một dòng model “nhẹ, nhanh và rẻ”.

Trong các màn trình diễn do Google công bố, Gemini 3.8 Flash có thể xây dựng một trò chơi 3D thông qua prompt, tạo phiên bản Google Maps chạy trên DOS và phát triển các công cụ trực quan hóa dữ liệu khoa học. Những ví dụ này cho thấy Google đang hướng dòng Flash tới các AI agent có khả năng thực hiện một chuỗi công việc thay vì chỉ trả lời câu hỏi.

Điều này đặc biệt quan trọng với doanh nghiệp. Thay vì sử dụng AI cho từng thao tác riêng lẻ, các tổ chức có thể triển khai agent để phân tích dữ liệu, viết và kiểm tra mã nguồn, sử dụng công cụ bên ngoài rồi tiếp tục xử lý kết quả.

Gemini 3.8 Flash Cyber: tập trung vào phát hiện và vá lỗ hổng

Song song với Gemini 3.8 Flash, Google cũng giới thiệu Gemini 3.8 Flash Cyber, phiên bản được thiết kế chuyên biệt cho an ninh mạng.

Model này hiện được cung cấp cho các bên phòng thủ đáng tin cậy thông qua Fairwind Program, thay vì mở rộng đại trà. Google cho biết Gemini 3.8 Flash Cyber tập trung vào khả năng tự động phát hiện và sửa các lỗ hổng phần mềm, ưu tiên mục đích phòng thủ thay vì khai thác lỗ hổng.

Trên một benchmark nội bộ bao phủ các codebase sử dụng 20 ngôn ngữ lập trình, Google cho biết Gemini 3.8 Flash Cyber đạt tỷ lệ thành công trên 70% trong việc phát hiện nhiều loại lỗ hổng. Trên CWE-Bench, model đạt 47,2% pass@1, gần với mức 47,8% của một frontier model dẫn đầu nhưng có chi phí thấp hơn đáng kể.

Google cũng đã sử dụng Gemini 3.8 Flash Cyber để bảo vệ chính hệ thống của mình. Đội ngũ bảo mật Chrome cho biết model tạo ra số bản vá chính xác cao gấp 2,6 lần so với các commercial model tốt nhất có quy mô lớn hơn.

Trong một trường hợp khác, nhóm Cloud Vulnerability Research của Google sử dụng model để phát hiện một lỗ hổng nền tảng nghiêm trọng trong chưa đầy 2 giờ, trong khi quá trình nghiên cứu và phát hiện dạng lỗ hổng này thông thường có thể mất nhiều tháng.

“Flash” đang trở thành dòng AI chủ lực giá rẻ của Google

Việc Gemini 3.8 Flash xuất hiện chỉ khoảng ba tuần sau Gemini 3.7 Flash và là lần phát hành Flash thứ ba của Google trong vòng sáu tuần cho thấy tốc độ phát triển của dòng model này đang được đẩy nhanh.

Quan trọng hơn, Google đang thay đổi vai trò của chữ “Flash”. Đây không còn chỉ là những mô hình được tối ưu cho tốc độ phản hồi và chi phí thấp, mà đang trở thành nhóm AI hướng tới reasoning, coding và các tác vụ agentic phức tạp.

Với mức giá đầu vào hiện tại chỉ 0,75 USD/1 triệu token, Gemini 3.8 Flash có thể tạo thêm áp lực cạnh tranh lên các mô hình AI frontier có chi phí vận hành cao hơn, đặc biệt trong những hệ thống cần xử lý hàng triệu hoặc hàng tỷ token.

Tuy nhiên, mức giá hiện tại là giá giới thiệu và sẽ tăng gấp đôi từ đầu năm 2027. Do đó, lợi thế chi phí của Gemini 3.8 Flash cần được đánh giá trong cả ngắn hạn và dài hạn, thay vì chỉ dựa trên mức giá ra mắt.

Hiện Gemini 3.8 Flash đã được Google cung cấp cho nhà phát triển thông qua Gemini API, Google AI Studio và một số nền tảng dành cho lập trình viên; người dùng Google AI Pro và Ultra cũng có thể tiếp cận model trong Gemini và một số sản phẩm của Google.

Phúc Hưng (t/h)