

Báo cáo Phân tích Chuyên sâu về Cơ chế Quản lý Tài nguyên, Cấp phát Quota và Tích hợp Tài khoản Gia đình trên Nền tảng Google Antigravity

Giới thiệu Tổng quan về Nền tảng Google Antigravity và Hệ sinh thái AI Agentic

Trong bối cảnh phát triển nhanh chóng của kỹ thuật phần mềm được hỗ trợ bởi trí tuệ nhân tạo (AI), nền tảng Google Antigravity đã tạo ra một sự dịch chuyển mang tính nền tảng về mặt kiến trúc. Khác với các công cụ hỗ trợ lập trình tuyến tính truyền thống (như GitHub Copilot hay các phiên bản IDE trước đây) vốn chỉ tập trung vào việc tự động hoàn thành mã nguồn (autocomplete) trong phạm vi một tệp duy nhất, Google Antigravity được thiết kế như một môi trường phát triển ưu tiên tác nhân (agent-first development environment).¹ Nền tảng này tích hợp một giao diện "Mission Control" (Trung tâm Điều khiển), cho phép người dùng điều phối nhiều tác nhân tự trị (autonomous agents) hoạt động song song để lên kế hoạch, thực thi mã lệnh trên terminal, và xác thực giao diện qua trình duyệt.¹ Việc chuyển đổi từ mô hình sinh mã phản ứng sang mô hình thực thi tác vụ chủ động này đòi hỏi một mức độ tài nguyên tính toán khổng lồ.⁴

Khi một lập trình viên yêu cầu một tác nhân thực hiện một tác vụ phức tạp, hệ thống không chỉ đơn thuần tạo ra một chuỗi văn bản phản hồi.¹ Thay vào đó, nó kích hoạt một chuỗi các hoạt động không đồng bộ: xây dựng kế hoạch triển khai, phân tích hàng nghìn dòng mã nguồn hiện có, viết mã cho cả giao diện người dùng (frontend) và hệ thống máy chủ (backend), thực thi các lệnh trên terminal để khởi chạy môi trường thử nghiệm cục bộ, và cuối cùng là điều hướng một trình duyệt không giao diện (headless browser) để kiểm tra trực quan các thay đổi.¹ Mỗi bước trong quy trình tự trị này đều đòi hỏi việc suy luận liên tục từ các mô hình ngôn ngữ lớn (LLM), liên tục cập nhật và nạp lại cửa sổ ngữ cảnh (context window), cũng như đánh giá các chỉ thị hệ thống.⁶ Do đó, tài nguyên máy chủ—đặc biệt là chu kỳ xử lý của GPU và TPU—được tiêu thụ bởi một câu lệnh duy nhất trong môi trường agentic như Antigravity cao hơn gấp nhiều lần so với các giao diện chatbot tiêu chuẩn.⁴

Sự gia tăng đột biến về nhu cầu điện toán này đã buộc các nhà cung cấp dịch vụ đám mây, cụ thể là Google, phải triển khai các hệ thống quản lý tài nguyên phức tạp và đa tầng nhằm duy trì sự ổn định của dịch vụ và đảm bảo phân phối công bằng cho tất cả người dùng.⁴ Đối với những người dùng đăng ký gói dịch vụ Google AI One Pro, chiến lược quản lý tài nguyên này được thể hiện thông qua một hệ thống hạn mức (quota) và tín dụng (credits) cực kỳ phức tạp, đôi khi gây

ra sự nhầm lẫn sâu sắc trong quá trình sử dụng.⁸

Báo cáo này được biên soạn nhằm phân tích một cách có hệ thống và giải mã toàn bộ các cơ chế vận hành bên trong khuôn khổ quản lý tài nguyên của Google Antigravity, đặc biệt tập trung vào các thách thức cụ thể liên quan đến tình trạng cạn kiệt hạn mức mô hình (model quota), chính sách chia sẻ gia đình (family sharing), hiện tượng khóa tài khoản nhiều ngày (multi-day lockouts), và sự khác biệt về mặt kinh tế học giữa hạn mức cơ sở và tín dụng AI (AI credits).⁹ Dựa trên việc phân tích kỹ lưỡng hình ảnh giao diện người dùng được cung cấp, kết hợp với các tài liệu kỹ thuật của nền tảng và dữ liệu đo lường từ cộng đồng lập trình viên vào tháng 3 năm 2026, báo cáo này cung cấp một cái nhìn toàn diện về cách thức đo lường, giới hạn và thương mại hóa tài nguyên tính toán trong hệ sinh thái Google Antigravity.

Phân tích Chi tiết Hình ảnh Giao diện Người dùng và Hiện trạng Quota

Dữ liệu hình ảnh được trích xuất từ giao diện của nền tảng Antigravity trên hệ điều hành macOS cung cấp những bằng chứng trực quan và định lượng vô cùng quan trọng để giải thích tình trạng mà hệ thống đang áp đặt lên tài khoản của người dùng. Hình ảnh hiển thị cửa sổ "Settings - Models" (Cài đặt - Mô hình), nơi cung cấp bức tranh toàn cảnh về việc phân bổ tài nguyên hiện tại.

Các dữ liệu được trích xuất từ hình ảnh cho thấy một bức tranh rõ ràng về tình trạng cạn kiệt tài nguyên diện rộng trên nhiều mô hình ngôn ngữ lớn khác nhau. Cụ thể, trong phần "MODEL QUOTA", giao diện hiển thị danh sách các mô hình mà người dùng có quyền truy cập thông qua gói Google AI One Pro, cùng với tình trạng hạn mức hiện tại:

- **Gemini 3.1 Pro (High) và Gemini 3.1 Pro (Low):** Cả hai thanh tiến trình đều hiển thị ở mức 0% (trống rỗng). Thời gian làm mới (refresh) được hệ thống ghi nhận là "Refreshes in 5 days, 12 hours" (Làm mới sau 5 ngày, 12 giờ).
- **Gemini 3 Flash:** Thanh tiến trình hiển thị một phần nhỏ dung lượng còn lại (khoảng 20%), tương đương với một vạch. Thời gian làm mới của mô hình nhẹ này ngắn hơn đáng kể, được ghi nhận là "Refreshes in 4 hours, 59 minutes" (Làm mới sau 4 giờ, 59 phút).
- **Claude Sonnet 4.6 (Thinking), Claude Opus 4.6 (Thinking), và GPT-OSS 120B (Medium):** Cả ba mô hình của bên thứ ba (được cung cấp thông qua Vertex AI Model Garden) đều hiển thị thanh tiến trình trống rỗng kèm theo một biểu tượng cảnh báo hình tam giác màu vàng. Đáng chú ý, thời gian làm mới của các mô hình này thậm chí còn dài hơn dòng Gemini Pro, cụ thể là "Refreshes in 5 days, 14 hours" (Làm mới sau 5 ngày, 14 giờ).

Bên cạnh đó, hình ảnh cũng cho thấy một tính năng quan trọng ở phần "MODEL CREDITS". Nút chuyển đổi (toggle) mang tên "Enable AI Credit Overages" (Cho phép vượt hạn mức bằng Tín dụng AI) hiện đang ở trạng thái tắt (màu xám). Ngay bên dưới nút này là một đoạn văn bản giải thích cơ chế hoạt động: "When toggled on, Antigravity will use your AI credits to fulfill model requests once you're out of model quota. Antigravity will always use your model quota first

before using AI credits." (Khi được bật, Antigravity sẽ sử dụng tín dụng AI của bạn để đáp ứng các yêu cầu mô hình khi bạn đã hết hạn mức mô hình. Antigravity sẽ luôn sử dụng hạn mức mô hình của bạn trước khi sử dụng tín dụng AI).

Tình trạng hiện tại của giao diện này xác nhận một số vấn đề cốt lõi. Đầu tiên, người dùng đã hoàn toàn cạn kiệt hạn mức cơ sở (baseline quota) cho tất cả các mô hình có hiệu suất cao, từ các mô hình do Google tự phát triển (Gemini 3.1 Pro) cho đến các mô hình đối tác (Claude 4.6 và GPT-OSS). Chỉ duy nhất mô hình tốc độ cao với chi phí thấp (Gemini 3 Flash) là còn một lượng nhỏ tài nguyên với chu kỳ làm mới ngắn. Thứ hai, sự xuất hiện của khoảng thời gian chờ lên tới hơn 5 ngày cho thấy tài khoản này đã kích hoạt một giới hạn an toàn cấp hệ thống, thay vì giới hạn ngắn hạn thông thường. Việc nút "Enable AI Credit Overages" đang bị tắt giải thích lý do tại sao hệ thống không tự động chuyển sang sử dụng phương thức thanh toán dự phòng, khiến người dùng bị khóa hoàn toàn khỏi việc thực thi các tác vụ phức tạp.

Giải mã Cơ chế Tính toán Quota của Antigravity: Mô hình "Trọng số Tính toán"

Để giải thích tại sao hạn mức mô hình lại bị cạn kiệt nhanh chóng đến vậy, cần phải đi sâu vào cơ sở hạ tầng kiến trúc đo lường tài nguyên của Google. Không giống như các nền tảng AI truyền thống hoặc các công cụ lập trình đối thủ như Cursor hay Kilo—vốn thường phân bổ một số lượng yêu cầu cố định (ví dụ: 500 yêu cầu/tháng hoặc 500 yêu cầu nhanh/ngày)—Antigravity hoàn toàn loại bỏ phương pháp đếm "tin nhắn" hoặc "lời nhắc" (prompts) đơn thuần.⁵

Thay vào đó, hệ thống hậu trường (backend) của Antigravity đánh giá và khấu trừ hạn mức dựa trên "trọng số tính toán" (computational weight) hay còn gọi là "khối lượng công việc đã hoàn thành" (work done).⁴ Khái niệm này là yếu tố cốt lõi để hiểu được tốc độ tiêu hao tài nguyên. Trọng số tính toán được xác định bởi nhiều yếu tố, bao gồm độ phức tạp của mô hình được gọi, kích thước của cửa sổ ngữ cảnh (bao gồm toàn bộ mã nguồn của dự án đang được phân tích), và số lượng các hành động tự trị mà tác nhân thực hiện.⁴

Các tác vụ khác nhau sẽ có chi phí điện toán khác nhau đáng kể:

- **Trò chuyện cơ bản (Simple Chat):** Mức độ tiêu thụ thấp. Việc yêu cầu mô hình giải thích một đoạn mã ngắn mà không yêu cầu phân tích toàn bộ kho lưu trữ (repository) sẽ tiêu tốn rất ít hạn mức.⁴
- **Chế độ Lập kế hoạch (Planning Mode):** Mức độ tiêu thụ trung bình. Tác nhân "suy nghĩ" và liệt kê các bước trước khi hành động, tạo ra các hiện vật (artifacts) như tài liệu nhưng chưa can thiệp sâu vào hệ thống tệp.¹
- **Chỉnh sửa Đa tệp (Multi-File Edits):** Mức độ tiêu thụ cao. Tác nhân phải quét toàn bộ kho lưu trữ, nạp mã nguồn vào ngữ cảnh và viết lại nội dung của nhiều bộ đệm (buffers) khác nhau.⁴
- **Tích hợp Trình duyệt/Terminal:** Mức độ tiêu thụ rất cao. Việc tác nhân tự động mở trình duyệt, điều hướng web, chụp ảnh màn hình và đối chiếu mã nguồn tạo ra một chu kỳ suy

luận liên tục và cực kỳ tốn kém.²

Điểm nghẽn nguy hiểm nhất dẫn đến việc cạn kiệt hạn mức nhanh chóng là hiện tượng "vòng lặp vô hạn" (infinite loops) hoặc chuỗi "độc thoại nội tâm" (inner monologues) của tác nhân.⁴ Khi một tác nhân được giao nhiệm vụ tự trị, nó có thể thực hiện hàng chục lệnh gọi API (API calls) ẩn dưới nền. Nếu tác nhân gặp lỗi trong quá trình biên dịch mã, nó sẽ tự động đọc thông báo lỗi, gửi lại toàn bộ ngữ cảnh khổng lồ đó cho LLM để suy luận giải pháp, sau đó thử viết lại mã.⁴ Quá trình này được lặp đi lặp lại. Về mặt toán học, lượng token tiêu thụ không tăng theo cấp số cộng mà tăng theo cấp số nhân khi độ dài của chuỗi ngữ cảnh ngày càng phình to. Điều này giải thích tại sao một lập trình viên có thể chỉ nhập một lời nhắc duy nhất, nhưng tác nhân xử lý ngầm đằng sau lại đốt sạch toàn bộ hạn mức của gói Pro chỉ trong một vài giờ làm việc căng thẳng.⁴

Nghịch lý Thời gian Làm mới: Hệ thống Quota Đa tầng (Chu kỳ 5 Giờ và Cầu dao 7 Ngày)

Thắc mắc cốt lõi và cũng là sự phẫn nộ phổ biến nhất của cộng đồng người dùng liên quan đến thời gian làm mới lên tới 5 ngày, thay vì 5 giờ như đã được quảng cáo.⁷ Theo các tài liệu chính thức từ Google, những người dùng đăng ký gói Google AI Pro sẽ nhận được "hạn mức cao, hào phóng, được làm mới mỗi 5 giờ cho các mô hình Gemini 3 Pro và các mô hình Vertex AI Model Garden khác".⁸ Sự khác biệt giữa cam kết "5 giờ" và thực tế "5 ngày 14 giờ" như trong hình ảnh cung cấp không phải là một lỗi hiển thị đơn thuần, mà là kết quả của một cấu trúc hạn mức đa tầng (multi-layered quota architecture) được thiết kế tinh vi nhằm ngăn chặn lạm dụng và cân bằng tải máy chủ.⁴

Kiến trúc này hoạt động dựa trên hai "xô" (buckets) tài nguyên độc lập, hoạt động đồng thời nhưng có giới hạn khác biệt:

1. Chu kỳ Nước rút 5 Giờ (The 5-Hour Sprint Window)

Đây là nguồn nhiên liệu ngắn hạn và là hạn mức mà người dùng nhìn thấy nhiều nhất khi sử dụng ở mức độ vừa phải.⁴ Mỗi khi người dùng bắt đầu một phiên làm việc mới sau một khoảng thời gian nghỉ, một chu kỳ 5 giờ sẽ bắt đầu đếm ngược.¹⁷ Nếu người dùng sử dụng hết dung lượng của "xô" ngắn hạn này, hệ thống sẽ tạm khóa và thông báo rằng hạn mức sẽ làm mới sau 5 giờ.⁴ Sau đúng 5 giờ, xô này được làm đầy lại 100%, cho phép người dùng tiếp tục công việc.⁴ Cơ chế này lý giải cho việc tại sao mô hình Gemini 3 Flash trong hình ảnh vẫn tuân theo chu kỳ ngắn hạn (Refreshes in 4 hours, 59 minutes), vì đây là mô hình nhẹ, chưa chạm đến giới hạn hệ thống.¹⁶

2. Giới hạn Cơ sở Hàng tuần (The Weekly Baseline Circuit Breaker)

Ẩn bên dưới chu kỳ 5 giờ là một "cầu dao" toàn cầu (global circuit breaker) có chu kỳ 7 ngày.⁴

Mỗi một token, mỗi một trọng số điện toán được tiêu thụ trong các đợt "nước rút" 5 giờ đều được cộng dồn vào xô "chạy marathon" 7 ngày này.⁴ Giới hạn hàng tuần này được thiết lập để ngăn chặn các "người dùng quyền lực" (power users) liên tục làm trống xô 5 giờ ngay khi nó vừa được làm mới, suốt ngày đêm.¹⁷

Sự kiện khóa tài khoản nhiều ngày (multi-day lockout) xảy ra khi tổng lượng tài nguyên tiêu thụ trong vòng 7 ngày qua chạm đến trần của Giới hạn Cơ sở Hàng tuần.⁴ Khi sự cố này được kích hoạt, hệ thống Antigravity sẽ ngay lập tức vô hiệu hóa chu kỳ làm mới 5 giờ đối với các mô hình nặng.⁴ Giao diện người dùng sẽ chuyển đổi và hiển thị thời gian chờ đếm ngược dựa trên thời điểm mà xô 7 ngày sẽ được đặt lại.⁴

Điều này giải thích chính xác những con số xuất hiện trong hình ảnh của người dùng: 5 ngày 12 giờ cho Gemini 3.1 Pro và 5 ngày 14 giờ cho Claude 4.6.¹¹ Người dùng này đã thực hiện các tác vụ agentic quá phức tạp, liên tục làm cạn kiệt các đợt 5 giờ trong những ngày trước đó, dẫn đến việc chạm đỉnh giới hạn hàng tuần.¹⁶ Hệ thống đang buộc người dùng phải chờ cho đến khi chu kỳ 7 ngày tính từ đợt sử dụng đầu tiên kết thúc.¹⁶ Các đại diện của Google đã xác nhận trên diễn đàn dành cho nhà phát triển rằng, để cân bằng việc cung cấp hạn mức tốt nhất và duy trì sự công bằng, các giới hạn hàng tuần này được áp dụng nghiêm ngặt đối với gói AI Pro.⁷ Gói duy nhất được miễn trừ khỏi giới hạn hàng tuần này là Google AI Ultra, được định vị dành cho những nhà phát triển chuyên nghiệp với cường độ cao nhất.⁷

Đặc điểm	Chu kỳ Nước rút (Sprint)	Giới hạn Cơ sở Hàng tuần (Marathon)
Thời gian làm mới	Cuốn chiếu 5 giờ	7 ngày (168 giờ)
Bản chất	Hạn mức khả dụng cho một phiên làm việc ngắn hạn.	Tổng dung lượng tích lũy tối đa trong một tuần.
Hệ quả khi cạn kiệt	Khóa tạm thời, đếm ngược ~5 giờ trên giao diện.	Kích hoạt "Cầu dao", khóa hệ thống trong nhiều ngày (5-7 ngày).
Áp dụng	Mọi mô hình (Gemini, Claude, v.v.)	Tích lũy chéo giữa các mô hình nặng (Vertex AI).
Gói Ultra	Có áp dụng	Không bị giới hạn hàng tuần. ⁸

Chính sách Chia sẻ Gia đình (Family Sharing) và Tính Độc lập của Quota

Một trong những nghi vấn lớn nhất được đặt ra là liệu tình trạng cạn kiệt hạn mức lên tới 5 ngày này có phải do 4 thành viên khác trong nhóm Gia đình (Family members) đã sử dụng quá nhiều và tiêu tốn chung một "quỹ" tài nguyên hay không.¹⁰ Để trả lời câu hỏi này, cần phải phân tích kiến trúc phân bổ tài nguyên của Google One áp dụng riêng cho các dịch vụ AI tổng quát.

Các tài liệu kỹ thuật, kiến trúc phân bổ và các xác nhận từ các chuyên gia sản phẩm của Google trên các diễn đàn cộng đồng đã thiết lập một cách dứt khoát rằng: **Hạn mức và giới hạn mô hình trong Google Antigravity hoạt động theo cơ chế độc lập và không dùng chung (independent and unpooled quota allocation).**¹⁹

Khi người sở hữu tài khoản chính (người mua gói Google AI One Pro) bật tùy chọn "Share Google One with Family" (Chia sẻ Google One với Gia đình), các đặc quyền về AI thực sự được mở rộng cho tối đa 5 thành viên khác.¹⁴ Tuy nhiên, các hạn mức điện toán, số lượng yêu cầu hàng ngày, chu kỳ làm mới 5 giờ, và **giới hạn cơ sở hàng tuần 7 ngày được cách ly hoàn toàn cho từng cá nhân dựa trên tài khoản Google riêng biệt của họ.**¹⁰

Nói một cách dễ hiểu, một gói đăng ký Google AI Pro đóng vai trò như một bộ nhân (multiplier), có khả năng cấp tối đa 6 bộ hạn mức Pro riêng biệt và hoàn chỉnh.²² Nếu một thành viên trong gia đình sử dụng Antigravity để biên dịch các dự án lớn, tiêu thụ toàn bộ hạn mức 5 giờ và chạm trần 7 ngày của họ, **điều đó sẽ tạo ra tác động bằng không (zero impact) đối với thanh tiến trình và dung lượng còn lại của người mua gói chính.**²⁰ Việc người dùng này hết quota hoàn toàn không phải do 4 thành viên gia đình sử dụng quá mức, **mà là do chính bản thân người dùng đã tự tiêu thụ hết hạn mức hàng tuần của chính mình thông qua các luồng công việc (workflows) do bản thân tạo ra trên môi trường IDE cục bộ.**²⁰

Rào cản Xác thực và Sự cố Rớt hạng (Free Tier Anomalies)

Mặc dù quota là độc lập, nhưng các thành viên gia đình không tự động nhận được đặc quyền của gói Pro chỉ bằng việc nằm trong nhóm gia đình.²⁰ Google áp đặt các rào cản xác thực vô cùng nghiêm ngặt đối với các môi trường AI tổng quát có khả năng duyệt web và tạo mã tự động.²⁰

Điều kiện tiên quyết quan trọng nhất là xác minh độ tuổi. Antigravity áp dụng chính sách cấm người dùng dưới 18 tuổi.⁹ Các thành viên trong gia đình bắt buộc phải trải qua quá trình xác minh độ tuổi chính thức qua cổng thông tin tài khoản Google (bằng giấy tờ tùy thân, thẻ tín dụng hoặc ảnh selfie xác thực).¹⁵ Nếu một thành viên chưa hoàn tất quá trình này, hoặc hệ thống gặp sự cố mất đồng bộ phiên đăng nhập (OAuth desynchronization), máy chủ hậu trường sẽ tự động xếp tài khoản đó vào hạng "Miễn phí" (Free tier).²⁰ Tài khoản hạng Miễn phí phải chịu một hạn mức hàng tuần cực kỳ thấp và hoàn toàn không có cơ chế làm mới 5 giờ,

khiến thành viên đó bị khóa gần như ngay lập tức sau vài lần sử dụng.⁸ Đây là một hiện tượng phổ biến khiến nhiều người quản lý nhóm gia đình lầm tưởng rằng hệ thống đang chia sẻ chung một mức quota nhỏ giọt.²⁰

Sự Khác biệt Cốt lõi Giữa Model Quota và AI Credits

Để tối ưu hóa doanh thu trong khi vẫn duy trì lời hứa tiếp thị về các dịch vụ "hào phóng" (generous), Google đã phân chia năng lực điện toán trên nền tảng Antigravity thành hai loại tiền tệ riêng biệt: Hạn mức Mô hình Cơ sở (Baseline Model Quota) và Tín dụng AI (AI Credits).⁸ Hiểu được sự phân đôi này là chìa khóa để giải mã tình năng "Enable AI Credit Overages" mà người dùng đang thắc mắc.⁸

Hệ thống được thiết kế theo cấu trúc thác nước (waterfall). Bất cứ khi nào một tác nhân Antigravity gửi yêu cầu, máy chủ sẽ luôn kiểm tra số dư của Hạn mức Mô hình trước tiên.⁸ Chỉ khi hạn mức này chạm mức 0%, hệ thống mới đánh giá đến tính khả dụng của Tín dụng AI.⁸

Hạn mức Mô hình (Baseline Model Quota)

- **Bản chất:** Đây là nguồn tài nguyên phi tiền tệ, được cấp phép theo thời gian và là giá trị cốt lõi đi kèm với các gói đăng ký định kỳ (Free, AI Pro, hoặc AI Ultra).⁸ Người dùng không trả tiền trực tiếp cho mỗi lần gọi API trong phạm vi hạn mức này.
- **Đơn vị đo lường:** Không được đo lường bằng số lượng yêu cầu API tuyến tính hay số đô la, mà sử dụng thuật toán "trọng số tính toán" nội bộ mờ đục như đã phân tích.⁴
- **Cơ chế phục hồi:** Không thể mua thêm bằng tiền. Hạn mức mô hình chỉ được phục hồi dựa trên thời gian, tuân theo hệ thống đa tầng: chu kỳ nước rút 5 giờ hoặc đợt tái thiết lập marathon toàn cầu 7 ngày.⁹
- **Biểu hiện giao diện:** Khi người dùng bị khóa 5 ngày như trong hình, chính là lúc Hạn mức Mô hình này đã bị vắt kiệt.¹¹

Tín dụng AI (AI Credits)

- **Bản chất:** Khác với hạn mức thời gian, Tín dụng AI là một dạng tiền tệ định lượng, hoạt động theo mô hình thanh toán trả sau (pay-as-you-go).⁸ Nó đóng vai trò là hệ thống dự phòng (fallback) để người dùng tiếp tục công việc khi Hạn mức Mô hình bị khóa.⁹
- **Đơn vị đo lường và Tiêu thụ:** Tín dụng AI minh bạch và trực tiếp hơn. Chúng bị khấu trừ một cách chính xác dựa trên bảng giá API tiêu chuẩn của Google Vertex AI cho từng mô hình cụ thể (được tính bằng số lượng token đầu vào và đầu ra).⁸
- **Cơ chế cung cấp:** Những người đăng ký Google AI Pro được cung cấp một khoản trợ cấp cố định là 1.000 tín dụng AI vào mỗi tháng.⁹ Khoản trợ cấp này không được cộng dồn nếu không sử dụng hết trong tháng.⁹ Tuy nhiên, nếu dùng hết, người dùng có thể mua thêm các gói nạp (top-ups) trực tiếp từ bảng điều khiển Google One, với các mức giá dao động từ \$24.99 cho 2.500 tín dụng, lên tới \$199 cho 20.000 tín dụng.⁹

Việc thiết lập hệ thống kép này cho phép Google đáp ứng nhu cầu thử nghiệm nguyên mẫu của người dùng cơ bản thông qua Hạn mức Mô hình, đồng thời chuyển đổi một cách tinh tế các nhà phát triển chuyên nghiệp (power users) sang mô hình khách hàng thanh toán API (thông qua AI Credits) mà không buộc họ phải rời khỏi giao diện IDE để thiết lập thanh toán Google Cloud Platform.⁸

Phân tích Cơ chế Hoạt động của "Enable AI Credit Overages"

Câu lệnh được trích dẫn trong câu hỏi của người dùng—*"Enable AI Credit Overages. When toggled on, Antigravity will use your AI credits to fulfill model requests once you're out of model quota. Antigravity will always use your model quota first before using AI credits"*—chính là bản chất của chiếc cầu nối giữa hai hệ thống kinh tế đã được phân tích ở trên.⁸ Nút chuyển đổi (toggle) này, hiện đang bị tắt đi (màu xám) trong hình ảnh, là cơ chế trao quyền kiểm soát cho người dùng.

Logic Vận hành của Nút chuyển đổi

Cài đặt "Enable AI Credit Overages" định đoạt hành vi của hệ thống Antigravity khi Hạn mức Mô hình (thanh tiến trình) chạm mức 0%.⁸

1. **Trạng thái TẮT (Never - Không bao giờ):** Đây là trạng thái an toàn mặc định để bảo vệ ví tiền của người dùng.⁸ Khi Hạn mức Mô hình cạn kiệt, IDE sẽ chặn đứng mọi yêu cầu tạo mã.⁸ Người dùng bị buộc phải nhận thông báo lỗi "Baseline model quota reached" và chờ đếm ngược (như 5 ngày 12 giờ).⁸ Ngay cả khi người dùng đang có sẵn hàng nghìn Tín dụng AI chưa sử dụng trong tài khoản, hệ thống cũng tuyệt đối không đụng đến chúng.⁸ Đây chính là tình trạng hiện tại của người dùng.
2. **Trạng thái BẬT (Always - Luôn luôn):** Nếu người dùng gạt nút này sang trạng thái bật, rào cản thời gian sẽ bị gỡ bỏ.⁸ Ngay khoảnh khắc Hạn mức Mô hình chạm đáy, hệ thống hậu trường sẽ âm thầm và mượt mà định tuyến lại các lệnh gọi API.⁸ Các tác vụ agentic tiếp theo sẽ được thực thi bằng cách "đốt" trực tiếp vào số dư Tín dụng AI của người dùng (bắt đầu từ 1.000 tín dụng miễn phí hàng tháng, sau đó đến tín dụng mua thêm).⁸

Đoạn văn bản *"Antigravity will always use your model quota first..."* là một cam kết về mặt kỹ thuật rằng Google luôn ưu tiên sử dụng tài nguyên miễn phí (hạn mức thời gian) trước.⁸ Hệ quả cực kỳ quan trọng là: Nếu người dùng đang làm việc bằng Tín dụng AI (trả phí), và thời gian đếm ngược 5 giờ (hoặc 7 ngày) của Hạn mức Mô hình kết thúc, hệ thống sẽ **ngay lập tức tự động chuyển về** sử dụng Hạn mức Mô hình vừa được làm mới, dừng việc trừ tiền từ quỹ Tín dụng AI.⁸

Tốc độ Tiêu hao Tín dụng Khủng khiếp và Sự cố Giao diện

Mặc dù tính năng này cung cấp một lối thoát để không bị gián đoạn công việc, nhưng thực tế

sử dụng cho thấy một tốc độ "đốt" tài nguyên kinh ngạc khi sử dụng AI Credits.²⁷ Khác với hạn mức mô hình mở đọc, AI Credits bị trừ trực tiếp dựa trên token API Vertex. Do các tác nhân Antigravity phải đẩy toàn bộ tệp mã nguồn khổng lồ vào cửa sổ ngữ cảnh để "suy nghĩ" (inner monologue) trước khi thực hiện hành động, số lượng token đầu vào (input tokens) luôn ở mức khổng lồ.¹

Nhiều lập trình viên đã báo cáo về sự hao hụt chóng mặt khi bật tính năng Overages. Một trường hợp ghi nhận rằng chỉ với một lời nhắc kiểm tra cực ngắn ("Test") gửi cho mô hình Claude Opus, số dư Tín dụng AI đã lập tức bốc hơi 1%.²⁸ Các tác vụ gỡ lỗi phức tạp hơn có thể thiêu rụi hàng trăm tín dụng chỉ trong vài phút, biến một luồng công việc vốn dĩ "miễn phí" dưới góc nhìn của người dùng Hạn mức Mô hình trở thành một hóa đơn đắt đỏ khi chạy qua Tín dụng AI.⁹

Thêm vào đó, tại thời điểm tháng 3 năm 2026, tính năng Overages này đang gặp phải những lỗi giao diện nghiêm trọng khi làm việc với các mô hình của bên thứ ba.²⁹ Cộng đồng đã ghi nhận lỗi khi Hạn mức Mô hình cho Claude Sonnet 4.6 cạn kiệt, ngay cả khi người dùng BẬT "Enable AI Credit Overages" và có sẵn 1.000 tín dụng, nút "Send" (Gửi) trong giao diện Antigravity vẫn bị tê liệt hoàn toàn (unresponsive).³⁰ Nút này ghi nhận thao tác nhấp chuột nhưng không khởi tạo bất kỳ yêu cầu mạng nào.³⁰ Tuy nhiên, nếu chuyển sang mô hình "cây nhà lá vườn" như Gemini 3 Flash, hệ thống lại hoạt động và trừ tín dụng bình thường.³⁰ Điều này cho thấy sự thiếu đồng bộ nghiêm trọng trong cơ chế định tuyến API và kiểm soát hạn mức giữa các máy chủ của Google và các máy chủ của đối tác Anthropic trên Vertex Model Garden.²⁹

Hướng dẫn Vận hành tại Vùng lãnh thổ, Điều kiện Ngôn ngữ và Tính năng Tích hợp Toàn diện

Khi sử dụng hệ sinh thái Google AI One Pro và Antigravity, người dùng cần nhận thức rõ những ràng buộc về mặt địa lý và ngôn ngữ, đặc biệt khi yêu cầu phân tích được đặt ra bằng tiếng Việt ("Tôi muốn bạn tìm kiếm...").

Mặc dù Google AI Pro và Antigravity đã khả dụng tại nhiều thị trường quốc tế, bao gồm cả Việt Nam (như một số người dùng tại Thành phố Hồ Chí Minh đã ghi nhận)³¹, tuy nhiên, về mặt kỹ thuật, nền tảng này hiện tại chỉ hỗ trợ chính thức và tối ưu hóa duy nhất cho ngôn ngữ tiếng Anh đối với các câu lệnh (prompts) và chỉ thị (instructions).⁹

Việc điều hướng các tác nhân Antigravity bằng tiếng Việt hoặc ngôn ngữ khác tiếng Anh tuy có thể nhận được phản hồi, nhưng lại tiềm ẩn nhiều rủi ro đối với hệ thống hạn mức. Lý do là các bộ phân tích cú pháp lệnh terminal, kịch bản điều hướng trình duyệt và các chỉ thị hệ thống nội bộ của LLM đều được tinh chỉnh cho tiếng Anh.²⁴ Khi người dùng nhập tiếng Việt, mô hình phải thực hiện thêm thao tác dịch thuật nội bộ trước khi lên kế hoạch và tạo mã. Điều này làm gia tăng đáng kể tải lượng token cho mỗi lời nhắc, khiến "trọng số tính toán" tăng vọt và làm cạn kiệt Hạn mức Mô hình nhanh hơn nhiều so với việc sử dụng tiếng Anh.¹³ Hơn nữa, độ chính xác

của các tác nhân (agent accuracy) sẽ giảm sút, dễ dẫn đến hiện tượng vòng lặp ảo giác (logic loops) khi tác nhân không hiểu rõ lỗi biên dịch, càng làm hao hụt nhanh chóng Hạn mức Cơ sở hàng tuần.⁴

Ngoài ra, cần lưu ý rằng gói Google AI One Pro không chỉ giới hạn ở Antigravity. Việc đăng ký gói này còn cấp cho người dùng, bao gồm cả các thành viên gia đình (đã xác minh 18+), quyền truy cập vào các giới hạn cao hơn trên toàn bộ hệ sinh thái AI của Google. Những tiện ích này bao gồm tính năng NotebookLM nâng cao (cho phép phân tích tới 300 nguồn tài liệu kích thước lớn), trợ lý lập trình Gemini Code Assist, khả năng gọi AI cho doanh nghiệp địa phương (hiện chỉ ở Mỹ), quyền tạo video giới hạn cao với Google Vids, và tích hợp AI vào Google Earth, Google Docs, Sheets, và Slides.⁹ Hệ sinh thái này cũng cấp cho người dùng Pro một khoản tín dụng Google Cloud trị giá \$10 mỗi tháng để hỗ trợ việc triển khai mã nguồn tạo ra từ Antigravity lên môi trường sản xuất (production).²⁶

Chiến lược Tối ưu hóa Nguồn tài nguyên và Quản lý Luồng công việc (Workflows)

Trước sự khắt khe của hệ thống "cầu dao" 7 ngày và chi phí đắt đỏ của Tín dụng AI, người dùng Antigravity thuộc gói AI Pro không thể tiếp tục sử dụng nền tảng này như một giao diện chat ChatGPT hay Gemini nền web thông thường.⁴ Việc áp dụng các chiến lược quản lý tài nguyên vi mô là bắt buộc.

1. Giám sát Hệ thống Đo lường (Telemetry)

Phương pháp phòng thủ đầu tiên chống lại lệnh cấm 5-7 ngày là giám sát chặt chẽ tỷ lệ Hạn mức Mô hình. Trong IDE, người dùng nên thường xuyên kiểm tra tab Agent Manager → Settings → Models.⁴ Khi dung lượng của các mô hình nặng (Gemini 3.1 Pro, Claude Opus) giảm xuống dưới mức 20%, người dùng cần thay đổi mô hình ngay lập tức.⁴ Việc chuyển hướng các tác vụ lập trình đơn giản (như giải thích mã, viết hàm đơn lẻ) sang các mô hình nhẹ hơn (như Gemini 3 Flash) sẽ giúp bảo tồn "xô" hạn mức hàng tuần, kéo dài thời gian sống sót cho đến đợt làm mới 5 giờ tiếp theo.⁴ Để theo dõi chi tiết hơn, cộng đồng đã phát triển các tiện ích mở rộng như "Antigravity Quota Watcher" (AGQ) để hiển thị thanh tiến trình quota theo thời gian thực ngay trên thanh trạng thái (status bar) của VS Code, giúp lập trình viên không bị bất ngờ trước các giới hạn ẩn.³²

2. Quản lý Độ tự trị của Tác nhân (Agent Autonomy)

Bởi vì hệ thống tính điểm dựa trên "khối lượng công việc đã hoàn thành", lập trình viên phải quản lý chặt chẽ sự tự trị của AI.⁴

- **Chế độ Lập kế hoạch (Planning Mode):** Thay vì sử dụng Chế độ Nhanh (Fast mode) và để AI tự do viết mã, sai, và sửa lại liên tục, hãy sử dụng Chế độ Lập kế hoạch.¹ AI sẽ tạo ra các Hiện vật (Artifacts) như danh sách tác vụ và kế hoạch triển khai trước khi động vào mã

nguồn.¹ Việc người dùng để lại bình luận (giống như Google Docs) để chỉnh sửa logic của AI ở bước này sẽ giúp tiết kiệm hàng triệu token so với việc để AI tự do biên dịch mã lỗi trên terminal.¹

- **Can thiệp Thủ công:** Lập trình viên phải sẵn sàng nhấn nút dừng tác nhân ngay khi nhận thấy nó đang rơi vào vòng lặp sửa lỗi không hiệu quả, nhằm bảo vệ Hạn mức Cơ sở.⁴ Việc lạm dụng các công cụ tự động hóa của bên thứ ba để kích hoạt Antigravity liên tục cũng sẽ kích hoạt bộ lọc lạm dụng của Google, dẫn đến lệnh khóa tài khoản 7 ngày ngay lập tức.⁴

3. Xử lý Lỗi Bất đồng bộ (Desynchronization Bugs)

Trong một số trường hợp, việc hệ thống hiển thị thời gian làm mới nhiều ngày mặc dù người dùng chưa hề sử dụng nền tảng có thể là một lỗi đồng bộ hóa token xác thực chứ không phải do dùng cạn quota.⁴ Để xử lý sự cố này, quy trình "hòa giải" (reconciliation) được khuyến nghị như sau:

- Đăng xuất hoàn toàn khỏi ứng dụng IDE Antigravity và đóng ứng dụng.⁴
- Truy cập vào tài khoản Google trên trình duyệt web, tìm mục kết nối bảo mật (OAuth) và thu hồi quyền truy cập (Remove all connections) đối với Antigravity.¹⁵
- Khởi động lại IDE và thực hiện đăng nhập lại từ đầu.¹⁵ Thao tác này buộc hệ thống phải gửi một truy vấn mới đến máy chủ thanh toán, đôi khi giúp xóa bỏ trạng thái "Miễn phí" bị lỗi và khôi phục lại chu kỳ làm mới 5 giờ chính xác của gói Pro.⁴

Kết luận

Các cơ chế vận hành của Google Antigravity đại diện cho một cách tiếp cận cực kỳ tinh vi, dù đôi khi thiếu minh bạch, nhằm quản lý lượng tài nguyên máy chủ khổng lồ bị tiêu thụ bởi các tác vụ phát triển phần mềm tự trị do AI điều khiển. Đối với người dùng Google AI One Pro đang phải đối mặt với tình trạng cạn kiệt hạn mức nhanh chóng và bị khóa tài khoản nhiều ngày, logic nền tảng đã cung cấp những câu trả lời dứt khoát.

Thứ nhất, việc cạn kiệt Hạn mức Mô hình hoàn toàn là kết quả từ các hoạt động sử dụng trong môi trường IDE cục bộ của chính người dùng. Kiến trúc chia sẻ gia đình của Google One phân bổ các hạn mức độc lập và cách ly cho từng thành viên.¹⁹ Việc một trong 4 thành viên gia đình sử dụng AI không thể tác động, cấu thành hay gây ra tình trạng khóa hạn mức trên tài khoản của người mua gói chính.²⁰ Người dùng phải chịu trách nhiệm hoàn toàn về lượng token do các tác nhân của mình tiêu thụ.

Thứ hai, thời gian làm mới kéo dài từ 5 đến 7 ngày không phải là một lỗi ngẫu nhiên, mà là sự thực thi có chủ đích của hệ thống quản lý tài nguyên đa tầng.⁴ Mặc dù gói dịch vụ quảng bá về chu kỳ làm mới mỗi 5 giờ (đợt nước rút), nhưng việc triển khai các tác nhân tự trị đa công cụ sẽ nhanh chóng làm cạn kiệt giới hạn cơ sở hàng tuần ẩn (đợt marathon).⁴ Một khi trần hàng tuần này bị xuyên thủng, tính năng làm mới 5 giờ sẽ bị vô hiệu hóa, và người dùng bị áp đặt một

khoảng thời gian chờ (cooldown) toàn cầu lên tới nhiều ngày.⁴

Thứ ba, nền tảng tự bảo vệ lợi ích kinh tế của mình thông qua sự phân nhánh tài nguyên nghiêm ngặt. Hạn mức Mô hình Cơ sở là phần tài nguyên phi tiền tệ, phụ thuộc vào thời gian và được bao gồm trong phí đăng ký định kỳ.⁹ Ngược lại, Tín dụng AI đóng vai trò là một đơn vị tiền tệ định lượng, thanh toán trực tiếp dựa trên biểu giá API Vertex.⁹ Nút chuyển đổi "Enable AI Credit Overages" là cầu nối linh hoạt giữa hai nền kinh tế này.⁸ Khi được bật, nó mang lại tính liên tục mượt mà cho dịch vụ khi Hạn mức Mô hình cạn kiệt, nhưng đi kèm với tốc độ tiêu hao tài chính khủng khiếp, do đặc thù cửa sổ ngưỡng cảnh khổng lồ của các tác vụ agentic chuyển hóa thành việc trừ tiền tín dụng nhanh chóng.⁸

Để tồn tại và làm việc hiệu quả trong hệ sinh thái Antigravity, lập trình viên buộc phải chuyển đổi từ tư duy "chat không giới hạn" sang vai trò của một người điều phối tài nguyên vi mô. Bằng cách giám sát các chỉ số hạn mức liên tục thông qua tiện ích mở rộng, định tuyến các tác vụ cơ bản sang các mô hình nhẹ, chủ động ngăn chặn các vòng lặp sửa lỗi vô hạn của AI, và chỉ kích hoạt Tín dụng AI vượt mức trong những trường hợp thực sự cần thiết, người dùng có thể lách qua khe hở của "cầu dao" hàng tuần và tận dụng toàn bộ sức mạnh tự trị của nền tảng mà không phải hứng chịu những lệnh khóa tài khoản kéo dài nhiều ngày. Môi trường phát triển của tương lai không chỉ đòi hỏi kỹ năng viết mã, mà còn đòi hỏi kỹ năng vận hành và quản lý ví điện toán của chính người dùng.

Works cited

1. Getting Started with Google Antigravity, accessed March 12, 2026, <https://codelabs.developers.google.com/getting-started-google-antigravity>
2. Build with Google Antigravity, our new agentic development platform, accessed March 12, 2026, <https://developers.googleblog.com/build-with-google-antigravity-our-new-agentic-development-platform/>
3. An Honest Review of Google Antigravity - DEV Community, accessed March 12, 2026, <https://dev.to/fabianfrankwerner/an-honest-review-of-google-antigravity-4g6f>
4. Navigating Antigravity Pro Quota Limits - Google AI Developers Forum, accessed March 12, 2026, <https://discuss.ai.google.dev/t/navigating-antigravity-pro-quota-limits/130212>
5. Cursor vs Trae Pricing 2025: Which AI IDE Subscription Wins?, accessed March 12, 2026, <https://zoer.ai/posts/zoer/cursor-vs-trae-subscription-pricing-comparison>
6. Antigravity AI, accessed March 12, 2026, <https://www.antigravityai.io/>
7. [BUG] Google AI Pro - Baseline quota refresh showing multi-day lock, accessed March 12, 2026, <https://discuss.ai.google.dev/t/bug-google-ai-pro-baseline-quota-refresh-showing-multi-day-lock-instead-of-advertised-5-hour-reset-in-antigravity/127678>
8. Plans - Google Antigravity Documentation, accessed March 12, 2026, <https://antigravity.google/docs/plans>

9. Use Google AI Pro benefits - Google One Help, accessed March 12, 2026,
<https://support.google.com/googleone/answer/14534406?hl=en>
10. how google ai pro family share works, accessed March 12, 2026,
<https://support.google.com/googleone/thread/399249550/how-google-ai-pro-family-share-works?hl=en>
11. Baseline model quota reached, 7 DAYS !?! - Google Antigravity, accessed March 12, 2026,
<https://discuss.ai.google.dev/t/baseline-model-quota-reached-7-days/131147>
12. first hour on antigravity, limit reached : r/google_antigravity - Reddit, accessed March 12, 2026,
https://www.reddit.com/r/google_antigravity/comments/1qg77o5/first_hour_on_antigravity_limit_reached/
13. Sitemap - 2025 - Kilo Blog, accessed March 12, 2026,
<https://blog.kilo.ai/sitemap/2025>
14. Family Sharing of Antigravity Limits on Google AI Ultra Plan - Reddit, accessed March 12, 2026,
https://www.reddit.com/r/google_antigravity/comments/1qvwtr5/family_sharing_of_antigravity_limits_on_google_ai/
15. [BUG] Google AI Pro in Antigravity shows multi-day lockouts instead, accessed March 12, 2026,
<https://discuss.ai.google.dev/t/bug-google-ai-pro-in-antigravity-shows-multi-day-lockouts-instead-of-advertised-5-hour-quota-refresh/129215>
16. Google AI Pro – Antigravity quota shows multi-day lockouts instead ..., accessed March 12, 2026,
<https://discuss.ai.google.dev/t/google-ai-pro-antigravity-quota-shows-multi-day-lockouts-instead-of-5-hour-reset/130202>
17. Pro has 5 hour refresh again? NICE!!! : r/google_antigravity - Reddit, accessed March 12, 2026,
https://www.reddit.com/r/google_antigravity/comments/1r8hhh2/pro_has_5_hour_refresh_again_nice/
18. Google Antigravity implementing weekly rate limits for all models, accessed March 12, 2026,
https://www.reddit.com/r/google_antigravity/comments/1q82mt3/google_antigravity_implementing_weekly_rate/
19. Google One AI Pro Family Sharing : r/google_antigravity - Reddit, accessed March 12, 2026,
https://www.reddit.com/r/google_antigravity/comments/1qew8mh/google_one_ai_pro_family_sharing/
20. Question regarding Family Sharing: Antigravity access and Token ..., accessed March 12, 2026,
<https://discuss.ai.google.dev/t/question-regarding-family-sharing-antigravity-access-and-token-quotas/115617>
21. Google AI Pro plan quota when sharing with family, accessed March 12, 2026,
<https://support.google.com/googleone/thread/381869963/google-ai-pro-plan-quota-when-sharing-with-family?hl=en>

22. Do Family Accounts Receive Antigravity Quota? : r/google_antigravity, accessed March 12, 2026,
https://www.reddit.com/r/google_antigravity/comments/1r13tn9/do_family_accounts_receive_antigravity_quota/
23. FAQ - Google Antigravity Documentation, accessed March 12, 2026,
<https://antigravity.google/docs/faq>
24. Get Google AI Ultra benefits, accessed March 12, 2026,
<https://support.google.com/googleone/answer/16286513?hl=en>
25. Looks like There is New Pricing etc for Quota Limits Check It Out, accessed March 12, 2026,
<https://discuss.ai.google.dev/t/looks-like-there-is-new-pricing-etc-for-quota-limits-check-it-out/130758>
26. New developer tools for Google AI Pro and Ultra subscribers, accessed March 12, 2026,
<https://blog.google/innovation-and-ai/technology/developers-tools/gdp-premium-ai-pro-ultra/>
27. You can use your AI credit now : r/google_antigravity - Reddit, accessed March 12, 2026,
https://www.reddit.com/r/google_antigravity/comments/1rr3uey/you_can_use_your_ai_credit_now/
28. Antigravity - Reddit, accessed March 12, 2026,
https://www.reddit.com/r/google_antigravity/new/
29. Partial refresh? Claude model "punishment"? : r/GoogleAntigravityIDE, accessed March 12, 2026,
https://www.reddit.com/r/GoogleAntigravityIDE/comments/1rqwoz9/partial_refresh_claude_model_punishment/
30. AI Credit Overages toggle not working — send button unresponsive, accessed March 12, 2026,
<https://discuss.ai.google.dev/t/ai-credit-overages-toggle-not-working-send-button-unresponsive-with-claude-non-gemini-models/130907>
31. Critical quota bug on Google AI Pro subscription – 2-day lock on, accessed March 12, 2026,
<https://discuss.ai.google.dev/t/critical-quota-bug-on-google-ai-pro-subscription-2-day-lock-on-gemini-3-pro-7-day-on-claude-sonnet-despite-active-pro-plan/119973>
32. Antigravity Quota (AGQ) - Open VSX Registry, accessed March 12, 2026,
<https://open-vsx.org/extension/henrikdev/ag-quota>