AI cục bộ trong an ninh mạng

Ứng dụng, tác động, rủi ro và giá trị phần cứng

ViWeb Technology | viweb.net Ấn bản nghiên cứu 1.0 | Ngày 17 tháng 11 năm 2026 Dành cho kỹ thuật viên và người dùng AI | Tham chiếu thị trường Mỹ, đơn vị USD

Tóm tắt điều hành

AI cục bộ là một lựa chọn thiết thực để hỗ trợ xử lý nhật ký nhạy cảm, mã nguồn và tài liệu nội bộ khi tổ chức có thể bảo vệ toàn bộ ứng dụng. Điểm khởi đầu phù hợp nhất là trợ lý có người giám sát: trợ lý giải thích bằng chứng và đề xuất công việc; kỹ thuật viên xác minh kết luận và kiểm soát hành động có tác động đáng kể. Chạy suy luận trên thiết bị cục bộ thay đổi nơi dữ liệu được xử lý; điều đó không làm mô hình trở nên đáng tin cậy hay tự bảo vệ máy chủ. [1] [2]

Bắt đầu bằng công việc có phạm vi rõ ràng. Ưu tiên tóm tắt sự cố, hỏi đáp trên quy trình vận hành được duyệt, rà soát cấu hình và soạn truy vấn phát hiện. Giữ bằng chứng gốc, yêu cầu dẫn nguồn và thử đầu ra trước khi sử dụng. Đây là đề xuất thử nghiệm, không phải kết quả ViWeb đã đo được.

Mua theo tác vụ vượt qua đánh giá. Mac nhỏ gọn là ứng viên hợp lý cho trợ lý một người dùng. Cần so sánh các hệ thống Ryzen AI Max+ 395 khi mô hình lớn hơn đòi hỏi 64GB hoặc 128GB bộ nhớ dùng chung. Cần đánh giá máy có GPU rời khi phần mềm đã chọn yêu cầu CUDA. Dung lượng bộ nhớ, tốc độ sinh token và kết quả an ninh hữu ích là các thước đo khác nhau; báo cáo không xác lập một lựa chọn tối ưu về giá/hiệu năng cho mọi trường hợp.

Đặt kiểm soát quyền hạn bên ngoài mô hình. Chèn lệnh vào đầu vào (prompt injection) có thể xuất hiện ngay trong tài liệu kỹ thuật viên đang điều tra. Kiểm soát truy xuất, giới hạn công cụ, xác thực đầu ra và cô lập máy chạy AI cần hạn chế hậu quả nếu tấn công thành công. Chỉ dặn trợ lý cẩn thận là không đủ. [2]

Hướng dẫn đọc

Bắt đầu tại	Câu hỏi được giải đáp
Mục 01-05	Tác vụ hữu ích, bằng chứng về tác động, rủi ro và triển khai có kiểm soát
Mục 06-10	Nhu cầu bộ nhớ, giá phần cứng tại Mỹ, giới hạn số liệu đo và chi phí vận hành
Mục 11-14	Thử trước khi mua, quy trình cho kỹ thuật viên, điều kiện áp dụng và giới hạn bằng chứng
Tài liệu tham khảo	Liên kết nguồn gốc và ghi chú mua sắm theo thời điểm

Nghiên cứu từ tài liệu có AI hỗ trợ. Không mua hay đo hiệu năng thiết bị cho báo cáo này. Giá là bằng chứng theo thời điểm, không phải chào bán. Việc công bố không tạo cam kết dịch vụ mới của ViWeb.

01 / AI cục bộ hỗ trợ kỹ thuật viên như thế nào?

Xác định ranh giới trước khi chọn mô hình

Trong báo cáo này, **AI cục bộ** là suy luận trên thiết bị do người dùng hoặc tổ chức kiểm soát. Máy tính cá nhân, máy chủ LAN nội bộ và máy trạm cô lập có ranh giới truy cập và sự cố khác nhau. **Ngoại tuyến** còn đòi hỏi toàn bộ quy trình hoạt động được khi không có kết nối bên ngoài; **cách ly mạng vật lý (air-gapped)** đòi hỏi sự tách biệt được thực thi và quy trình chuyển dữ liệu có kiểm soát. Giao diện cục bộ của sản phẩm không chứng minh được bất kỳ điều kiện nào trong số đó.

Phạm vi chủ yếu là mô hình ngôn ngữ sinh nội dung, truy xuất và quy trình có công cụ hỗ trợ. Phát hiện dựa trên thống kê, phần mềm chống mã độc và phân tích an ninh quy mô lớn vẫn hữu ích, nhưng không thể suy ra hiệu quả của chúng từ so sánh LLM. Huấn luyện mô hình nền tảng nằm ngoài phạm vi sử dụng máy tính để bàn được đề xuất.

Tác vụ nên thử nghiệm

Tác vụ	Đầu ra hữu ích	Kiểm tra bắt buộc
Nhật ký web, xác thực hoặc tường lửa	Dòng thời gian, nhóm sự kiện, giả thuyết	Đối chiếu bản ghi gốc, múi giờ và truy vấn xác định
Rà soát cấu hình, mã nguồn	Giải thích, mẫu đáng ngờ, đề xuất sửa	Kiểm tra đúng đoạn mã và tái hiện vấn đề trong môi trường thử
Đọc thông tin tình báo về đe dọa	Tóm tắt liên kết đến thông báo an ninh	Xác minh sản phẩm, phiên bản, điều kiện và ngày của thông báo
Xây dựng logic phát hiện	Bản nháp SQL, Sigma hoặc truy vấn khác	Xác minh lược đồ, cú pháp, báo động giả và mẫu độc hại đã biết
Tìm trong quy trình nội bộ	Câu trả lời kèm tham chiếu tài liệu	Thực thi quyền trước truy xuất; kiểm tra tính cập nhật và đoạn nguồn
Báo cáo sự cố, khôi phục	Bản nháp dựa trên bằng chứng	Xác minh diễn biến, phạm vi và mọi tuyên bố về xâm nhập hay khôi phục

Đây là các quy trình được đề xuất. Giá trị phụ thuộc vào mô hình, chất lượng đầu vào và thời gian sửa đầu ra. Giải thích thuyết phục không xác nhận một lỗ hổng hay sự cố.

Ưu tiên truy xuất trước khi tinh chỉnh mô hình

Sinh nội dung kết hợp truy xuất (RAG) đưa tài liệu được chọn vào lúc suy luận. RAG có thể giúp truy nguồn câu trả lời và cập nhật kiến thức vận hành, nhưng quyền của tài liệu gốc phải được giữ trong chỉ mục. Tinh chỉnh (fine-tuning) thay đổi hành vi mô hình và tạo thêm gánh nặng quản lý dữ liệu, mô hình. Cả hai kỹ thuật đều không loại bỏ prompt injection. [3] [2]

Trong thử nghiệm đầu tiên, nên chọn tập kiến thức nhỏ đã duyệt và quyền chỉ đọc thay vì lập chỉ mục rộng cho mọi khách hàng và máy chủ.

02 / Tác động: lợi ích phụ thuộc vào tác vụ

Đo năng suất cần đường cơ sở tương đương

Bản thảo nghiên cứu về phân loại email lừa đảo do tác giả thuộc Microsoft thực hiện tuyển 167 chuyên viên an ninh và loại 13 câu trả lời. Nghiên cứu báo cáo số mẫu độc hại được nhận diện đúng trên mỗi phút làm việc của chuyên viên tăng tới 6.5 lần khi có tác tử chuyên dụng. Người tham gia xử lý hàng đợi được chọn lọc, gồm lỗi tác tử được mô phỏng; thứ tự hàng đợi đóng góp đáng kể vào kết quả. Điều này hỗ trợ thử một quy trình cụ thể có AI trợ giúp, không chứng minh mô hình có trọng số công khai chạy cục bộ đạt mức cải thiện tương tự. [4]

Một đối chứng hữu ích là thử nghiệm lập trình đầu năm 2025 của METR: 16 lập trình viên có kinh nghiệm xử lý 246 tác vụ thực tế mất thêm 19% thời gian khi được dùng công cụ AI. Kết quả áp dụng cho bộ công cụ lịch sử và những kho mã cụ thể. Bản cập nhật tháng 2 năm 2026 mô tả hiệu ứng chọn mẫu khiến ước lượng về sau khó diễn giải. Không nên lấy trung bình các nghiên cứu này hay coi chúng là ước lượng hiện tại cho mọi kỹ thuật viên. [5] [6]

Hệ quả thực hành: đo thời gian hoàn thành và độ đúng của tác vụ thực tế, gồm chờ đợi, xác minh và làm lại. Bản nháp nhanh hơn vẫn có thể làm tăng chi phí của cuộc điều tra đã hoàn tất.

Triển khai cục bộ thay đổi sự đánh đổi vận hành

Lợi ích tiềm năng	Chi phí hoặc điều kiện đi kèm
Giảm việc thường xuyên gửi đầu vào bí mật đến mô hình thuê ngoài	Ứng dụng khách, embedding, nhật ký, bản sao lưu và công cụ cũng phải tuân thủ ranh giới dữ liệu
Tiếp tục hoạt động khi mất internet	Mô hình, phụ thuộc phải có sẵn; toàn bộ quy trình phải được thử ngoại tuyến
Kiểm soát phiên bản và thời điểm cập nhật	Tổ chức chịu trách nhiệm vá lỗi, thử hồi quy và quay lại phiên bản trước
Chi phí sở hữu dễ dự tính khi dùng ổn định	Công suất nhàn rỗi, quản trị, điện và rà soát có thể chiếm ưu thế khi ít sử dụng
Trợ lý nội bộ dùng chung	Xác thực, tách biệt khách hàng, điều phối và tính sẵn sàng trở thành trách nhiệm dịch vụ

Đây là cơ hội và sự đánh đổi về kiến trúc, không phải lợi ích đo được của thiết bị cụ thể.

Kẻ tấn công cũng hưởng lợi từ AI

Đánh giá năm 2026 của NCSC dự kiến AI làm tăng quy mô, tốc độ và hiệu quả của hoạt động tấn công hiện có; mức độ áp dụng không đồng đều và con người vẫn chỉ đạo. Tài liệu cũng nêu rủi ro từ AI tự phát ngoài kiểm soát tổ chức (shadow AI). Đây là đánh giá đe dọa, không phải số đếm tấn công do mô hình cục bộ gây ra. [7]

Công cụ cục bộ có trọng số công khai có thể giảm phụ thuộc vào nhà cung cấp dịch vụ và kiểm soát sử dụng của họ. Đó là tính chất lưỡng dụng, không chứng minh AI cục bộ là nguyên nhân riêng gây tội phạm mạng. Bên phòng vệ vẫn cần vá lỗi, xác thực mạnh, khôi phục đã thử và thu thập bằng chứng tin cậy.

03 / Rủi ro vẫn tồn tại trong ranh giới cục bộ

Coi trợ lý là bộ diễn giải chưa đáng tin

Bảng dưới ánh xạ các nhóm rủi ro đã công bố sang tình huống kỹ thuật. Tình huống và biện pháp kiểm soát là đề xuất thử nghiệm, không phải tuyên bố ViWeb hay sản phẩm được nêu đã gặp các sự cố này. Tham chiếu OWASP chỉ rõ những mục thuộc ấn bản 2025 đã được xem xét.

Rủi ro	Tình huống kỹ thuật	Kiểm soát cần thử
Prompt injection gián tiếp [2]	Trường nhật ký hoặc tài liệu hỗ trợ do kẻ tấn công kiểm soát đổi hướng trợ lý	Coi nội dung là bằng chứng; giới hạn năng lực được kích hoạt và thử đầu vào độc hại
Quyền tự hành quá mức [8]	Trợ lý sửa tài khoản, chặn lưu lượng hoặc xóa tệp vượt phạm vi tác vụ	API công cụ hẹp; mã thông thường kiểm tra quyền cho từng hành động; rà soát thay đổi có tác động lớn
Xử lý đầu ra thiếu an toàn [9]	Lệnh shell, SQL hoặc HTML do AI sinh được chuyển đến nơi thực thi	Xác thực kiểu và đường dẫn; tham số hóa truy vấn; mã hóa đầu ra; thử lệnh riêng
Rò rỉ hoặc đầu độc truy xuất [3]	Truy vấn của khách hàng này lấy bản ghi của khách hàng khác hoặc quy trình bị đầu độc	Lọc danh tính và khách hàng trước truy xuất; kiểm soát nạp và xóa dữ liệu
Nạp mô hình thiếu an toàn [10]	Gói tải về thực thi mã trong lúc khởi tạo	Ưu tiên định dạng chỉ chứa dữ liệu nếu được hỗ trợ; cố định phiên bản; kiểm tra mã tùy biến bắt buộc
Xâm nhập chuỗi cung ứng [11]	Adapter, gói chạy mô hình hoặc cập nhật công cụ đổi hành vi được tin cậy	Ghi nguồn gốc, phụ thuộc; thử cập nhật theo giai đoạn; giữ khả năng quay lui
Cạn tài nguyên [12]	Yêu cầu dài hoặc vòng lặp công cụ chiếm máy phân tích duy nhất	Giới hạn ngữ cảnh, đầu ra, đồng thời, hàng đợi, thời gian chạy và tổng hành động công cụ
Bịa thông tin và phụ thuộc quá mức [1]	Mô hình bịa chỉ báo, thông báo, chẩn đoán hoặc tuyên bố làm sạch thành công	Yêu cầu bằng chứng xác minh được; đánh giá lỗi thực tế; cho phép trả lời chưa chắc chắn

Ngoại tuyến không loại kẻ tấn công khỏi dữ liệu

Máy ngắt mạng vẫn có thể nhận tệp độc hại, tài liệu bị đầu độc hoặc gói bị xâm nhập qua quy trình chuyển dữ liệu. Hạn chế kết nối ra ngoài có thể giảm đường rò rỉ dữ liệu, nhưng không làm câu trả lời sai thành đúng hay ngăn thao tác phá hoại mà trợ lý được phép làm cục bộ.

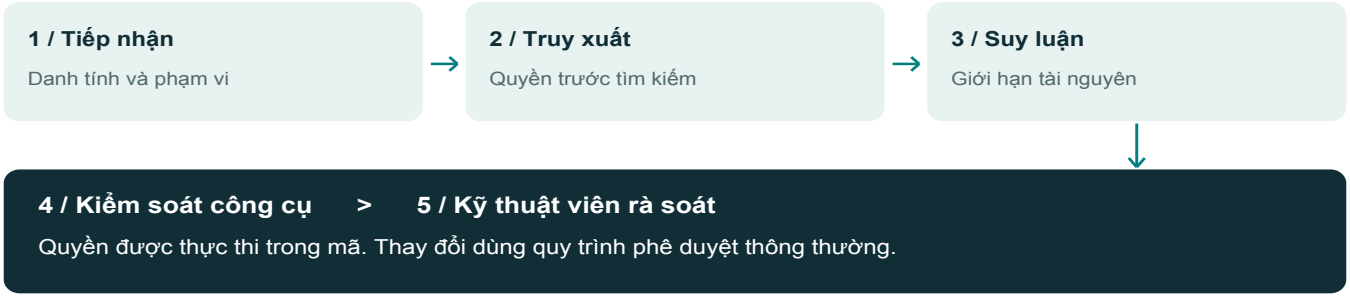
Dữ liệu nhạy cảm còn có thể tồn tại trong lịch sử chat, tệp tạm, bản ghi sự cố, chỉ mục và bản sao lưu thông thường. Bảo vệ chúng như dữ liệu vận hành, với quy tắc lưu giữ và truy cập rõ ràng. Phạm vi triển khai phải bao gồm giao diện và các kho dữ liệu phụ trợ, không chỉ trọng số mô hình. [13]

Mô hình từ chối yêu cầu gây hại là hành vi hữu ích, nhưng không đủ làm ranh giới duy nhất bảo vệ thông tin xác thực, dữ liệu khách hàng hoặc công cụ sản xuất.

04 / Thiết kế tham chiếu với ranh giới được thực thi

Tách bằng chứng, suy luận và hành động

Sau đây là mô hình triển khai được đề xuất cho nhóm kỹ thuật nhỏ. Đây là khuyến nghị thiết kế, không mô tả hạ tầng hiện tại của ViWeb.



1. Tiếp nhận: xác thực kỹ thuật viên, phân loại tác vụ và chọn phạm vi khách hàng hoặc dự án. Loại bí mật không cần để trả lời.

2. Bằng chứng: nạp tệp đã duyệt qua bộ phân tích bị giới hạn. Giữ bản gốc, định danh nguồn và dấu thời gian. Áp quyền tài liệu trước truy xuất.

3. Suy luận: chạy mô hình đã duyệt và dịch vụ embedding cục bộ bằng tài khoản riêng. Giới hạn bộ nhớ, số yêu cầu và thời gian. Trả lời kèm đoạn nguồn và mã bằng chứng.

4. Hành động: kiểm tra từng lời gọi công cụ theo danh sách cho phép, lược đồ tham số và quyền người dùng. Không đưa shell đa dụng và thông tin xác thực sản xuất vào thử nghiệm đầu tiên.

5. Rà soát: hiển thị đúng thao tác và mục tiêu cho kỹ thuật viên. Dùng quy trình thay đổi thông thường cho hành động có tác động lớn; giữ quyết định và kết quả.

Thiết kế áp dụng các nguyên tắc cô lập, kiểm soát truy cập và giám sát thông thường cho quy trình AI. Nó nhằm giảm hậu quả của lỗi mô hình mà không giả định mô hình có thể tự giám sát đáng tin cậy. [13] [8]

Quy tắc xử lý dữ liệu cần nêu rõ

Dữ liệu	Quy tắc thử nghiệm
Thông tin xác thực, token phiên, khóa riêng	Loại khỏi prompt và truy xuất; công cụ lấy thông tin xác thực có phạm vi hẹp bên ngoài mô hình
Nhật ký và mã nguồn khách hàng	Chỉ dùng trong mục đích và ranh giới khách hàng đã thống nhất
Chỉ mục truy xuất và embedding	Kế thừa độ nhạy của nguồn; thực thi truy cập, lưu giữ và lan truyền thao tác xóa
Nhật ký kiểm toán mô hình, công cụ	Ghi phiên bản, siêu dữ liệu hành động và mã bằng chứng; giảm nội dung nhạy cảm thô
Đầu ra mô hình và báo cáo	Áp cùng phân loại dữ liệu với bằng chứng bên trong

05 / Thiết lập trợ lý nhỏ một cách an toàn

Chọn môi trường chạy theo mô hình và nền tảng

llama.cpp là đường cơ sở suy luận đa nền tảng hữu ích với nhiều backend tăng tốc. **MLX** là lựa chọn cho Apple silicon. **Ollama** đơn giản hóa vận hành mô hình cục bộ và cung cấp API tại máy. Vẫn phải xác minh kiến trúc mô hình, lượng tử hóa, backend và cách gọi công cụ cụ thể. Cài đặt thành công không chứng minh GPU được dùng hay các nền tảng có tính năng tương đương. [14] [15] [16]

Với Ryzen AI Max, kiểm tra hướng dẫn tương thích hiện hành của AMD cho hệ điều hành, kernel, driver và framework dự định dùng. Tài liệu RDNA3.5 mô tả cấu hình bộ nhớ và yêu cầu kernel. Suy luận Vulkan chạy được không tự động xác nhận tác vụ huấn luyện hay phục vụ mô hình bằng ROCm. [17]

Ví dụ: xác minh quy trình Ollama chỉ chạy cục bộ

Tài liệu hiện tại cung cấp thiết lập dưới đây. Áp vào môi trường của đúng tiến trình máy chủ, khởi động lại và xác minh hành vi với phiên bản đã cài. [16]

```
OLLAMA_NO_CLOUD=1
```

Cấu hình máy chủ được tài liệu hóa cũng có thể chứa:

```
{
  "disable_ollama_cloud": true
}
```

Ollama nêu địa chỉ lắng nghe mặc định là `127.0.0.1:11434`. Giữ thử nghiệm một người dùng trên loopback. Chia sẻ qua LAN cần ranh giới dịch vụ có xác thực và quy tắc mạng phù hợp. Thiết lập tắt tính năng đám mây không kiểm soát mọi kết nối của giao diện, plugin hay hệ điều hành riêng biệt.

Kiểm tra trước khi dùng đầu vào bí mật

• Xác nhận đúng mô hình cục bộ được chọn; không bật chuyển sang đám mây, dịch vụ embedding bên ngoài hoặc công cụ tìm web.

• Kiểm tra lưu lượng ra ngoài khi dùng toàn bộ giao diện và quy trình. Thử suy luận sau khi ngắt kết nối ngoài, với tài nguyên cần thiết đã tải sẵn.

• Kiểm tra địa chỉ lắng nghe, quyền tài khoản và truy cập thư mục khách hàng. Thử yêu cầu không được phép từ người hoặc thiết bị thứ hai.

• Xác nhận mô hình thực sự được nạp lên bộ tăng tốc và mức dùng bộ nhớ. Với Ollama, xem `ollama ps`; đồng thời giám sát tiến trình.

• Kiểm tra lịch sử, nhật ký debug và kho truy xuất xem có sao chép bí mật thử nghiệm ngoài ý muốn. Xác nhận thao tác xóa và chính sách lưu giữ dùng được.

• Thử cập nhật theo giai đoạn và quay lui. Ghi phiên bản mô hình, hash tệp, môi trường chạy, driver, cấu hình và kết quả đánh giá.

Khi phân tích mã độc, chuyển văn bản hoặc siêu dữ liệu đã trích xuất từ môi trường phân tích được cô lập phù hợp. Không chạy tệp thực thi chưa rõ nguồn trên máy AI thông thường chỉ để có lời giải thích tốt hơn.

06 / Tính nhu cầu bộ nhớ trước khi mua năng lực tính toán

Kích thước trọng số là mức tối thiểu

Ước lượng đầu tiên hữu ích là **số byte trọng số = số tham số x số bit mỗi trọng số / 8**. Bảng dùng GB thập phân và trọng số lý tưởng của mô hình dense. Tệp lượng tử hóa thực tế còn chứa hệ số, siêu dữ liệu và tensor có độ chính xác hỗn hợp. Bộ nhớ lúc chạy, bộ đệm khóa/giá trị (KV), hệ điều hành và ứng dụng khác làm tăng nhu cầu.

Mô hình danh nghĩa	Trọng số 4-bit lý tưởng	Trọng số 8-bit lý tưởng	Trọng số 16-bit lý tưởng
8 tỷ tham số	4GB	8GB	16GB
14 tỷ	7GB	14GB	28GB
32 tỷ	16GB	32GB	64GB
70 tỷ	35GB	70GB	140GB
120 tỷ	60GB	120GB	240GB

Dự trù minh họa: mô hình 32B 4-bit danh nghĩa có thể có tệp 20GB. Cộng 4GB bộ đệm khi chạy, 8GB KV cache và 8GB cho hệ điều hành cùng dịch vụ khác, tổng là 40GB. Đây là giả định, không phải số đo hay mức cấp phát chung. Đo mô hình thực tế với ngữ cảnh và số yêu cầu đồng thời dự kiến trước khi chọn cấu hình 48GB hoặc 64GB.

Ngữ cảnh và mức đồng thời có thể đổi kết quả

Với bộ đệm transformer thông thường, nhu cầu KV có thể ước lượng:

```
2 x layers x KV heads x head dimension
  x bytes per cache element x cached tokens x sequences
```

Công thức lần lượt dùng số lớp, số KV head, kích thước mỗi head, số byte mỗi phần tử đệm, số token đã lưu và số chuỗi. Với 32 lớp, 8 KV head, kích thước head 128 và cache 16-bit, 32,768 token chiếm khoảng 4GiB cho một chuỗi; bốn chuỗi như vậy chiếm khoảng 16GiB. Chia sẻ, nén cache và kiến trúc khác có thể đổi kết quả. GB và GiB là hai đơn vị khác nhau.

Mô hình mixture-of-experts có thể chỉ kích hoạt một phần tham số mỗi token, nhưng vẫn cần giữ toàn bộ trọng số các expert trong bộ nhớ hoặc quản lý bằng chiến lược chuyển tải. Ít tham số hoạt động không đồng nghĩa tệp tải về hay tập dữ liệu làm việc nhỏ tương ứng.

Từ mức vừa bộ nhớ đến quyết định mua

Có thể bắt đầu thử khoảng 24-32GB bộ nhớ dùng chung cho trợ lý nhỏ một người dùng; đánh giá 48-64GB khi mô hình hoặc ngữ cảnh lớn hơn có lý do; so sánh hệ 128GB khi nhu cầu đo được vượt mức nhỏ hơn. Đây là điểm xuất phát thử nghiệm, không bảo đảm mô hình sẽ vừa. Chuyển xử lý sang CPU hoặc dùng swap có thể giúp mô hình khởi động nhưng không đạt mục tiêu độ trễ.

07 / Danh sách phần cứng và giá tham chiếu tại Mỹ

So sánh đúng cấu hình và loại giá

Kiểm tra ngày 17 tháng 11 năm 2026. Đơn vị USD, chưa gồm thuế bán hàng. Phí vận chuyển, ngoại vi, hỗ trợ và giấy phép bổ sung không được tính trừ khi nêu rõ. Giá Apple lấy từ cột **Consumer (MSRP)** trong bảng tháng 9 năm 2026; các số khác là giá hiển thị tại cửa hàng nhà cung cấp. Chưa xác minh tổng thanh toán hay cam kết giao hàng.

Ứng viên	Bộ nhớ / lưu trữ	Bằng chứng giá tại Mỹ	Giá trị cần kiểm chứng
Mac mini M6	24GB / 512GB	$1,299 MSRP [18]	Trợ lý nhỏ gọn, một người dùng, mô hình nhỏ
Mac mini M5 Pro	24GB / 512GB	$1,699 MSRP [18]	Mac có băng thông cao hơn; cùng dung lượng với hàng trên
Framework Desktop DIY, Max+ 395	64GB / SSD tính thêm	$1,959 cho hệ thống [19]	Ứng viên Linux/Windows nhiều bộ nhớ hơn; tính đủ cấu hình
GMKtec EVO-X2, Max+ 395	64GB / 1TB	$2,199.99 niêm yết [20]	Ứng viên mức 64GB đã có bộ lưu trữ
Mac Studio M5 Max	36GB / 512GB	$2,499 MSRP [18]	Đánh giá khi cần năng lực nền tảng; riêng dung lượng chưa đủ lý do nâng cấp
Framework Desktop DIY, Max+ 395	128GB / SSD tính thêm	$3,449 cho hệ thống [19]	Dung lượng cao hơn; cộng lưu trữ, hệ điều hành và yêu cầu lắp ráp
MINISFORUM MS-S1 MAX, Max+ 395	128GB / 2TB	Hiển thị $3,959; cần báo giá [21]	Mạng/mở rộng; thông báo hết đợt giảm giá khiến giá chưa chắc chắn
Beelink GTR9 Pro, Max+ 395	128GB / 2TB	Giá giảm hiển thị $4,199 [22]	Ứng viên dung lượng cao đã đủ lưu trữ; xác nhận giá cuối và hỗ trợ
GMKtec EVO-X5 Pro, Max+ PRO 495	192GB / 2TB	$6,799 niêm yết [23]	Lựa chọn chuyên biệt vì dung lượng; tác vụ phải đáng với phần chênh giá

Giá nổi bật có thể thuộc mức bộ nhớ khác

Lựa chọn $2,199.99 trên trang EVO-X2 là **64GB/1TB**, dù trang cũng có lựa chọn 128GB. Trang MINISFORUM hiển thị giá khuyến mại kèm thông báo đợt bán đã kết thúc. Những giới hạn này có thể đổi quyết định mua và cần được giữ khi trích vào báo giá hay website.

Bộ cấu hình Framework niêm yết thêm SSD 1TB ở mức $215. Cộng với hệ 64GB được **$2,174 trước tùy chọn khác**, so với EVO-X2 niêm yết $2,199.99. Chênh lệch $25.99 quá nhỏ để quyết định nếu chưa xác nhận cấu hình đủ, hệ điều hành, hỗ trợ, hàng sẵn có và kết quả tác vụ. Đây là phép cộng linh kiện, không phải đơn hàng đã thực hiện. [19] [20]

Trang Apple hiện liệt kê Mac mini M6 và M5 Pro. Kết quả M4 cũ ở phần sau là số đo lịch sử; không được đổi nhãn thành hiệu năng M6 hoặc M5 Pro. [24]

08 / Nền tảng nào đáng xem xét kỹ hơn?

Mac mini: ứng viên nhỏ gọn cho tác vụ có giới hạn

Apple công bố băng thông 153GB/s cho cấu hình bộ nhớ M6 cơ bản, 170GB/s với 32GB và 307GB/s cho M5 Pro. M6 hỗ trợ tối đa 32GB, M5 Pro tối đa 64GB. Bộ nhớ dùng chung còn phục vụ hệ điều hành và ứng dụng khác, không dành toàn bộ cho trọng số. Cần báo giá riêng cho cấu hình 48GB hoặc 64GB; không áp giá bản 24GB. [24]

Khuyến nghị: bắt đầu bằng Mac mini khi macOS hỗ trợ công cụ dự định dùng và mô hình vừa bộ nhớ với phần dự phòng đủ. Chỉ thử cấu hình đắt hơn khi số đo về độ trễ, bộ nhớ hoặc công việc hữu ích khác giải thích được chi phí. Không suy ra hỗ trợ CUDA chỉ vì máy có GPU Apple.

Ryzen AI Max+ 395: so bộ nhớ dùng được và cấu hình đủ

AMD công bố giao tiếp LPDDR5x-8000 rộng 256-bit, bộ nhớ tối đa 128GB và Radeon 8060S. Giao tiếp này suy ra băng thông lý thuyết 256GB/s; đó là trần từ thông số, không phải tốc độ ứng dụng đã đo. AMD mô tả khả năng cấp tối đa 96GB cho đồ họa ở hệ 128GB. Firmware OEM, hệ điều hành và môi trường chạy quyết định cách cấp phát và dung lượng dùng được thực tế. [25] [26]

Khuyến nghị: đưa hệ 64GB vào danh sách cho tác vụ cỡ vừa, hệ 128GB khi tập dữ liệu làm việc đo được vượt mức đó. Xác nhận làm mát kéo dài, việc dùng bộ tăng tốc, cập nhật firmware, thay bộ lưu trữ và kênh hỗ trợ khả thi. Bộ nhớ LPDDR5x hàn sẵn khiến việc mua thiếu dung lượng khó sửa về sau.

GPU NVIDIA rời: tính giá cả máy trạm

RTX 5060 Ti có lựa chọn 16GB; RTX 5070 Ti có 16GB và RTX 5090 có 32GB. Nên đánh giá khi ứng dụng hưởng lợi từ CUDA hoặc bộ nhớ GPU riêng. VRAM vẫn là giới hạn; RAM hệ thống thông thường không trở thành VRAM có cùng tốc độ. Báo cáo chưa xác minh báo giá hiện tại cho máy hoàn chỉnh; giá GPU khởi điểm quảng cáo không phải chi phí của máy trạm dùng được. [27]

Nhãn AI không thay cho phép đo thông lượng

TOPS hoặc FLOPS quảng cáo phụ thuộc độ chính xác, độ thưa và thành phần được đo. Môi trường chạy dùng GPU không tự động dùng NPU. Xác nhận backend và tốc độ tác vụ thực tế, thay vì so chỉ số NPU với số tiếp thị của GPU hoặc cả hệ thống.

DGX Spark và hệ bộ nhớ dùng chung lớn hơn

Cấu hình Spark GB10 được xem xét có 128GB bộ nhớ dùng chung, băng thông 273GB/s và CPU Arm. Kiểm tra phụ thuộc trên Arm, bộ phần mềm phục vụ mô hình dự kiến và hiệu năng kéo dài. Trang mua hàng NVIDIA đã truy xuất chỉ đưa liên kết đối tác, không có giá Mỹ hiện tại; vì vậy đây là lựa chọn **cần báo giá** trong so sánh. Năng lực FP4 quảng cáo không xác lập tốc độ trả lời. [28] [29]

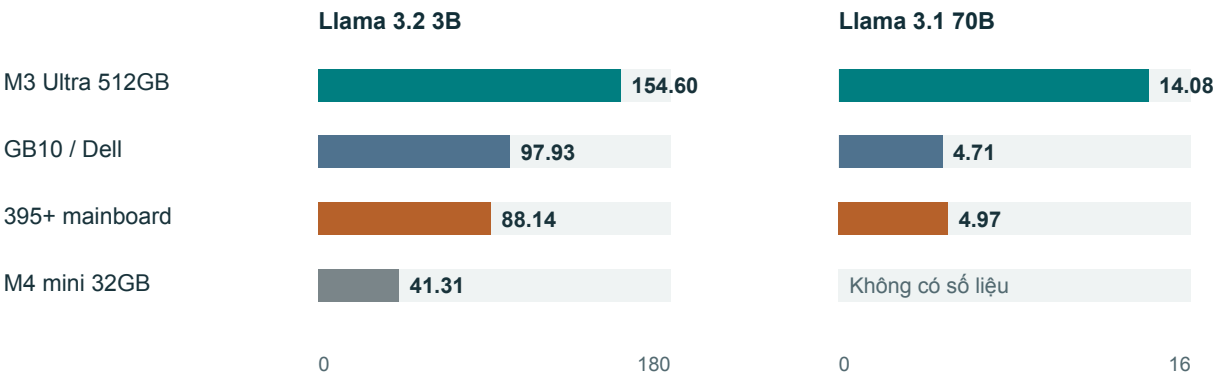
Khuyến nghị: cân nhắc Spark cho quy trình phát triển NVIDIA đã xác minh, hoặc hệ 192GB cho nhu cầu dung lượng đã chứng minh. Cả hai đều không mặc nhiên là lựa chọn tiết kiệm nếu thiếu báo giá đầy đủ và lợi thế đo được.

09 / Số đo đã công bố là tham khảo, không phải bảng xếp hạng mua sắm

Số đo lịch sử cho thấy vai trò của kích thước mô hình

Các số dưới đây là tốc độ sinh được báo cáo trong tập hợp phép đo của Jeff Geerling, cố định tại phiên bản README `2f56f19`. ViWeb không thực hiện các phép đo này. Tập hợp gồm nhiều hệ thống, phần mềm và thiết lập chạy khác nhau; đây không phải so sánh có kiểm soát của danh sách mua sắm hiện tại. [30]

Tốc độ sinh token được báo cáo (token/s)



Hai thang đo khác nhau. Số liệu lịch sử từ nhiều hệ; không phải thử nghiệm mua sắm có đối chứng.

Hệ thống được báo cáo	Llama 3.2 3B, token/giây	Llama 3.1 70B, token/giây
Mac Studio M3 Ultra, 512GB	154.60	14.08
Dell Pro Max với GB10	97.93	4.71
Bo mạch Framework, 128GB / 395+	88.14	4.97
Mac mini M4, 32GB	41.31	Không có số liệu

Nguồn gắn nhãn GPU/CPU cho hàng Framework 70B, nhưng GPU cho hàng 3B. Giá trị Mac mini 70B không có không đồng nghĩa bằng không. Chỉ tên/tag mô hình không xác nhận cùng hash tệp, thiết lập cache hay lượng tử hóa. Xem liên kết của từng hệ trong nguồn trước khi so sánh sâu hơn. [30]

Quyết định mua còn cần những gì?

• **Xử lý prompt và thời gian đến token đầu tiên:** nhật ký dài có thể nạp chậm dù đầu ra sau đó có vẻ nhanh.

• **Độ trễ toàn tác vụ:** truy xuất, chờ hàng đợi, công cụ và rà soát đều góp vào công việc hoàn tất.

• **Chất lượng tại mức lượng tử hóa đã chọn:** sinh nhanh hơn ít giá trị nếu bỏ sót bằng chứng quan trọng.

• **Hành vi khi nhiều người dùng đồng thời:** một phiên phản hồi nhanh vẫn có thể đi kèm độ trễ cao cho cả nhóm.

• **Điện năng và nhiệt khi chạy ổn định:** công suất adapter, TDP chip và công suất đỉnh không phải điện năng cho tác vụ.

Báo cáo không chia tốc độ lịch sử cho giá hôm nay để tuyên bố bên thắng. Tỷ lệ giá/hiệu năng có cơ sở cần cùng tác vụ, tệp mô hình, ngưỡng chất lượng và toàn bộ chi phí mua. Mục 11 đưa ra phép thử tạo được so sánh đó.

10 / Giá trị phụ thuộc công việc được chấp nhận và chi phí vận hành

Hai thước đo giá trị, kèm ngưỡng chất lượng

Giá trị phần cứng: số tác vụ được chấp nhận mỗi giờ chia cho toàn bộ chi phí mua. Chỉ so giữa hệ vượt cùng yêu cầu chất lượng và độ trễ. Báo cáo cơ cấu tác vụ thực tế; tóm tắt nhật ký và sửa mã không phải đơn vị hoán đổi được.

Giá trị vận hành: tổng chi phí tháng chia cho số tác vụ được chấp nhận hoàn tất trong tháng. Tính cả thời gian rà soát, sửa để câu trả lời dùng được. Truy cập dùng chung, sao lưu và hỗ trợ vẫn tốn chi phí dù không có hóa đơn theo token.

Mô hình chi phí cơ bản:

```
Monthly cost = (purchase cost - residual value) / useful months
  + energy + administration + licenses + storage/backup
  + review/rework + other allocated costs
```

Tức là phân bổ giá mua sau khi trừ giá trị còn lại theo số tháng sử dụng, rồi cộng điện, quản trị, giấy phép, lưu trữ/sao lưu, rà soát/làm lại và chi phí phân bổ khác.

Minh họa: mức sử dụng đổi chi phí mỗi tác vụ

Đây là phép tính theo giả định minh bạch, không dự báo lợi tức đầu tư, không phải giá điện hiện tại hay số đo của thiết bị trong bảng. Giả sử máy giá $2,200, không có giá trị còn lại, dùng 36 tháng, công suất trung bình tại ổ cắm 100W khi hoạt động, 8 giờ/ngày trong 22 ngày/tháng và điện $0.15/kWh.

Khoản theo tháng	Phép tính	USD
Phân bổ phần cứng	$2,200 / 36	$61.11
Điện	0.100kW x 176h x $0.15	$2.64
Quản trị	2 giờ x $50/giờ	$100.00
Tạm tính ba khoản trên	Tính bằng số chưa làm tròn	$163.75
Với 2,000 tác vụ được chấp nhận/tháng	Tạm tính / 2,000	$0.082/tác vụ
Với 200 tác vụ được chấp nhận/tháng	Tạm tính / 200	$0.82/tác vụ

Tạm tính chưa gồm giấy phép, sao lưu, rà soát, làm lại, thuế, tài chính và thời gian ngừng hoạt động; do đó **không phải TCO đầy đủ**. Nếu chạy liên tục 30 ngày với cùng giả định 100W, điện là $10.80/tháng. Cần đo công suất nhàn rỗi và khi hoạt động thực tế.

So cục bộ và thuê ngoài trên cùng điều kiện

Cục bộ hấp dẫn hơn khi sử dụng ổn định, mô hình đạt ngưỡng chất lượng và tổ chức coi trọng kiểm soát xử lý nhạy cảm. Suy luận thuê ngoài có thể có lợi kinh tế cho công việc ít hoặc tăng giảm đột biến khi cách xử lý dữ liệu chấp nhận được. Thiết kế kết hợp cần định tuyến rõ để yêu cầu cục bộ thất bại không âm thầm gửi dữ liệu bí mật ra ngoài.

Thử thiết bị hiện có trước khi thay. Chỉ điện năng chưa xác lập lợi ích môi trường; sản xuất phần cứng, mức sử dụng và phương án thay thế cũng quan trọng.

11 / Phép thử lặp lại được trước khi chi tiền

Cố định tác vụ và so hệ thống hoàn chỉnh

Chọn tập nghiệm thu nhỏ đại diện cho công việc, có đáp án được kiểm tra độc lập. Gồm trường hợp lành tính, độc hại và ví dụ cần trả lời chưa chắc chắn. Dùng tài liệu được phép hoặc dữ liệu tổng hợp; giữ dữ liệu khách hàng trong ranh giới đã thống nhất.

Lớp thử	Cần ghi và so sánh
Định danh hệ thống	Đúng thiết bị, báo giá, bộ nhớ, lưu trữ, hệ điều hành, firmware, driver, chế độ điện
Mô hình	Nhà phát hành, phiên bản bất biến, hash tệp, tokenizer, lượng tử hóa, ngữ cảnh và cache
Suy luận	Commit/phiên bản môi trường chạy, backend, chuyển tải, mẫu prompt, lấy mẫu, vị trí nạp trên bộ tăng tốc
Hiệu năng	Khởi động nguội/ấm, tốc độ xử lý prompt, sinh đầu ra, độ trễ toàn tác vụ p50/p95, bộ nhớ và điện tại ổ cắm
Dịch vụ	Một người dùng và mức đồng thời dự kiến, hàng đợi, timeout, phục hồi lỗi và làm mát kéo dài
Chất lượng	Cơ sở thực tế, bỏ sót đe dọa, báo động giả, truy vấn hợp lệ, độ đúng, công làm lại của người rà soát
Ranh giới	Tách biệt người dùng, hành vi với tài liệu độc hại, lời gọi công cụ bị từ chối, kết nối ra ngoài và xóa dữ liệu

Dùng cả phép đo nhỏ và chạy lại ứng dụng

Công cụ `llama-bench` của dự án gốc hỗ trợ đo xử lý prompt và sinh token, lặp nhiều lần và xuất dạng máy đọc được. Bắt đầu bằng cùng tệp mô hình và cấu hình, rồi chạy lại các tác vụ đại diện đầy đủ. Có thể thử tối ưu riêng cho từng nền tảng ở lượt khác và gắn nhãn tương ứng. [31]

Quy trình thử đề xuất: làm nóng hệ thống nhất quán, chạy ít nhất năm lần cho mỗi phép thử đã chọn và giữ kết quả thô. Thử độ dài ngữ cảnh dự kiến, phiên chạy đủ lâu để lộ giảm xung. Báo trung vị và độ phân tán; mẫu năm lần chạy không hỗ trợ kết luận đáng tin về độ trễ ở phần đuôi phân bố. Dùng lượt chạy lại ứng dụng dài hơn để đo p95 theo nhịp yêu cầu thực tế.

Đặt điều kiện đạt trước khi xem kết quả

Quy tắc khởi đầu hữu ích: không quan sát thấy rò rỉ chéo khách hàng hoặc hành động trái phép trong tập đối kháng xác định; không giảm đáng kể độ bao phủ phát hiện đe dọa (recall) so với đường cơ sở đã kiểm tra; và đạt cải thiện đã thống nhất trước về chi phí hoặc độ trễ của tác vụ hoàn tất. Đây là điều kiện nghiệm thu đề xuất, không chứng minh rủi ro thực tế bằng không.

Ghi kích thước mẫu, số đếm và độ bất định. Rà soát riêng các lần bỏ sót nghiêm trọng. Điểm tốt trên tập nhỏ là lý do mở rộng thử nghiệm có giám sát, không phải cho phép tự động hóa không giới hạn.

Mua hệ hoàn chỉnh ít tốn kém nhất vượt các yêu cầu chất lượng, dung lượng, độ trễ, an toàn và hỗ trợ. Nếu không hệ nào đạt, đổi quy trình hoặc mô hình trước khi tăng ngân sách phần cứng.

12 / Ba quy trình an ninh mạng thực hành

A. Giải thích mẫu lưu lượng web bất thường

Kỹ thuật viên xuất mẫu nhật ký có giới hạn và được phép, giữ bản gốc. Truy vấn xác định thiết lập số đếm, khoảng thời gian, mã phản hồi và đường dẫn yêu cầu. Trợ lý nhóm mẫu và đề xuất giải thích kèm tham chiếu bản ghi.

Kỹ thuật viên đối chiếu giải thích với yêu cầu gốc, cấu hình máy chủ và hoạt động bảo trì đã biết. Lưu lượng tăng chỉ trở thành phát hiện khi có bằng chứng hỗ trợ. Chặn nguồn hay đổi quy tắc đi theo quy trình vận hành thông thường. Văn bản của mô hình không bao giờ là lý do duy nhất để đổi lưu lượng sản xuất.

Thước đo thành công: thời gian đến lời giải thích đúng, có bằng chứng, tính cả xác minh; theo dõi bỏ sót trường hợp độc hại và báo động giả. Giữ chuỗi do kẻ tấn công kiểm soát như dữ liệu, và đưa trường hợp prompt injection vào đánh giá.

B. Rà soát thay đổi PHP hoặc JavaScript đáng ngờ

Dùng bản sao được phép trong môi trường thử. Nếu có, so với phiên bản gốc đáng tin; chạy kiểm tra tĩnh thông thường và quan sát hành vi. Yêu cầu trợ lý giải thích luồng điều khiển lạ, xác định rủi ro có thể có và đề xuất bản vá hẹp.

Xác minh độc lập đường dẫn, phụ thuộc và lỗ hổng được dẫn. Thử bản vá, kiểm tra hồi quy trước triển khai. Coi tên gói và lệnh cài đặt do mô hình sinh là chưa đáng tin cho đến khi xác nhận. Không chạy mẫu chưa rõ nguồn trên máy làm việc thường ngày. [9]

Thước đo thành công: số phát hiện và bản sửa an toàn được xác nhận độc lập mỗi giờ, gồm công sửa sai. Trợ lý tuyên bố tệp sạch không chứng minh đã gỡ hết mã độc.

C. Soạn báo cáo sự cố hoặc khôi phục

Cung cấp dòng thời gian đã xác minh, quan sát trong phạm vi xác định và bản ghi khắc phục đã duyệt. Yêu cầu bản nháp tách quan sát, suy luận và điều chưa biết. Mọi phát biểu quan trọng phải trỏ về bằng chứng.

Kỹ thuật viên xác minh hệ thống bị ảnh hưởng, ngày và kết quả kiểm tra thực tế. Nếu chỉ áp bản vá mà chưa điều tra số đầy đủ hay thử khôi phục, báo cáo phải nói đúng điều đó. Báo cáo khách hàng phải được bảo mật như bằng chứng gốc.

Thước đo thành công: thời gian đến bản nháp chính xác có thể công bố và số sửa đổi nội dung đáng kể. Viết trôi chảy không được tính là điều tra thành công.

Giữ năng lực của nhóm vận hành

Luân phiên người xác minh công việc có mô hình hỗ trợ và giữ bài tập cần đọc bằng chứng gốc. Khi mô hình gián đoạn, kỹ thuật viên vẫn phải điều tra, trao đổi và khôi phục bằng công cụ đã thiết lập. Ghi nơi trợ lý giúp ích, gây phân tâm và mắc lỗi tốn kém. Các quy trình là đề xuất, không mô tả công việc ViWeb đã hoàn thành cho khách hàng.

13 / Quyết định áp dụng theo giai đoạn

Giai đoạn 1: đánh giá một trường hợp trên phần cứng hiện có

Chọn quy trình chỉ đọc có đáp án rõ, như tập câu hỏi về quy trình vận hành hoặc tóm tắt nhật ký có giới hạn. Chỉ định người vận hành; xác định dữ liệu, quyền, yêu cầu bằng chứng và điều kiện dừng. Ghi đường cơ sở không có trợ lý và đánh giá trợ lý theo đó.

Điều kiện kết thúc: trợ lý tạo đầu ra hữu ích, xác minh được trên mẫu xác định và các kiểu lỗi đã được hiểu. Dừng nếu mô hình cần dùng không đạt ngưỡng chất lượng; riêng nâng phần cứng không sửa được lỗi thực tế hay phân quyền.

Giai đoạn 2: chỉ so phần cứng giải quyết giới hạn đã đo

Giới hạn quan sát được	Ứng viên đáng thử	Điều gì biện minh chi phí?
Mô hình nhỏ vừa; một kỹ thuật viên cần máy gọn	Máy hiện có hoặc Mac mini ở mức dung lượng đã thử	Chi phí tác vụ hoàn tất thấp hơn hoặc phản hồi đủ nhanh, hỗ trợ chấp nhận được
Tập dữ liệu làm việc vượt máy dùng chung bộ nhớ nhỏ	Max+ 395 64GB; báo giá riêng Mac mini nhiều bộ nhớ hơn	Vừa bộ nhớ và đạt độ trễ đo được, có dự phòng hệ điều hành và cache
Mô hình cùng ngữ cảnh vượt mức 64GB	Max+ 395 128GB; báo giá Mac lớn hơn tương đương	Cải thiện chất lượng hữu ích với tổng chi phí duy trì được
Phần mềm đã chọn yêu cầu CUDA	Báo giá máy trạm GPU rời hoặc hệ GB10	Tương thích đã xác minh, đủ VRAM/bộ nhớ dùng chung và thông lượng chấp nhận được
Tập dữ liệu làm việc vượt 128GB	Báo giá Mac nhiều bộ nhớ hơn hoặc nền tảng 192GB	Lợi ích tác vụ lớn hơn phần chênh chi phí dung lượng và gánh nặng vận hành

Đây là danh sách theo điều kiện. Không hàng nào khẳng định thiết bị nhanh nhất hay bảo đảm kích thước mô hình. Giữ báo giá nâng bộ nhớ và lưu trữ tương đương cùng hồ sơ đánh giá.

Giai đoạn 3: từ thiết bị sang dịch vụ được quản lý

Trước khi chia sẻ trợ lý, thêm kiểm soát danh tính, tách biệt khách hàng, hạn mức, bảo vệ truyền tải, giám sát và khả năng khôi phục. Thử lại ranh giới và tải sau thay đổi đáng kể về mô hình, môi trường chạy hoặc công cụ. Giữ cấu hình đã biết hoạt động và cách được ghi rõ để tắt công cụ hoặc quay lại quy trình thông thường.

Tự động hóa dần: chỉ soạn nháp, rồi công cụ chỉ đọc có phạm vi hẹp, sau đó thay đổi có rà soát nếu bằng chứng hỗ trợ. Hành động có tác động lớn vẫn cần phân quyền được thực thi và chủ thể vận hành chịu trách nhiệm rõ ràng.

Mặc định khuyến nghị cho độc giả ViWeb: bắt đầu bằng một trường hợp nhỏ có giám sát trên thiết bị hiện có phù hợp, chi phí thấp nhất. Lấy báo giá đầy đủ và chạy phép thử ở mục 11 trước khi coi Mac, Ryzen AI Max hay máy trạm GPU mới là khoản mua đáng giá.

14 / Bằng chứng, giới hạn và thuật ngữ

Ấn bản này xác lập điều gì?

Đây là hướng dẫn kỹ thuật nghiên cứu từ tài liệu, với nguồn được kiểm tra vào **ngày 17 tháng 11 năm 2026, múi giờ Asia/Ho_Chi_Minh**. Báo cáo kết hợp tài liệu gốc, thông số và giá từ nhà cung cấp, nghiên cứu nguyên bản và tập hợp phép đo lịch sử được định danh. Khuyến nghị, quy trình đề xuất và tình huống số được gắn nhãn.

Không có số đo thiết bị của ViWeb, dữ liệu năng suất khách hàng, mức giảm xâm nhập đã đo hay so sánh có kiểm soát của danh sách hiện tại. Vì vậy không thể xếp hạng tốc độ/giá chung. Không đặt hàng hay kiểm tra tồn kho ở bước thanh toán.

Giá tham chiếu thị trường Mỹ, đơn vị USD. Bằng chứng MSRP Apple có từ tháng 9 năm 2026; ngày kiểm tra nguồn không biến bảng cũ thành chào bán trực tiếp. Khuyến mại, mức bộ nhớ, thuế, bảo hành theo vùng, phí nhập khẩu và hàng sẵn có có thể đổi kết quả. Lấy báo giá giao đến nơi hiện tại trước khi mua.

Đọc bằng chứng thế nào?

Loại bằng chứng	Cách dùng phù hợp	Giới hạn
Hướng dẫn an ninh chính thức	Nhận diện đe dọa, thiết kế kiểm soát	Không chứng nhận thiết kế hay chứng minh rủi ro còn lại bằng không
Thông số và giá nhà cung cấp	Xác lập cấu hình có tên và giá hiển thị	Không xác lập tốc độ tác vụ, TCO đầy đủ hay giao hàng
Thí nghiệm nguyên bản	Hiểu kết quả cho người và quy trình được nghiên cứu	Không xác lập tác động tương tự với mô hình cục bộ
Báo cáo đo lịch sử	Minh họa hành vi được đo trên các hệ đã báo cáo	Khác lượt chạy và phần mềm hạn chế so sánh nhân quả
Tình huống số học	Khảo sát bộ nhớ hoặc chi phí dưới giả định rõ	Không phải quan sát, dự báo thị trường hay hứa hẹn hiệu năng

Thuật ngữ trong báo cáo

Suy luận (inference): dùng mô hình sẵn có để sinh câu trả lời. **Trọng số công khai (open-weight):** tham số mô hình được cung cấp theo giấy phép; không mặc nhiên cho phép dùng không hạn chế hay công khai đầy đủ quá trình huấn luyện. **Lượng tử hóa:** lưu trữ hoặc tính ở độ chính xác thấp hơn. **Prefill:** xử lý prompt đầu vào. **Decode:** sinh token tiếp theo. **KV cache:** trạng thái attention đã lưu khi sinh. **TTFT:** thời gian đến token đầu tiên. **TCO:** tổng chi phí sở hữu. **RAG:** sinh nội dung kết hợp truy xuất.

Công bố về biên tập

Chuẩn bị cho ViWeb Technology với AI hỗ trợ nghiên cứu, soạn thảo và dàn trang. Nguồn và số học đã được kiểm tra khi chuẩn bị; quy trình không có thẩm định kỹ thuật riêng bởi con người hay thử nghiệm trong phòng của ViWeb. Ấn bản nghiên cứu dành cho công chúng; khuyến nghị vẫn chịu các giới hạn bằng chứng ở trên. Báo cáo không tuyên bố đã dùng quy trình bản tin tự động riêng của website.

Cần kiểm tra giấy phép mô hình và điều khoản xử lý dữ liệu khách hàng cho mục đích dự kiến. Tài liệu tham khảo được xếp theo lần trích đầu tiên. Gửi đính chính kèm bằng chứng qua trang liên hệ ViWeb.

Tài liệu tham khảo

Tất cả nguồn được kiểm tra ngày 17/11/2026. Số thứ tự theo lần trích đầu tiên. Trang nhà cung cấp và tài liệu có thể thay đổi; ngày dưới đây giữ rõ loại bằng chứng. Giữ tên gốc của tài liệu để thuận tiện tra cứu.

[1] Generative Artificial Intelligence Profile (NIST AI 600-1)

NIST | 26/07/2024; trang công bố cập nhật 08/04/2026

Khung rủi ro tự nguyện; hỗ trợ nhận diện vấn đề về bịa thông tin, quyền riêng tư và giám sát. Không chứng nhận một triển khai cục bộ.

[2] LLM01:2025 Prompt Injection

OWASP Gen AI Security Project | Ấn bản 2025; trang trực tuyến

Chèn lệnh trực tiếp và gián tiếp; RAG và tinh chỉnh không loại bỏ vấn đề này.

[3] LLM08:2025 Vector and Embedding Weaknesses

OWASP Gen AI Security Project | Ấn bản 2025; trang trực tuyến

Truy xuất có xét quyền, tính toàn vẹn của nguồn và rò rỉ giữa các ngữ cảnh.

[4] Randomized Controlled Trials for Phishing Triage Agent

James Bono, Microsoft; bản thảo arXiv v1 | 17/11/2025

Tuyển 167 chuyên viên; loại 13 câu trả lời; hàng đợi email chọn lọc và lỗi tác tử được mô phỏng. Kết quả của sản phẩm cụ thể không chứng minh hiệu năng mô hình cục bộ.

[5] Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity

METR | 10/07/2025

16 lập trình viên và 246 tác vụ; nghiên cứu lập trình trong quá khứ. Đọc cùng bản cập nhật năm 2026, không dùng làm ước lượng hiện tại cho mọi trường hợp.

[6] We are Changing our Developer Productivity Experiment Design

METR | 24/02/2026

Thí nghiệm về sau gặp hiệu ứng chọn mẫu. Không xác lập ước lượng năng suất hiện tại rõ ràng cho toàn bộ nhóm người dùng.

[7] Annual Review 2026: The impact of AI on cyber threat

Trung tâm An ninh Mạng Quốc gia Anh | 2026; kỳ tổng kết 01/09/2025-31/08/2026

Đánh giá tình báo, không phải tỷ lệ sự cố toàn cầu được đo hay nghiên cứu quy trách nhiệm cho AI cục bộ.

[8] LLM06:2025 Excessive Agency

OWASP Gen AI Security Project | Ấn bản 2025; trang trực tuyến

Giới hạn công cụ, quyền và mức tự chủ; thực thi kiểm soát quyền tại hệ thống phía sau.

Tài liệu tham khảo / tiếp theo

[9] LLM05:2025 Improper Output Handling

OWASP Gen AI Security Project | Ấn bản 2025; trang trực tuyến

Đầu ra được sinh cần xác thực thông thường, mã hóa theo ngữ cảnh và ranh giới thực thi an toàn.

[10] Transformers Security Policy

Hugging Face | Tài liệu kho mã trực tuyến

Ưu tiên safetensors, cố định bản sửa đổi đã rà soát và kiểm tra mọi mã được bật qua trust_remote_code.

[11] LLM03:2025 Supply Chain

OWASP Gen AI Security Project | Ấn bản 2025; trang trực tuyến

Mô hình, adapter, dữ liệu, runtime và phụ thuộc mở rộng chuỗi cung ứng.

[12] LLM10:2025 Unbounded Consumption

OWASP Gen AI Security Project | Ấn bản 2025; trang trực tuyến

Giới hạn yêu cầu, tài nguyên, độ dài hàng đợi và vòng lặp công cụ.

[13] Deploying AI Systems Securely

NSA, CISA, FBI và đối tác quốc tế; bản do FBI lưu trữ | Tháng 4/2024, phiên bản 1.0

Hướng dẫn bảo mật triển khai; hỗ trợ gia cố, cô lập, giám sát và ứng phó sự cố. Không phải chứng chỉ tuân thủ.

[14] llama.cpp

ggml-org | Tài liệu dự án trực tuyến

Suy luận cục bộ với nhiều backend tăng tốc. Có hỗ trợ backend không chứng minh tốc độ hay tính năng ngang nhau.

[15] MLX documentation

Apple ML research | Tài liệu trực tuyến

Framework học máy cho Apple silicon, bộ nhớ dùng chung và ví dụ LLM.

[16] Ollama Local FAQ

Ollama | Tài liệu trực tuyến

Cài đặt chỉ cục bộ, hành vi khởi động lại, lắng nghe localhost và kiểm tra mô hình trên GPU. Cài đặt phải khớp bản phát hành đã cài.

Tài liệu tham khảo / tiếp theo

[17] AMD RDNA3.5 system optimization

Tài liệu AMD ROCm | Tài liệu trực tuyến

Cấu hình bộ nhớ APU Ryzen và yêu cầu nhân Linux; kiểm tra ma trận tương thích tương ứng trước khi mua.

[18] NASPO PSS: Apple-branded price list

Apple Mỹ | Tháng 9/2026; trang 1

Dùng cột Consumer (MSRP), không dùng giá hợp đồng giảm giá. Hỗ trợ cấu hình tiêu chuẩn có tên; không bảo đảm thanh toán hay tồn kho.

[19] Framework Desktop DIY configuration: Ryzen AI Max 300

Bộ cấu hình Framework Mỹ | Trang cấu hình trực tuyến

Hệ 395/64GB giá $1,959; hệ 395/128GB giá $3,449. Ổ lưu trữ và hệ điều hành tính thêm; chưa xác nhận độc lập khả năng mua cấu hình được chọn.

[20] EVO-X2 Ryzen AI Max+ 395 mini PC

Cửa hàng GMKtec Mỹ | Trang bán hàng trực tuyến

Biến thể 64GB/1TB được chọn có giá $2,199.99. Không gán giá nổi bật này cho biến thể 128GB; trang có thông báo một biến thể không sẵn có.

[21] MINISFORUM MS-S1 MAX 128GB Max AI Compute Edition

Cửa hàng MINISFORUM Mỹ | Trang bán hàng trực tuyến

128GB/2TB hiển thị $3,959 và $4,949 cùng thông báo đợt giảm giá đã kết thúc. Tín hiệu khuyến mại mâu thuẫn: cần báo giá trước khi mua.

[22] Beelink GTR9 Pro Ryzen AI Max+ 395

Cửa hàng Beelink, đơn vị USD | Trang bán hàng trực tuyến

128GB LPDDR5x hàn trên bo và SSD 2TB: giá giảm hiển thị $4,199, giá đối chiếu $4,699; chưa xác minh tồn kho cuối cùng và tổng giá.

[23] EVO-X5 Pro Ryzen AI Max+ PRO 495

Cửa hàng GMKtec Mỹ | Trang bán hàng trực tuyến

Biến thể 192GB/2TB được chọn hiển thị $6,799. Nhà cung cấp cho biết còn hàng; chưa xác minh việc giao hàng.

[24] Mac mini technical specifications

Apple Mỹ | Trang sản phẩm trực tuyến

Trang hiện tại liệt kê M6 và M5 Pro, tùy chọn bộ nhớ và băng thông. Không phải bằng chứng đo hiệu năng M4.

Tài liệu tham khảo / tiếp theo

[25] AMD Ryzen AI Max+ 395 specifications

AMD | Trang sản phẩm trực tuyến

Bộ nhớ hệ thống tối đa 128GB, LPDDR5x-8000, giao tiếp 256-bit và Radeon 8060S.

[26] Ryzen AI Max+ 395: Breakthrough AI Performance in Thin and Light

Blog nhà cung cấp AMD | 2025

Nhà cung cấp mô tả khả năng phân bổ tới 96GB cho đồ họa ở cấu hình 128GB. Không bảo đảm ứng dụng dùng được mức bộ nhớ đó trên mọi máy OEM hoặc hệ điều hành.

[27] GeForce RTX 50 Series specifications

NVIDIA Mỹ | Trang sản phẩm trực tuyến

RTX 5060 Ti có tùy chọn 16GB; RTX 5070 Ti có 16GB và RTX 5090 có 32GB. Giá khởi điểm quảng cáo không phải báo giá hiện tại cho máy tính hoàn chỉnh.

[28] DGX Spark hardware overview

NVIDIA | Hướng dẫn sử dụng trực tuyến

GB10, CPU Arm, bộ nhớ dùng chung 128GB, băng thông 273GB/s và thông số điện. Năng lực tính toán quảng cáo không phải tốc độ token đo được.

[29] DGX Spark US marketplace listing

NVIDIA | Trang sản phẩm trực tuyến

Trang 128GB/4TB chuyển tới đối tác được ủy quyền và không cung cấp giá Mỹ hiện tại trong nội dung truy xuất. Chỉ đưa vào diện cần báo giá.

[30] AI/LLM benchmarks: Findings

Jeff Geerling và người đóng góp kho mã | Bản README ngày 22/09/2026; các lượt chạy riêng cũ hơn

Tổng hợp lịch sử từ nhiều hệ thống, không phải thử nghiệm mua sắm tháng 11/2026 với điều kiện tương đương. Theo liên kết issue từng hệ để xem cài đặt; không tính xếp hạng mua hàng từ bảng này.

[31] llama-bench performance testing tool

ggml-org | Tài liệu dự án trực tuyến

Thử xử lý prompt và sinh token, lặp phép đo và đầu ra máy đọc được. Vẫn cần chạy lại tác vụ trong ứng dụng thực tế.