Bản tin công nghệ - 19/01/2026

Mở đầu

Trong bối cảnh AI đã bước ra khỏi giai đoạn hype và trở thành lực lượng cốt lõi thúc đẩy chuyển đổi số ở mọi ngành, các tiến bộ gần đây đang mở ra những khả năng chưa từng có cho việc triển khai mô hình lớn trên môi trường tài nguyên hạn chế. Từ việc tách các bộ điều chỉnh LoRA thành mô-đun toàn cục-cục bộ (SDFLoRA) giúp giảm tải truyền tải tới 40 % và duy trì độ chính xác khi áp dụng vi sai khác (DP), đến các giải pháp VLM không cần tiền huấn luyện lại (MHA2MLA-VLM) giảm KV cache hơn 94 %, chúng ta đang chứng kiến một làn sóng tối ưu hoá kiến trúc sâu sắc. Các khung can thiệp thời gian thực như Think-with-Me và Think-Clip-Sample giảm độ dài suy luận tới hơn 80 % và rút ngắn thời gian inference hơn 50 %, trong khi LoRA-cle cung cấp công cụ kiểm toán nhẹ nhàng để phát hiện backdoor và membership. Song song đó, các doanh nghiệp toàn cầu như JPMorgan hay IBM đã chứng minh lợi nhuận thực tế khi áp dụng trợ lý AI và kiến trúc lai giảm bộ nhớ lên tới 10×. Những xu hướng này không chỉ hứa hẹn tiết kiệm chi phí và tài nguyên mà còn đặt nền móng cho việc xây dựng các agent tự động hoá quy trình đa dạng.

Điểm nhấn: SDFLoRA: Selective Dual-Module LoRA for Federated Fine-tuning with Heterogeneous Clients

Đột phá SDFLoRA tách bộ điều chỉnh LoRA thành mô-đun toàn cục và mô-đun cục bộ, cho phép chỉ gộp các hướng chuyển giao qua “selective stacking” trong khi giữ riêng các đặc thù khách hàng. Nhờ cách tiếp cận này, mô hình đạt cải thiện độ chính xác trung bình 3-5 % trên các benchmark GLUE so với FedAvg và giảm sai số chuẩn dưới 0.1 so với phương pháp padding. Khi áp dụng bảo mật vi sai khác (DP) với $\varepsilon = 1$, SDFLoRA duy trì độ chính xác cao hơn 7 % so với việc gộp toàn bộ tham số. Đối với FPT, kiến trúc giảm tải truyền tải lên tới 40 % và hỗ trợ đa dạng thiết bị nhờ xếp hạng LoRA không đồng nhất, mở đường cho triển khai LLM trên môi trường tài nguyên hạn chế và tuân thủ quy định dữ liệu.

Ngày xuất bản: 16 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.11219v1

Tin nhanh công nghệ (7 bài)

News #1: 🐱 🎙️ We're LIVE now talking AI's impact across 16 industries...

AI đã chuyển từ giai đoạn hype sang thực thi thực tế trong nhiều ngành. Ví dụ, JPMorgan triển khai trợ lý GPT-4 cho 200 000 nhân viên, giảm chi phí trung tâm cuộc gọi 30 % và thời gian nghiên cứu phân tích 83 %. Trong xây dựng, công cụ dự đoán AI rút ngắn thời gian lấy mẫu 76 %. IBM Granite 4.0 dùng kiến trúc lai giảm bộ nhớ và độ trễ tới 10×, cho phép chạy trên laptop thay $40K GPU.

Ngày xuất bản: 16 Jan, 2026 **Nguồn:** NeuronDaily **URL:** https://www.theneurondaily.com/p/we-re-live-now-talking-ai-s-impact-across-16-industries

News #2: 🐱 Claude: Months of research, done in 20

minutes

Claude (Anthropic) đã hoàn thành nghiên cứu gen kéo dài hàng tháng chỉ trong 20-35 phút, giảm thời gian phân tích từ 3 tuần xuống dưới giờ; OpenAI ký hợp đồng $10 tỷ với Cerebras để cung cấp 750 MW siêu tốc tính toán tới 2028; Higgsfield huy động $80 triệu, đạt doanh thu ARR $200 triệu, tăng gấp đôi trong 2 tháng.

Ngày xuất bản: 16 Jan, 2026 **Nguồn:** NeuronDaily **URL:** https://www.theneurondaily.com/p/claude-months-of-research-done-in-20-minutes

News #3: Why Autonomous AI Agents Will Redefine Enterprise IT Strategy

Sự chuyển đổi từ trợ lý AI dạng copilot sang các agent tự động là thay đổi mới nhất: các agent không cần lệnh liên tục, có khả năng thực thi chuỗi công việc, truy cập dữ liệu thời gian thực và ghi nhớ liên phiên. Đối với FPT, việc xây dựng nền tảng dữ liệu thống nhất và khung quản trị cho agent sẽ cho phép tích hợp nhanh CRM, ERP, hỗ trợ bán hàng và marketing, nâng cao tốc độ cá nhân hoá khách hàng và tạo lợi thế cạnh tranh.

Ngày xuất bản: 16 Jan, 2026 **Nguồn:** TechRepublic **URL:** https://www.techrepublic.com/article/news-autonomous-ai-agents-redefine-enterprise-it-strategy/

News #4: Introducing ChatGPT Go, now available worldwide

ChatGPT Go đã được triển khai toàn cầu, mở rộng quyền truy cập vào mô hình GPT-5.2 Instant với giới hạn sử dụng cao hơn và bộ nhớ dài hơn, giúp giảm chi phí sử dụng AI tiên tiến. Đối với FPT, việc tiếp cận dễ dàng hơn vào công nghệ ngôn ngữ lớn cho phép tăng tốc phát triển các giải pháp SaaS, chatbot và tự động hoá quy trình nội bộ mà không gặp rào cản tài chính.

Ngày xuất bản: 16 Jan, 2026 **Nguồn:** OpenAI News **URL:** https://openai.com/index/introducing-chatgpt-go

News #5: Daily Tech Insider Unpacks AI Assistants' Leap From Code to Chores

Google bật Gemini Personal Intelligence cho Gmail để truy vấn email, ảnh & YouTube; Anthropic ra mắt Claude Cowork – trợ lý desktop tự động sắp xếp file & tạo báo cáo; Salesforce triển khai Agentforce Slackbot trên Workspace Business+/Enterprise+ (cuối Feb) để soạn tài liệu & lên lịch họp; Claude for Health đạt chuẩn HIPAA; LG giới thiệu robot CLOiD gập khăn trong 30 s. Đối với FPT, chúng giảm tới 20h/tuần công việc, nâng năng suất & bảo mật dữ liệu.

Ngày xuất bản: 16 Jan, 2026 **Nguồn:** TechRepublic **URL:** https://www.techrepublic.com/article/news-best-of-dti-jan-12-16-2026/

News #6: Amazon Starts Auto-Upgrading Prime Members to Alexa+ Without Warning

Amazon tự động nâng cấp một số thành viên Prime lên Alexa+, trợ lý AI sinh sinh mà không hỏi ý kiến. Phiên bản mới trả lời dài hơn, duy trì ngữ cảnh trên các thiết bị Echo và Fire TV. Đối với FPT, việc triển khai AI mặc định gây lo ngại về quyền lựa chọn và niềm tin khách hàng; chúng ta cần cơ chế bật/tắt rõ ràng khi tích hợp AI vào dịch vụ đám mây.

Ngày xuất bản: 16 Jan, 2026 **Nguồn:** TechRepublic **URL:** https://www.techrepublic.com/article/news-amazon-auto-enables-alexa-plus-prime-members/

News #7: 😼 ChatGPT ads are here...

OpenAI sẽ chèn quảng cáo vào tầng miễn phí của ChatGPT và cả gói "ChatGPT Go" giá $8/tháng, đồng thời duy trì các gói trả phí không có quảng cáo. Đây là phản ứng trước mức tiêu hao tài nguyên khổng lồ: OpenAI đã tiêu tốn khoảng $5 tỷ năm 2024 và ~$9 tỷ năm 2025 trên doanh thu $13 tỷ (tỷ lệ cháy ~70%). Đối với FPT, việc này báo hiệu xu hướng "AI-ads" sẽ lan rộng, yêu cầu chúng ta chuẩn bị mô hình doanh thu hỗ trợ quảng cáo cho các dịch vụ AI nội bộ.

Ngày xuất bản: 18 Jan, 2026 **Nguồn:** NeuronDaily **URL:** https://www.theneurondaily.com/p/chatgpt-ads-are-here

Nghiên cứu khoa học nổi bật (4 bài)

Article #1: MHA2MLA-VLM: Enabling DeepSeek's Economical Multi-Head Latent Attention across Vision-Language Models

MHA2MLA-VLM cho phép chuyển các VLM hiện có sang kiến trúc MLA mà không cần tiền huấn luyện lại. Hai đổi mới chính: (1) chiến lược Partial-RoPE thích ứng đa mô-đun, giữ lại các chiều quay quan trọng cho ảnh và văn bản; (2) SVD phân tách mô-đun giảm KV cache lên tới 94.6 % với chỉ mất <2 % độ chính xác. Quá trình tinh chỉnh chỉ cập nhật ~10 % tham số và dùng 1.8 B token, giảm chi phí GPU đáng kể cho FPT.

Ngày xuất bản: 16 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.11464v1

Article #2: Beyond Model Scaling: Test-Time Intervention for Efficient Deep Reasoning

Think-with-Me giới thiệu khung can thiệp thời gian thực cho Large Reasoning Models, dùng liên từ chuyển tiếp làm điểm dừng và nhận phản hồi từ LLM hoặc con người. Nhờ đó độ dài suy luận giảm tới 81 % (AIME24: 322 token vs 1199) và độ chính xác tăng 7,19 % so với QwQ-32B. Đối với FPT, giảm token và latency tiết kiệm tài nguyên đám mây, nâng năng suất AI và mở rộng ứng dụng trong bài toán phức tạp.

Ngày xuất bản: 16 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.11252v1

Article #3: Think-Clip-Sample: Slow-Fast Frame Selection for Video Understanding

Think-Clip-Sample (TCS) mang lại hai đổi mới: **Multi-Query Reasoning** – tự động sinh tới 4

truy vấn phụ từ câu hỏi để khai thác đa góc nhìn (đối tượng, hành động, cảnh); và **Clip-level Slow-Fast Sampling** – phân bổ khoảng 75 % khung vào các clip có độ tương đồng cao và 25 % khung đều khắp video còn lại. Nhờ đó các mô hình MLLM tăng độ chính xác lên tới **6.9 %** và giảm thời gian suy luận hơn **50 %** ($\approx 2\times$ tốc độ), giúp FPT Smart Cloud xử lý video dài hiệu quả hơn, tiết kiệm tài nguyên và nâng cao dịch vụ AI cho khách hàng.

Ngày xuất bản: 16 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.11359v1

Article #4: LoRA as Oracle

LoRA-cle (LoRA as Oracle) giới thiệu khung kiểm toán mô hình dựa trên **adapter low-rank**: chỉ cần gắn LoRA vào mô hình đóng băng, sau đó đo độ biến đổi và độ ổn định của các cập nhật để phát hiện **backdoor** và **membership**. Kết quả cho thấy độ chính xác > 90 % trên cả hai nhiệm vụ (ví dụ: 95 % F1 cho CIFAR-10) và tiêu thụ chỉ 6 W/6 GB GPU – đủ chạy trên máy tính cá nhân. Đối với FPT, công nghệ này cho phép rà soát nhanh, chi phí thấp các mô hình pre-trained trước khi triển khai dịch vụ đám mây hoặc AI nội bộ, giảm rủi ro dữ liệu rò rỉ và tấn công nhúng.

Ngày xuất bản: 16 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.11207v1

Kết luận

Những đổi mới vừa nêu – từ mô-đun hoá LoRA, tối ưu hoá VLM đa mô-đun đến khung kiểm toán dựa trên adapter – tạo nên một hệ sinh thái AI linh hoạt hơn cho FPT Smart Cloud. Chúng giúp chúng ta giảm băng thông truyền tải lên tới 40 %, khai thác hiệu quả tài nguyên GPU trên thiết bị đầu cuối và nâng cao an ninh dữ liệu trước các mối đe dọa tiêm nhiễm mô hình. Đồng thời, việc tích hợp agent tự động hoá và các dịch vụ AI dạng SaaS sẽ tăng tốc phát triển CRM/ERP nội bộ cũng như mở rộng dịch vụ cho khách hàng mà không gặp rào cản chi phí. Để tận dụng tối đa tiềm năng này, FCI cần tập trung vào xây dựng nền tảng dữ liệu thống nhất, chuẩn hoá quy trình kiểm toán mô hình và triển khai các giải pháp modular trên cả cloud và edge.

Tiêu đề	Ngày xuất bản	URL
SDFLoRA: Selective Dual-Module LoRA for Federated Fine-tuning with Heterogeneous Clients	16 Jan, 2026	Link
MHA2MLA-VLM: Enabling DeepSeek's Economical Multi-Head Latent Attention across Vision-Language Models	16 Jan, 2026	Link
Beyond Model Scaling: Test-Time Intervention for Efficient Deep Reasoning	16 Jan, 2026	Link
Think-Clip-Sample: Slow-Fast Frame Selection for Video Understanding	16 Jan, 2026	Link
LoRA as Oracle	16 Jan, 2026	Link
🐱🎙️ We're LIVE now talking AI's impact across 16 industries...	16 Jan, 2026	Link
🐱 Claude: Months of research, done in 20 minutes	16 Jan, 2026	Link
Why Autonomous AI Agents Will Redefine Enterprise IT Strategy	16 Jan, 2026	Link

Introducing ChatGPT Go, now available worldwide	16 Jan, 2026	Link
Daily Tech Insider Unpacks AI Assistants’ Leap From Code to Chores	16 Jan, 2026	Link
Amazon Starts Auto-Upgrading Prime Members to Alexa+ Without Warning	16 Jan, 2026	Link
🐱 ChatGPT ads are here...	18 Jan, 2026	Link

Bản tin được tạo tự động bởi hệ thống FCI News Agents.