Bản tin công nghệ - 19/01/2026

Mở đầu

Trong tuần này, chuỗi các đột phá vừa được công bố – từ việc nén mô hình Vision-Language bằng kiến trúc Multi-Head Latent Attention (MHA2MLA-VLM) chỉ với 10 % tham số fine-tune, tới LoRA kép trong SDFLoRA giúp triển khai Federated Learning trên thiết bị đa dạng – đang vẽ lại bức tranh khả năng mở rộng AI trên hạ tầng đám mây nội bộ của FPT. Các cải tiến như Think-Clip-Sample giảm hơn 50 % thời gian suy luận video và Think-with-Me cắt giảm 81 % độ dài token cho Large Reasoning Model không chỉ nâng cao độ chính xác mà còn hạ thấp chi phí tính toán. Song song đó, các công cụ an ninh như LoRA-oracle và xu hướng AI agents tự động hoá quy trình đang mở ra cơ hội kiểm soát rủi ro và cung cấp dịch vụ AI-native cho khách hàng doanh nghiệp.

Điểm nhấn: MHA2MLA-VLM: Enabling DeepSeek's Economical Multi-Head Latent Attention across Vision-Language Models

MHA2MLA-VLM cho phép chuyển đổi các mô hình Vision-Language (VLM) hiện có sang kiến trúc Multi-Head Latent Attention (MLA) chỉ với $\approx 10\,\%$ tham số được fine-tune và $< 0.002\,\%$ dữ liệu huấn luyện, giảm kích thước KV cache tới 94.6 % mà hiệu năng chỉ giảm ≤ 1.3 điểm trên các benchmark đa dạng. Phương pháp “modality-adaptive partial-RoPE” giữ lại các chiều quay quan trọng cho cả ảnh và văn bản, trong khi “modality-decoupled SVD” nén riêng không gian KV của từng modality, giảm mất mát lên tới 35 % ở lớp sâu. Đối với FPT, việc giảm bộ nhớ GPU và thời gian inference (giảm 59 % thời gian chuyển đổi) mở đường cho triển khai VLM quy mô lớn trên hạ tầng đám mây nội bộ, tăng khả năng phục vụ khách hàng AI đa phương tiện mà không cần đầu tư thêm phần cứng.

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

Tin nhanh công nghệ (6 bài)

News #1: Why Autonomous AI Agents Will Redefine Enterprise IT Strategy

AI copilots đang chuyển sang giai đoạn mới: các autonomous AI agents tự động thực thi chuỗi công việc, phân tích dữ liệu và điều phối quy trình đa hệ thống mà không cần lệnh. Đối với FPT, xây dựng nền tảng dữ liệu thống nhất, khung quản trị quyền truy cập cho agent và tích hợp sâu CRM-ERP-marketing sẽ cho phép cung cấp dịch vụ AI-native cho ngân hàng, 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 #2: Introducing ChatGPT Go, now available worldwide

ChatGPT Go đã được triển khai toàn cầu, cho phép truy cập ngay vào mô hình GPT-5.2 Instant với giới hạn sử dụng cao hơn và bộ nhớ dài hơn. Điều này giúp FPT giảm chi phí triển khai AI tiên tiến, mở rộng quy mô dịch vụ khách hàng và tăng tốc độ phát triển các giải pháp đám mây nội bộ.

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

News #3: Daily Tech Insider Unpacks AI Assistants' Leap From Code to Chores

Google mở Gemini Personal Intelligence trên Gmail để truy vấn email、ảnh、YouTube；Anthropic phát hành Cowork—desktop agent tự động đổi tên file、lập bảng chi phí；Salesforce triển khai Agentforce Slackbot cắt giảm 2-20 giờ làm việc mỗi tuần；Claude for Health đạt chuẩn HIPAA hỗ trợ đối tác tăng tốc 10×；LG trình diễn robot CLOiD gấp khăn trong 30 giây。FPT có thể dùng AI để cải thiện dịch vụ & tăng cường an ninh

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

News #4: A business that scales with the value of intelligence

OpenAI mở rộng mô hình kinh doanh từ thuê bao sang API cho nhà phát triển,chúng tôi thêm quảng cáo trong ChatGPT,tích hợp thương mại và cung cấp sức mạnh tính toán. Điều này chứng minh khả năng tăng doanh thu qua việc khai thác trí tuệ nhân tạo. Đối với FPT,mô hình đa kênh đề xuất cách xây dựng dịch vụ AI có thể bán nền tảng,chạy quảng cáo và giải pháp tính toán để giảm rủi ro phụ thuộc vào nguồn duy nhất。

Ngày xuất bản: 18 Jan, 2026 **Nguồn:** OpenAI News **URL:** https://openai.com/index/a-business-that-scales-with-the-value-of-intelligence

News #5: 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 dựa trên mô hình ngôn ngữ lớn, thay đổi giọng và phản hồi trên các thiết bị Echo, Fire TV và tablet mà không hỏi ý kiến. Việc bật mặc định gây tranh cãi về quyền lựa chọn – cảnh báo FPT cần có cơ chế bật/tắt rõ ràng khi triển khai dịch vụ AI nội bộ.

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

News #6: London Mayor Urges Stronger Regulation as AI Threatens Workforce

London công bố đào tạo AI miễn phí cho mọi công dân và thành lập Taskforce AI để bảo vệ việc làm; 56 % người lao động dự đoán ảnh hưởng trong năm tới và 70 % kỹ năng sẽ thay

đổi đến 2030. Đối với FPT, xu hướng này tăng nhu cầu giải pháp đám mây an toàn, tư vấn quản trị AI và nâng cao năng lực nội bộ.

Ngày xuất bản: 16 Jan, 2026 **Nguồn:** TechRepublic **URL:** https://www.techrepublic.com/article/news-london-mayor-ai-warning/

Nghiên cứu khoa học nổi bật (4 bài)

Article #1: SDFLoRA: Selective Dual-Module LoRA for Federated Fine-tuning with Heterogeneous Clients

SDFLoRA giới thiệu LoRA kép (global + local module) với 'selective stacking' để gộp cập nhật low-rank đa dạng mà không ép đồng nhất toàn bộ tham số. Giúp FPT triển khai FL LLM trên thiết bị tài nguyên khác nhau, giảm băng thông và bảo vệ dữ liệu bằng cách chỉ thêm nhiễu DP vào module global. Thử nghiệm GLUE tăng tới 6.71 % MNLI, 4.26 % RTE; ở $\varepsilon = 1$ vẫn giữ > 84 % accuracy。

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

Article #2: Think-Clip-Sample: Slow-Fast Frame Selection for Video Understanding

Thay đổi mới: Think-Clip-Sample (TCS) giới thiệu **Multi-Query Reasoning** và **Clip-level Slow-Fast Sampling** – hai cơ chế không cần huấn luyện để chọn khung hình cho video dài. Quan trọng với FPT vì nó nâng độ chính xác lên tới **6.9 %** (ví dụ MiMo-VL-7B trên MLVU) và giảm thời gian suy luận hơn **50 %**, giúp các giải pháp AI video của công ty xử lý nội dung giờ-giờ hiệu quả hơn, tiết kiệm tài nguyên đám mây và tăng tính cạnh tranh.

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

Article #3: Beyond Model Scaling: Test-Time Intervention for Efficient Deep Reasoning

Think-with-Me là khung can thiệp thời gian chạy cho Large Reasoning Model, dùng liên từ chuyển tiếp làm điểm dừng và nhận phản hồi từ LLM proxy hoặc con người; mô hình được huấn luyện lại bằng GRPO. Với FPT, giảm overthinking/overshoot giúp cắt chi phí token và độ trễ, nâng hiệu suất AI đám mây. Kết quả: AIME24 ↑7.19% so với QwQ-32B đồng thời độ dài suy luận ↓81% (từ 1199 → 322 token).

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

Article #4: LoRA as Oracle

LoRA-oracle mới cho phép phát hiện backdoor và thực hiện membership inference chỉ bằng cách gắn adapter low-rank vào mô hình cố định. Đạt > 90 % độ chính xác trên ResNet, VGG, ViT và DenseNet, tiêu thụ tối đa 6 GB VRAM và < 270 W – vượt trội so với các phương pháp thường gây OOM. Giúp FCI nhanh kiểm tra an toàn mô hình tiền-huấn luyện mà không cần dữ liệu gốc.

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

Kết luận

Những tiến bộ vừa nêu cho thấy xu hướng ba mặt: tối ưu tài nguyên (bộ nhớ GPU, băng thông), tăng cường độ tin cậy (phát hiện backdoor, giảm overthinking) và mở rộng khả năng tự động hoá (autonomous agents, multi-query reasoning). Đối với FPT, việc áp dụng MHA2MLA-VLM và SDFLoRA sẽ giảm đáng kể chi phí hạ tầng khi triển khai VLM và LLM quy mô lớn; Think-Clip-Sample cùng Think-with-Me hứa hẹn cải thiện dịch vụ AI video và xử lý ngôn ngữ phức tạp mà không cần đầu tư phần cứng mới. Chúng ta nên lập kế hoạch thí điểm các giải pháp này trên môi trường cloud nội bộ, đồng thời xây dựng khung quản trị an ninh mô hình để nhanh chóng đưa ra các sản phẩm AI an toàn và cạnh tranh hơn.

Tiêu đề	Ngày xuất bản	URL
MHA2MLA-VLM: Enabling DeepSeek's Economical Multi-Head Latent Attention across Vision-Language Models	16 Jan, 2026	Link
SDFLoRA: Selective Dual-Module LoRA for Federated Fine-tuning with Heterogeneous Clients	16 Jan, 2026	Link
Think-Clip-Sample: Slow-Fast Frame Selection for Video Understanding	16 Jan, 2026	Link
Beyond Model Scaling: Test-Time Intervention for Efficient Deep Reasoning	16 Jan, 2026	Link
LoRA as Oracle	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
A business that scales with the value of intelligence	18 Jan, 2026	Link
Amazon Starts Auto-Upgrading Prime Members to Alexa+ Without Warning	16 Jan, 2026	Link
London Mayor Urges Stronger Regulation as AI Threatens Workforce	16 Jan, 2026	Link

---