Bản tin công nghệ - 19/01/2026

Mở đầu

Tuần này chúng ta chứng kiến một loạt tiến bộ quan trọng trong lĩnh vực AI và điện toán đám mây: SDFLoRA giới thiệu kiến trúc dual-module cho federated learning giúp giảm đáng kể mất mát hữu ích và nâng cao độ chính xác trên GLUE; ChatGPT Go mở rộng toàn cầu với GPT-5.2 Instant giảm chi phí tiếp cận mô hình mạnh; MHA2MLA-VLM nén KV cache lên tới 94.6 % và chỉ cần tinh chỉnh 10 % tham số; công nghệ HORSE cải thiện tốc độ và độ đặc thù của transformer bằng hypernetwork trên dư lượng. Đồng thời, LoRA-Oracle cung cấp giải pháp phát hiện backdoor và membership inference nhẹ nhàng; SD-RAG đưa ra khung Selective Disclosure tăng Privacy lên 58 %; các tác nhân AI tự động đang thay thế trợ lý copilot trong quy trình KYC và marketing; và xu hướng quy định cùng mô hình kinh doanh đa dạng của OpenAI, Google, Amazon,… mở ra cơ hội mới cho FPT Smart Cloud.

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

SDFLoRA mang lại cách tiếp cận “dual-module” cho LoRA trong môi trường federated learning: mỗi client sở hữu một **global module** để chia sẻ kiến thức và một **local module** giữ riêng các đặc thù cá nhân. Nhờ **selective stacking** chỉ gộp global module, phương pháp giải quyết triệt để vấn đề *rank heterogeneity* mà không ép buộc các hướng cập nhật riêng biệt phải đồng nhất. Khi áp dụng DP-SGD chỉ vào global module, độ mất mát hữu ích giảm đáng kể; trên GLUE, SDFLoRA đạt $+\mathbf{6.7\,\%}$ độ chính xác MNLI và $+\mathbf{4.3\,\%}$ RTE so với zero-padding, đồng thời duy trì hiệu suất khi $\varepsilon = 1$ (privacy mạnh). Đối với FPT, giải pháp này cho phép triển khai nhanh LLM trên thiết bị đầu cuối đa dạng (y tế, tài chính) với chi phí truyền tải thấp, bảo mật dữ liệu tốt hơn và khả năng cá nhân hoá cao—một bước tiến chiến lược cho nền tảng AI đám mây của công ty.

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

Tin nhanh công nghệ (6 bài)

News #1: Introducing ChatGPT Go, now available worldwide

ChatGPT Go đã được triển khai toàn cầu, cho phép người dùng truy cập GPT-5.2 Instant, tăng giới hạn sử dụng và mở rộng bộ nhớ lưu trữ ngữ cảnh. Điều này giảm chi phí tiếp cận AI tiên tiến, giúp FPT nhanh chóng tích hợp mô hình mạnh hơn vào các giải pháp đám mây và dịch vụ AI nội bộ, nâng cao năng lực cạnh tranh và đáp ứng nhu cầu khách hàng Việt Nam.

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

News #2: Why Autonomous AI Agents Will Redefine Enterprise IT Strategy

Trong 2 năm qua CIO chuyển từ trợ lý AI (copilot) sang các tác nhân AI tự động, cho phép thực thi công việc mà không cần lệnh liên tục và có bộ nhớ liên phiên. Tác nhân hỗ trợ quy trình KYC ngân hàng, bán hàng và marketing. Đối với FPT, xây dựng nền dữ liệu thống nhất và khung quản trị cho tác nhân sẽ tạo lợi thế cạnh tranh và mở rộng dịch vụ AI doanh nghiệp.

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

News #3: London Mayor Urges Stronger Regulation as AI Threatens Workforce

Sir Sadiq Khan công bố đào tạo AI cho Londoners và thành lập Taskforce bảo vệ việc làm. 56 % người lao động dự đoán ảnh hưởng trong năm tới; 70 % kỹ năng sẽ thay đổi đến 2030. Đối với FPT, xu hướng quy định nghiêm ngặt và nhu cầu nâng cao năng lực AI tạo cơ hội cung cấp dịch vụ tư vấn chính sách, đào tạo doanh nghiệp và giải pháp an ninh AI ở châu Âu.

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

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, quảng cáo trong ChatGPT, thương mại điện tử và phí tính toán. Sự đa dạng hoá nguồn thu này phản ánh tiềm năng lợi nhuận khi mức độ chấp nhận AI tăng. Đối với FPT, đây là ví dụ để thiết kế gói AI-cloud linh hoạt, khai thác cả khách hàng doanh nghiệp và người dùng cuối.

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

Google bật Personal Intelligence cho Gmail—bot tra cứu email và ảnh/video. Anthropic ra Cowork macOS tự động đổi tên file και lập bảng chi phí sandbox. Salesforce đưa Agentforce vào Slack , giảm 2-20 giờ công việc tuần tới。Claude for Health đáp ứng HIPAA και nhanh hơn 10× xử lý hồ sơ y tế。LG demo robot CLOiD gập khăn dalam 30 giây。FPT có thể dùng tính năng này để xây dựng AI trợ lí doanh nghiệp an toàn。

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 đã bật tính năng Auto-Upgrade chuyển nhóm người dùng Prime sang Alexa+ — phiên bản generative-AI với tiếng nói mới dài hơn & khả năng giữ bối cảnh đa bước — ngay

trong mọi loa Echo & thiết bị Fire mà không hỏi ý kiến trước. Điều này nhấn mạnh nhu cầu quản trị quyền tùy chỉnh & bảo mật dữ liệu cá nhân mà FPT phải cân nhắc khi phát triển tính năng tương đồng trong nền tảng Cloud & Edge.

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

Nghiên cứu khoa học nổi bật (4 bài)

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

Đổi kiến trúc VLM sang Multi-Head Latent Attention (MLA) bằng MHA2MLA-VLM cho phép nén KV cache tới 94.6 % (ví dụ Qwen2.5-VL giảm bộ nhớ 94.64 %) và chỉ cần tinh chỉnh ∼10 % tham số với dữ liệu <0.01 % token, giảm thời gian chuyển đổi 59 %. Điều này giảm chi phí GPU và tăng khả năng triển khai dịch vụ đa phương tiện của FPT Smart Cloud.

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

Article #2: Hierarchical Orthogonal Residual Spread for Precise Massive Editing in Large Language Models

HORSE đưa ra cách lan truyền dư lượng phân cấp và trực giao trên các lớp transformer, chỉnh sửa ở mức token bằng hypernetwork học trên dư lượng thay vì toàn bộ trọng số. Kết quả đạt tốc độ 8.85 s cho 100 mẫu và cải thiện trung bình $+6.26$ % hiệu suất, $+10$ 12 % độ đặc thù. Đối với FPT, công nghệ này giảm rủi ro sai lệch mô hình và hỗ trợ cập nhật kiến thức nhanh trong AI

Ngày xuất bản: 16 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.11441v1

Article #3: LoRA as Oracle

LoRA-Oracle mới cho phép phát hiện backdoor và thực hiện membership inference chỉ bằng cách gắn adapter LoRA thấp-rank vào mô hình đã triển khai, không cần dữ liệu gốc hay tái huấn luyện toàn bộ. Đạt >90 % độ chính xác trên 4 000 thí nghiệm, tiêu thụ tối đa 6 GB bộ nhớ – giảm tới 2-3 lần so với các phương pháp hiện có. Giúp FCI nhanh chóng kiểm tra an toàn AI trong môi trường tài nguyên hạn chế.

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

Article #4: SD-RAG: A Prompt-Injection-Resilient Framework for Selective Disclosure in Retrieval-Augmented Generation

SD-RAG đưa ra khung Selective Disclosure cho RAG: tách làm sạch dữ liệu trước khi sinh câu trả lời và dùng đồ thị để gắn ràng buộc bảo mật bằng ngôn ngữ tự nhiên. Nhờ đó điểm Privacy tăng tới 58 % (0.50→0.83) và hệ thống chống được prompt-injection. Đối với FPT

Smart Cloud, giải pháp bảo vệ dữ liệu nhạy cảm của khách hàng, đáp ứng GDPR/PDPA và nâng cao độ tin cậy dịch vụ AI đám mây

Ngày xuất bản: 16 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.11199v1

Kết luận

Những đột phá này đồng thời củng cố ba trụ cột chiến lược cho FPT: (1) tối ưu hoá triển khai LLM trên thiết bị đầu cuối đa dạng bằng SDFLoRA và MLA để giảm chi phí truyền tải và GPU; (2) nâng cao an ninh và tuân thủ dữ liệu qua LoRA-Oracle và SD-RAG nhằm đáp ứng GDPR/PDPA cũng như phòng ngừa backdoor; (3) xây dựng nền tảng AI-agent linh hoạt dựa trên các mô hình tự động hoá công việc và học hỏi từ mô hình kinh doanh đa kênh của OpenAI cùng tính năng Personal Intelligence của Google. Việc tích hợp những công nghệ này sẽ giúp FPT duy trì lợi thế cạnh tranh trong thị trường AI đám mây khu vực Đông Nam Á.

Tiêu đề	Ngày xuất bản	URL
SDFLoRA: Selective Dual-Module LoRA for Federated Fine-tuning with Heterogeneous Clients	16 Jan, 2026	Link
Introducing ChatGPT Go, now available worldwide	16 Jan, 2026	Link
MHA2MLA-VLM: Enabling DeepSeek's Economical Multi-Head Latent Attention across Vision-Language Models	16 Jan, 2026	Link
Hierarchical Orthogonal Residual Spread for Precise Massive Editing in Large Language Models	16 Jan, 2026	Link
LoRA as Oracle	16 Jan, 2026	Link
SD-RAG: A Prompt-Injection-Resilient Framework for Selective Disclosure in Retrieval-Augmented Generation	16 Jan, 2026	Link
Why Autonomous AI Agents Will Redefine Enterprise IT Strategy	16 Jan, 2026	Link
London Mayor Urges Stronger Regulation as AI Threatens Workforce	16 Jan, 2026	Link
A business that scales with the value of intelligence	18 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

---