Bản tin công nghệ - 19/01/2026

Mở đầu

Trong bối cảnh các doanh nghiệp công nghệ đang tìm cách cân bằng giữa hiệu năng mô hình AI và hạn chế tài nguyên cũng như yêu cầu bảo mật dữ liệu, tuần này chúng ta chứng kiến một loạt tiến bộ đáng chú ý. **SDFLoRA** mang lại giải pháp federated learning linh hoạt cho các thiết bị biên với giảm sai lệch LoRA tới 6 % và chi phí truyền giảm 40 %; đồng thời kiến trúc MultiMax của Gemini duy trì hiệu suất khi mở rộng ngữ cảnh mà chi phí suy luận giảm hơn 10 000-x. Các cải tiến như **MHA2MLA-VLM**, **Think-Clip-Sample** và khung an ninh **LoRA-Oracle** hứa hẹn giảm đáng kể kích thước bộ nhớ và chi phí GPU cho các dịch vụ đa phương tiện và kiểm toán mô hình. Bên cạnh đó, xu hướng tích hợp trợ lý AI tự động hoá quy trình (AI agent), ra mắt các phiên bản mạnh mẽ hơn như **ChatGPT Go**/GPT-5.2 Instant và các chiến lược thương mại mới như quảng cáo trên ChatGPT miễn phí đang mở ra cơ hội và thách thức mới cho FPT trong việc triển khai giải pháp cloud/edge an toàn và hiệu quả.

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

SDFLoRA cung cấp cách tiếp cận mới cho fine-tuning mô hình ngôn ngữ lớn trong môi trường federated khi các client có rank LoRA khác nhau. Nó tách adapter thành **module toàn cục** (được stack và aggregated) và **module cục bộ** (giữ riêng), giảm sai lệch khi gộp LoRA hetero tới 6 % (RTE ↑ 4.26 %, MNLI ↑ 6.71 %) và nâng độ ổn định dưới DP $\varepsilon = 1$ lên 30 %. Chỉ chèn nhiễu DP vào module toàn cực còn giảm chi phí truyền xuống 40 % so với padding đầy đủ. Đối với FPT, kiến trúc này hỗ trợ triển khai FL trên thiết bị biên tài nguyên hạn chế, đồng thời bảo vệ dữ liệu khách hàng mà không làm mất hiệu năng mô hình.

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: Why Autonomous AI Agents Will Redefine Enterprise IT Strategy

Thay đổi: thay vì dùng copilot trả lời theo lời nhắc thì agent AI có thể tự chạy chuỗi nhiệm vụ liên tục mà không cần người nhập câu hỏi mỗi lần. Agent thu thập/phan tich data real-time rồi kết nối CRM-ERP để hoàn thiện workflow. Với FPI chuẩn hoá data lake + thiết lập quyền riêng cho agent sẽ tạo ưu thế trong banking & retail đồng thời nâng hiệu quả nhóm sales

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

News #2: Daily Tech Insider Unpacks AI Assistants' Leap From Code to Chores

Google bật Gemini Personal Intelligence tích hợp Gmail/Photos; Anthropic ra Cowork trên macOS tự động sắp xếp file và tạo báo cáo; Salesforce đưa Agentforce vào Slack giảm 2-20 giờ công việc mỗi tuần; Claude for Health đạt chuẩn HIPAA nhanh hơn tới 10×; LG demo robot CLOiD gập khăn trong 30 giây. Đối với FPT, các trợ lý AI này giúp tự động hoá quy trình, tăng năng suất và đáp ứng tiêu chuẩn bảo mật y tế.

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

News #3: Introducing ChatGPT Go, now available worldwide

Thay đổi: ChatGPT Go được ra mắt toàn cầu, cung cấp quyền truy cập vào GPT-5.2 Instant với giới hạn sử dụng cao hơn và bộ nhớ dài hơn, làm cho AI tiên tiến trở nên rẻ hơn.

Quan trọng với FPT: khả năng dùng mô hình mới giúp tăng tốc độ phát triển các giải pháp AI trên nền cloud/edge, giảm chi phí hạ tầng nhờ mức sử dụng linh hoạt và mở rộng quy mô nhanh chóng.

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

News #4: Our approach to advertising and expanding access to ChatGPT

OpenAI sẽ thử nghiệm hiển thị quảng cáo trên các gói ChatGPT miễn phí và Go tại Mỹ, nhằm tạo nguồn thu mới và giảm chi phí sử dụng cho người dùng toàn cầu đồng thời duy trì bảo mật và chất lượng trả lời. Đối với FPT, mô hình này mở ra cơ hội học hỏi cách cân bằng doanh thu quảng cáo với trải nghiệm AI, giúp chúng ta thiết kế sản phẩm AI giá cả phải chăng hơn và tăng sức cạnh tranh trên thị trường trong nước.

Ngày xuất bản: 16 Jan, 2026 **Nguồn:** OpenAI News **URL:** https://openai.com/index/our-approach-to-advertising-and-expanding-access

News #5: Amazon Starts Auto-Upgrading Prime Members to Alexa+ Without Warning

Amazon tự động chuyển một số thành viên Prime sang Alexa+ – trợ lý AI sinh sinh dựa trên mô hình ngôn ngữ lớn – mà không yêu cầu người dùng đồng ý, thay đổi giọng và cách phản hồi trên mọi thiết bị Echo, Fire TV và tablet. Điều này quan trọng với FPT vì nó nhấn mạnh rủi ro khi triển khai tính năng AI mặc định: mất niềm tin người dùng, yêu cầu cơ chế tắt rõ ràng và kiểm soát hiệu suất (độ trễ, lỗi nhận lệnh) để tránh ảnh hưởng tiêu cực tới sản phẩm và dịch vụ của chúng ta.

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

Lãnh đạo London ra mắt khóa đào tạo AI miễn phí cho cư dân và thành lập Taskforce giám sát tác động việc làm; 56 % người lao động dự đoán ảnh hưởng trong năm tới và 70 % kỹ năng có thể thay đổi đến 2030. Chính phủ Anh mục tiêu đào tạo 7,5 triệu người. Đối với FPT, xu hướng này thúc đẩy nhu cầu nâng cao năng lực AI nội bộ và chuẩn bị giải pháp tuân thủ khi mở rộng dịch vụ tại châu Âu.

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: Building Production-Ready Probes For Gemini

Triển khai probe kích hoạt cho Gemini dùng kiến trúc MultiMax giữ hiệu suất khi đổi từ ngữ cảnh ngắn sang dài; lỗi kiểm thử ≈ 2.5 % – gần bằng Gemini 2.5 Flash (∼²%). Chi phí chạy thấp hơn 10 000-x so với lớp giám sát LLM và tránh tăng chi phí huấn luyện lên 22-x như đào tạo trên dữ liệu dài. Kết hợp cascade giảm chi phí suy luận tới 1/50 LLM mà vẫn duy trì độ chính xác.

Ngày xuất bản: 16 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.11516v1

Article #2: 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, nhờ chiến lược Partial-RoPE đa mô-đun và MD-SVD tách riêng cho hình ảnh và văn bản. Kết quả giảm kích thước KV cache tới 94,6 % (Qwen2.5-VL) và chỉ tinh chỉnh ≈10 % tham số, với mất mát độ chính xác dưới 2 %. Điều này giảm chi phí GPU và thời gian triển khai dịch vụ đa phương tiện của FPT.

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

Article #3: 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 %** (MLVU) và giảm thời gian suy luận hơn **50 %**, cho phép các mô hình đa phương tiện của công ty xử lý video hàng giờ trong môi trường tài nguyên hạn chế, tăng năng suất và khả năng cạnh tranh.

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

Article #4: LoRA as Oracle

Thay đổi mới: giới thiệu **LoRA-Oracle**, một khung kiểm tra an ninh dựa trên mô-đun thích nghi low-rank (LoRA) để phát hiện backdoor và thực hiện membership inference mà không cần dữ liệu gốc hay tái huấn luyện toàn mô hình.

Tầm quan trọng với FCI: cho phép kiểm toán nhanh, nhẹ (≤ 8 % tham số huấn luyện) các mô hình pre-trained được dùng trong dự án cloud/AI, giảm chi phí GPU (6 GB VRAM) và đáp ứng yêu cầu bảo mật dữ liệu khách hàng.
Số liệu: độ chính xác > 90 % cho cả hai nhiệm vụ trên 4 kiến trúc và 4 bộ dữ liệu; tiêu thụ bộ nhớ chỉ 6 GB so với > 15 GB của các phương pháp truyền thống.

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

Kết luận

Những đột phá vừa nêu không chỉ nâng cao khả năng triển khai AI trên thiết bị hạn chế mà còn củng cố nền tảng bảo mật và tối ưu chi phí vận hành – những yếu tố then chốt cho FPT khi mở rộng dịch vụ tại thị trường nội địa và quốc tế. Để tận dụng lợi thế này, chúng ta cần nhanh chóng đánh giá tính khả thi của **SDFLoRA**, **MHA2MLA-VLM** và **LoRA-Oracle** trong các dự án hiện có, đồng thời xây dựng quy trình tích hợp AI agent vào workflow CRM-ERP nhằm tăng năng suất bán hàng. Cuối cùng, việc theo dõi sát sao các mô hình AI mới như **ChatGPT Go** và các chính sách quảng cáo sẽ giúp FPT cân bằng giữa trải nghiệm người dùng và mô hình doanh thu bền vững.

Tiêu đề	Ngày xuất bản	URL
SDFLoRA: Selective Dual-Module LoRA for Federated Fine-tuning with Heterogeneous Clients	16 Jan, 2026	Link
Building Production-Ready Probes For Gemini	16 Jan, 2026	Link
MHA2MLA-VLM: Enabling DeepSeek's Economical Multi-Head Latent Attention across Vision-Language Models	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
Why Autonomous AI Agents Will Redefine Enterprise IT Strategy	16 Jan, 2026	Link
Daily Tech Insider Unpacks AI Assistants’ Leap From Code to Chores	16 Jan, 2026	Link
Introducing ChatGPT Go, now available worldwide	16 Jan, 2026	Link
Our approach to advertising and expanding access to ChatGPT	16 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

---