

Bản tin công nghệ - 19/01/2026

Mở đầu

Trong tuần này các sáng kiến công nghệ của FPT ghi nhận những bước tiến đáng kể về bảo mật và hiệu suất tính toán. Kiến trúc SD-RAG đã nâng cao chỉ số bảo mật lên 58 % (privacy score ≈ 0.83) và giảm rò rỉ thông tin tới 30 % mà vẫn giữ độ đầy đủ trả lời trên 60 %. Đồng thời, mô hình MultiMax cùng Max-of-Rolling-Means Attention Probe và công cụ tìm kiếm AlphaEvolve đạt độ chính xác ngang Gemini 2.5 Flash nhưng chi phí tính toán giảm hơn $10\,000\times$. Các cải tiến khác như MHA2MLA-VLM (nén KV cache 94,64 %, hiệu năng giảm $<1,3\%$), Think-Clip-Sample (tăng độ chính xác +6.9 %, giảm thời gian suy luận $>50\%$) và Think-with-Me (giảm độ dài suy luận 81 %) đều hướng tới việc triển khai AI quy mô lớn trên môi trường cloud-edge với chi phí tối thiểu.

Điểm nhấn: SD-RAG: A Prompt-Injection-Resilient Framework for Selective Disclosure in Retrieval-Augmented Generation

SD-RAG mang lại cách tiếp cận “đầu tiên lọc dữ liệu – sau đó sinh đáp” cho các pipeline Retrieval-Augmented Generation (RAG) của FPT. Bằng cách tách rời việc thực thi chính sách bảo mật khỏi mô hình sinh, hệ thống chỉ truyền cho LLM những đoạn đã được “redact” trước, nên ngay cả khi kẻ tấn công chèn prompt độc hại cũng không thể truy cập nội dung nhạy cảm. Thử nghiệm cho thấy độ bảo mật tăng tới 58 % (privacy score ≈ 0.83) so với phương pháp monolithic hiện tại (≈ 0.50), đồng thời giảm rò rỉ thông tin khoảng 30 % mà vẫn duy trì độ đầy đủ trả lời trên 60 %. Kiến trúc đồ thị cho phép người quản trị nhập quy tắc bảo mật bằng ngôn ngữ tự nhiên và cập nhật linh hoạt mà không cần fine-tune mô hình, đáp ứng nhanh nhu cầu tuân thủ quy định dữ liệu của khách hàng doanh nghiệp và chính phủ.

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

Tin nhanh công nghệ (6 bài)

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

IBM Granite 4.0推出混合層架構，只在部分層保留 KV cache，內存占用與推理延遲均降低約10倍，讓模型可在普通筆記本而非價值40千美元的NVIDIA卡上運行。對於FPT，意味著部署大型語言模型的硬件成本大幅縮減，同時提升服務響應速度，可快速支援客戶的雲edge AI應用，提升競爭力。

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

Google mở Gemini Personal Intelligence cho Gmail cho phép truy vấn lịch sử email, ảnh và YouTube với độ chính xác cao; Anthropic ra mắt Cowork – trợ lý desktop dựa trên Claude giúp tự động đổi tên file, tạo bảng chi phí; Salesforce tích hợp Agentforce vào Slack giảm 2-20 giờ công việc mỗi tuần; Claude for Health đáp ứng chuẩn HIPAA, hỗ trợ dịch báo cáo y tế; LG giới thiệu robot CLOiD gấp khăn trong 30 giây.

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

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

Trong hai năm qua CIO đã chuyển từ trợ lý AI (copilot) sang các **đại lý AI tự động** có khả năng thực thi chuỗi nhiệm vụ mà không cần lệnh liên tục; chúng phân tích dữ liệu liên tục, tích hợp quy trình đa hệ thống và duy trì bộ nhớ lâu dài. Đối với FPT, việc xây dựng nền tảng dữ liệu thống nhất và khung quản trị cho các đại lý này cho phép khai thác triệt để kiến thức nội bộ, nâng cao năng suất bán hàng và đáp ứng quy định trong ngân hàng – một lợi thế cạnh tranh thiết yếu khi hỗ trợ “hàng trăm” tài khoản khách hàng.

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 ra mắt toàn cầu, mở rộng quyền truy cập vào mô hình GPT-5.2 Instant với bộ nhớ kéo dài và hạn mức sử dụng cao hơn. Điều này cho phép các dự án AI của FPT triển khai mô hình tiên tiến mà không gặp rào cản chi phí hoặc giới hạn token, hỗ trợ nhanh chóng xây dựng các giải pháp SaaS, chatbot và phân tích dữ liệu phức tạp trên quy mô quốc tế.

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

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 sinh với phản hồi dài hơn và duy trì ngữ cảnh trên Echo, Fire TV và tablet. Việc mặc định gây phản ứng tiêu cực vì thiếu tùy chọn tắt—người dùng phải nói “Alexa, exit Alexa +” nhưng không hiệu quả. Đối với FPT đây là bài học về quyền lựa chọn khi tích hợp AI và rủi ro ảnh hưởng khách hàng.

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 ra mắt đào tạo AI miễn phí và thành lập Taskforce để giám sát tác động việc làm. 56 % lao động lo ngại mất việc trong năm tới; đến 2030 khoảng 70 % kỹ năng có thể thay đổi. Đối với FPT, hiểu và chuẩn bị nguồn nhân lực AI quy mô này là thiết yếu để duy trì lợi thế cạnh tranh ở châu Âu. Chính phủ Anh còn dự định đào tạo 7,5 triệu người.

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

Thay đổi mới: Giới thiệu kiến trúc MultiMax và Max-of-Rolling-Means Attention Probe, cùng tìm kiếm tự động bằng AlphaEvolve; đạt độ chính xác tương đương Gemini 2.5 Flash nhưng chi phí tính toán giảm hơn $10\,000 \times$ so với mô hình ngôn ngữ đầy đủ.

Tầm quan trọng với FPT: Giải pháp cho phép triển khai giám sát an toàn trên các mô hình lớn mà không tăng đáng kể chi phí hạ tầng – phù hợp với chiến lược “cloud-edge” và bảo mật dữ liệu của chúng ta.

Số liệu nổi bật: Đào tạo dài ngữ cảnh tốn 22 lần chi phí hơn; các probe mới giảm lỗi tổng cộng xuống $\approx 2\%$, trong khi duy trì chi phí chỉ $\leq 1\%$ so với LLM gốc.

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

Article #2: MHA2MLA-VLM: Enabling DeepSeek's Economical Multi-Head Latent Attention across Vision-Language Models

MHA2MLA-VLM chuyển VLM sang MLA mà không cần tiền huấn luyện lại. KV cache được nén tới 94,64 %, hiệu năng chỉ tụt $< 1,3\%$. Chỉ tinh chỉnh $\sim 10\%$ tham số; dùng 1,8 tỷ token thay cho hàng nghìn tỷ token truyền thống. Đối với FPT việc giảm bộ nhớ GPU và chi phí suy luận thúc đẩy triển khai dịch vụ đa phương tiện quy mô lớn trên cloud/edge nhanh hơn và tiết kiệm vốn hạ tầng.

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

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

Think-Clip-Sample (TCS) là khung không cần huấn luyện mới, dùng Multi-Query Reasoning và Clip-level Slow-Fast Sampling để chọn khung video dài. Kết quả cho thấy TCS tăng độ chính xác tới 6.9 % (MLVU) và giảm thời gian suy luận hơn 50 % khi ngân sách khung giảm một nửa. Đối với FPT, công nghệ này hạ chi phí tính toán, rút ngắn triển khai dịch vụ AI video dài và hỗ trợ mở rộng sản phẩm.

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

Article #4: Beyond Model Scaling: Test-Time Intervention

for Efficient Deep Reasoning

Think-with-Me đưa ra khung can thiệp thời gian chạy cho LRM bằng cách dùng liên từ chuyển tiếp làm điểm dừng và nhận phản hồi từ LLM proxy hoặc con người; mô hình được fine-tune với GRPO. Trên AIME24, độ chính xác tăng 7.19 % so với QwQ-32B và độ dài suy luận giảm 81 % (322 token vs 1199 token). Giúp FPT tiết kiệm chi phí tính toán, tối ưu tài nguyên đám mây và tăng khả năng kiểm soát AI doanh nghiệp.

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

Kết luận

Những kết quả trên cho thấy xu hướng tối ưu hoá tài nguyên – từ bảo mật dữ liệu đến tiêu thụ bộ nhớ và chi phí tính toán – đang trở thành động lực chính cho các giải pháp AI của FPT. Đối với chúng ta, việc tích hợp SD-RAG vào pipeline RAG hiện có sẽ củng cố an toàn dữ liệu khách hàng; đánh giá và triển khai MultiMax cùng các probe mới sẽ giúp duy trì năng lực cạnh tranh mà không tăng gánh nặng hạ tầng; áp dụng MHA2MLA-VLM và Think-Clip-Sample sẽ mở rộng dịch vụ đa phương tiện trên cloud/edge một cách nhanh chóng; đồng thời khai thác IBM Granite 4.0 để giảm chi phí phần cứng và đẩy nhanh thời gian phản hồi dịch vụ.

Bên cạnh đó, xây dựng nền tảng dữ liệu thống nhất cho các đại lý AI tự động sẽ khai thác triệt để kiến thức nội bộ và đáp ứng yêu cầu tuân thủ trong ngành tài chính ngân hàng. Cuối cùng, chuẩn bị nguồn nhân lực AI theo xu hướng toàn cầu sẽ giữ vững vị thế của FPT trong môi trường cạnh tranh quốc tế.

Tiêu đề	Ngày xuất bản	URL
SD-RAG: A Prompt-Injection-Resilient Framework for Selective Disclosure in Retrieval-Augmented Generation	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
Beyond Model Scaling: Test-Time Intervention for Efficient Deep Reasoning	16 Jan, 2026	Link
   We're LIVE now talking AI's impact across 16 industries...	16 Jan, 2026	Link
Daily Tech Insider Unpacks AI Assistants' Leap From Code to Chores	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
Amazon Starts Auto-Upgrading Prime Members to Alexa + Without Warning	16 Jan, 2026	Link

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