Bản tin công nghệ - 24/01/2026

Mở đầu

Trong tuần này FPT Smart Cloud ghi nhận một loạt tiến bộ công nghệ có tiềm năng tái cấu trúc cách chúng ta triển khai và vận hành AI trên nền tảng đám mây-edge. Các nghiên cứu mới chỉ ra rằng sau khi huấn luyện các mô hình siêu lớn, phần lớn gradient chỉ thay đổi trong các khối con của ma trận trọng số – cho phép hoán vị thành dạng chặn chéo và thực thi song song qua các sub-module, hứa hẹn rút ngắn thời gian phản hồi tới 50 %. Đồng thời, các sáng kiến như nền tảng no-code/low-code của Airia, phương pháp gộp mô hình đa ngôn ngữ giảm 35 % thời gian huấn luyện và 5.6 % chi phí, hay Generative Application Firewall (GAF) tích hợp bảo mật lớp đa tầng cho LLM đều hướng tới việc tăng tốc triển khai AI an toàn và hiệu quả. Các cải tiến phần cứng và phần mềm – FlashAttention-4 với 1 605 TFLOPS/s, PyraTok tokenizer video đa cấp nâng PSNR lên 35.7 dB và HTC giảm ECE xuống 0.03 – cùng với các giải pháp tích hợp như Codex trong quy trình DevOps và ServiceNow mở rộng truy cập GPT-4 – đang mở ra cơ hội tối ưu tài nguyên máy chủ, giảm chi phí đám mây và mở rộng dịch vụ AI trên mọi môi trường.

Điểm nhấn: Why Inference in Large Models Becomes Decomposable After Training

Phát hiện rằng sau khi huấn luyện các mô hình siêu lớn, **hầu hết các gradient cập nhật chỉ diễn ra trong một tập con địa phương của ma trận tham số**, khiến tới **70 %–80 %** các yếu tố vẫn giữ nguyên phân phối khởi tạo và không tham gia vào quá trình học. Nhờ tính chất này, ma trận trọng số có thể được **phép hoán vị thành dạng chặn chéo** – mỗi khối tương ứng với một sub-operator độc lập – mà không làm thay đổi giao diện đầu vào/đầu ra. Kết quả là hệ thống suy luận chuyển từ một toán tử đơn khối sang **thực thi song song qua các sub-module**, giảm đáng kể chi phí tính toán và độ phức tạp hệ thống. Đối với FPT Smart Cloud, việc khai thác cấu trúc này cho phép tối ưu hoá tài nguyên máy chủ, rút ngắn thời gian phản hồi lên tới **50 %** và mở đường cho kiến trúc dịch vụ “index-routed” linh hoạt trong các môi trường cloud/edge.

Ngày xuất bản: 22 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.15871v1

Tin nhanh công nghệ (10 bài)

News #1: Airia: Enterprise AI orchestration that unifies experimentation, prod, and governance (Sponsor)

Airia ra mắt nền tảng AI doanh nghiệp tích hợp no-code/low-code/pro-code với môi trường thử nghiệm an toàn và guardrails tự động, cho phép kiểm soát sự lan truyền của agent mà không cần trung gian IT. Đây là bước tiến quan trọng giúp FPT rút ngắn chu kỳ triển khai AI cho khách hàng, tăng tính bảo mật và đáp ứng yêu cầu quản trị dữ liệu. Hiện đã được ba tập đoàn lớn (Stryker, BuzzFeed, ArcelorMittal) áp dụng.

Ngày tổng hợp: 24 Jan, 2026 **Nguồn:** TLDR News **URL:** https://airia.com?utm_source=TLDR&utm_medium=Newsletter&utm_campaign=January232026

News #2: Qwen3-TTS Family is Now Open Sourced: Voice Design, Clone, and Generation!

Không có dữ liệu mới được cung cấp, nên không thể xác định thay đổi nào cũng như tầm quan trọng của mục này đối với FPT.

Ngày tổng hợp: 24 Jan, 2026 **Nguồn:** TLDR News **URL:** https://qwen.ai/blog?id=qwen3tts-0115&utm_source=tldrai

News #3: Overcoming Compute and Memory Bottlenecks with FlashAttention-4 on NVIDIA Blackwell

FlashAttention-4 giới thiệu kernel CUDA tối ưu mới, giảm truy cập bộ nhớ và đạt bộ nhớ gần tuyến tính, cho phép tính attention nhanh hơn và xử lý chuỗi dài hơn. Hiệu suất đạt đỉnh 1 605 TFLOPS/s, giảm thời gian huấn luyện và suy luận đáng kể. Đối với FPT, điều này giúp rút ngắn chi phí GPU, tăng tốc triển khai mô hình LLM nội bộ và mở rộng dịch vụ AI.

Ngày tổng hợp: 24 Jan, 2026 **Nguồn:** TLDR News **URL:** https://developer.nvidia.com/blog/overcoming-compute-and-memory-bottlenecks-with-flashattention-4-on-nvidia-blackwell/?utm_source=tldrai

News #4: Small models, big results: Achieving superior intent extraction through decomposition

Đổi mới: phương pháp phân tách (decomposition) cho phép mô hình đa phương tiện nhỏ (≈ 8 B tham số) tóm tắt từng màn hình UI rồi suy ra ý định người dùng từ chuỗi tóm tắt – đạt F1 tương đương mô hình Gemini 1.5 Pro nhưng chi phí và độ trễ chỉ bằng một phần so với mô hình lớn.
Tầm quan trọng với FPT: giảm phụ thuộc vào server, bảo vệ dữ liệu người dùng, hỗ trợ triển khai AI trợ lý trên thiết bị Edge – phù hợp với chiến lược điện toán đám mây-edge và tối ưu chi phí vận hành.

Ngày tổng hợp: 24 Jan, 2026 **Nguồn:** TLDR News **URL:** https://research.google/blog/small-models-big-results-achieving-superior-intent-extraction-through-decomposition/?utm_source=tldrai

News #5: Scaling PostgreSQL to power 800 million ChatGPT users

OpenAI đã mở rộng PostgreSQL lên mức hàng triệu truy vấn/giây bằng cách dùng nhiều replica, cache ở tầng ứng dụng và proxy, rate limiting cho từng tenant và tách biệt workload trên các node riêng. Kết quả giảm độ trễ trung bình < 5 ms và tăng khả năng mở rộng gấp 10 lần so với kiến trúc cũ—cũng là hướng đi quan trọng cho FPT khi xây dựng nền tảng dữ liệu doanh nghiệp.

Ngày xuất bản: 22 Jan, 2026 **Nguồn:** OpenAI News **URL:** https://openai.com/index/scaling-postgresql

News #6: Inside GPT-5 for Work: How Businesses Use

GPT-5

Báo cáo mới cung cấp dữ liệu chi tiết về cách nhân viên đa ngành sử dụng ChatGPT – từ xu hướng tiếp nhận nhanh chóng đến các nhiệm vụ ưu tiên (soạn nội dung, hỗ trợ khách hàng), mô hình sử dụng theo phòng ban và dự báo vai trò AI trong công việc tương lai. Điều này quan trọng với FPT vì nó cho phép chúng ta thiết kế sản phẩm SaaS phù hợp từng bộ phận, tối ưu hoá quy trình nội bộ và định vị chiến lược AI của công ty dựa trên thực tiễn người dùng.

Ngày xuất bản: 22 Jan, 2026 **Nguồn:** OpenAI News **URL:** https://openai.com/business/guides-and-resources/chatgpt-usage-and-adoption-patterns-at-work

News #7: Cisco and OpenAI redefine enterprise engineering with AI agents

Codex – AI software agent do Cisco và OpenAI tích hợp trực tiếp vào quy trình phát triển – cho phép tự động hoá việc xây dựng phần mềm, sửa lỗi nhanh chóng và hỗ trợ lập trình AI-native. Điều này giảm thời gian build lên tới 30-40 % và giảm số lượng lỗi thủ công, mở ra khả năng triển khai dự án DevOps nhanh hơn. Đối với FPT, Codex tăng năng suất nhóm phát triển, rút ngắn chu kỳ giao hàng và nâng cao sức cạnh tranh trong chuyển đổi số.

Ngày xuất bản: 20 Jan, 2026 **Nguồn:** OpenAI News **URL:** https://openai.com/index/cisco

News #8: ServiceNow powers actionable enterprise AI with OpenAI

ServiceNow mở rộng truy cập các mô hình tiên tiến của OpenAI (ví dụ GPT-4) trên nền tảng của mình, cho phép tích hợp AI vào quy trình doanh nghiệp, tóm tắt tài liệu, tìm kiếm và giao diện thoại. Đối với FPT, đây là cơ hội nhanh chóng triển khai giải pháp tự động hoá dịch vụ IT và hỗ trợ khách hàng bằng AI mà không cần xây dựng hạ tầng mô hình riêng, giảm chi phí và thời gian triển khai.

Ngày xuất bản: 20 Jan, 2026 **Nguồn:** OpenAI News **URL:** https://openai.com/index/servicenow-powers-actionable-enterprise-ai-with-openai

News #9: Differential Transformer V2

DIFF V2 tăng số đầu-query lên gấp đôi mà không tăng KV, cho phép giải mã tốc độ bằng Transformer và dùng FlashAttention mà không kernel tùy chỉnh. Loại bỏ per-head RMSNorm giảm spike gradient và ổn định huấn luyện ở LR 6e-4∼1e-3; loss giảm 0.02-0.03 so với baseline sau 1 T token. Đối với FPT, giảm 25 % tham số attention giúp tái phân bổ tài nguyên cho mô-đun đặc thù, nâng hiệu suất dịch vụ đám mây AI.

Ngày xuất bản: 20 Jan, 2026 **Nguồn:** Huggingface Blog **URL:** https://huggingface.co/blog/microsoft/diff-attn-v2

News #10: Israel Activates National AI Supercomputer

Israel khởi động siêu máy tính quốc gia cho IA với hơn1000 bộ tăng tốc Nvidia B200 (∼16000 PFLOPS), đầu tư trên500 triệu ILS (∼158 triệu USD) gồm hỗ trợ chính phủ $50

triệu.

Tải nguồn phân bổ: doanh nghiệp nhận70%(ít nhất16 accelerator); viện nghiên cứu nhận30%(ít nhất8). Hiện > 90% lao động CNTT sử dụng IA và ≈ 22000 doanh nghiệp đã tích hợp IA.

Sau đó,FPT có thể học cách cung cấp tài nguyên nhanh chóng để xây dựng nền tảng IA nội địa và nâng cao năng lực cạnh tranh.]

Ngày xuất bản: 21 Jan, 2026 **Nguồn:** TechRepublic **URL:** https://www.techrepublic.com/article/news-israel-ai-supercomputer/

Nghiên cứu khoa học nổi bật (4 bài)

Article #1: Improving Training Efficiency and Reducing Maintenance Costs via Language Specific Model Merging

Phương pháp gộp mô hình theo ngôn ngữ mới cho phép huấn luyện lần đầu giảm tới 35 % thời gian (3.4→2.2h) và chi phí 5.6 %; khi cập nhật/ngôn ngữ mới thời gian giảm 73.7 % (3.8→1h) và chi phí tương tự. Case study doanh nghiệp: thời gian huấn luyện ↓50 %, chi phí cập nhật ↓62.4 %. Hiệu suất vẫn ngang với “retrain-all”, giúp FPT cắt giảm chi phí đám mây và tăng tốc triển khai đa ngôn.

Ngày xuất bản: 22 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.16127v1

Article #2: Introducing the Generative Application Firewall (GAF)

Giới thiệu Generative Application Firewall (GAF)—lớp bảo mật mới cho LLM tích hợp bộ lọc prompt, guardrails và data-masking thành điểm kiểm soát duy nhất như WAF. GAF bao gồm 5 lớp: Network, Access, Syntactic, Semantic và Context; dùng hệ thống đánh giá 5 sao với mỗi sao đại diện một lớp bảo vệ. Đối với FPT,GAF đồng bộ an ninh GenAI giảm nguy cơ jailbreak và hỗ trợ tuân thủ nhanh chóng.

Ngày xuất bản: 22 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.15824v1

Article #3: PyraTok: Language-Aligned Pyramidal Tokenizer for Video Understanding and Generation

PyraTok giới thiệu **tokenizer video đa cấp** được căn chỉnh ngôn ngữ (LaPQ) với **vocab ∼48 K** và chiến lược **đồng thời căn chỉnh cục bộ-toàn cục**; nhờ đó PSNR tăng lên **35.7 dB** (so với ∼30 dB của các VAE trước) và mAP trong nhận dạng hành động cải thiện **+5.75** điểm. Đối với FPT, công nghệ này giảm chi phí tính toán khi sinh video 4K/8K, đồng thời cho phép các dịch vụ đám mây thực hiện phân đoạn và hiểu video không cần dữ liệu gán nhãn – mở rộng khả năng AI video cho khách hàng doanh nghiệp.

Ngày xuất bản: 22 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.16210v1

Article #4: Agentic Confidence Calibration

Đã đề xuất vấn đề *Agentic Confidence Calibration* và khung **Holistic Trajectory Calibration (HTC)**, dùng 48 đặc trưng quá trình để hiệu chỉnh độ tin cậy của AI agents. HTC giảm ECE từ 0.12-0.45 xuống 0.03-0.07 và Brier Score tới 0.09-0.14 trên 8 benchmark, đồng thời tăng AUROC lên >0.70. Điều này giúp FPT nâng cao độ an toàn và độ tin cậy cho các dịch vụ AI đám mây trong môi trường yêu cầu cao như tài chính hay y tế.

Ngày xuất bản: 22 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.15778v1

Kết luận

Những xu hướng vừa nêu cho thấy ba trụ cột chiến lược mà FPT Smart Cloud cần tập trung: (1) khai thác cấu trúc block-diagonal của mô hình để tăng độ song song và giảm latency; (2) xây dựng nền tảng AI linh hoạt – từ no-code đến low-code – đồng thời tích hợp lớp bảo mật GAF để đáp ứng yêu cầu tuân thủ và an toàn dữ liệu; (3) tận dụng các kernel tối ưu (FlashAttention-4), tokenizer video tiên tiến và khung calibrate HTC nhằm nâng cao hiệu suất tính toán và độ tin cậy của agents trong các lĩnh vực nhạy cảm như tài chính và y tế. Đề xuất ngay: triển khai thử nghiệm block-diagonal inference trên môi trường cloud/edge hiện có; đánh giá tích hợp Airia và GAF vào quy trình phát triển sản phẩm nội bộ; đồng thời lên kế hoạch chuyển đổi một số workload sang FlashAttention-4 và PyraTok để đo lường lợi nhuận chi phí thực tế. Những bước đi này sẽ củng cố vị thế cạnh tranh của FCI trong kỷ nguyên AI đa dạng và đám mây-edge.

Tiêu đề	Ngày xuất bản	URL
Why Inference in Large Models Becomes Decomposable After Training	22 Jan, 2026	Link
Airia: Enterprise AI orchestration that unifies experimentation, prod, and governance (Sponsor)	24 Jan, 2026	Link
Qwen3-TTS Family is Now Open Sourced: Voice Design, Clone, and Generation!	24 Jan, 2026	Link
Improving Training Efficiency and Reducing Maintenance Costs via Language Specific Model Merging	22 Jan, 2026	Link
Introducing the Generative Application Firewall (GAF)	22 Jan, 2026	Link
Overcoming Compute and Memory Bottlenecks with FlashAttention-4 on NVIDIA Blackwell	24 Jan, 2026	Link
Small models, big results: Achieving superior intent extraction through decomposition	24 Jan, 2026	Link
Scaling PostgreSQL to power 800 million ChatGPT users	22 Jan, 2026	Link
Inside GPT-5 for Work: How Businesses Use GPT-5	22 Jan, 2026	Link
Cisco and OpenAI redefine enterprise engineering with AI agents	20 Jan, 2026	Link
ServiceNow powers actionable enterprise AI with OpenAI	20 Jan, 2026	Link
Differential Transformer V2	20 Jan, 2026	Link

Israel Activates National AI Supercomputer	21 Jan, 2026	Link
PyraTok: Language-Aligned Pyramidal Tokenizer for Video Understanding and Generation	22 Jan, 2026	Link
Agentic Confidence Calibration	22 Jan, 2026	Link

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