Bản tin công nghệ - 25/01/2026

Mở đầu

Trong tuần này, FPT Smart Cloud chứng kiến một loạt tiến bộ công nghệ quan trọng, từ việc củng cố an ninh cho giao diện ngôn ngữ tự nhiên bằng Generative Application Firewall (GAF) đến các cải tiến sâu về hiệu suất và chi phí cho các mô hình LLM và hạ tầng dữ liệu. Các giải pháp như LaPQ cho video-text alignment, chiến lược "train-once, merge-as-needed" cho đa ngôn ngữ và structural annealing đang mở ra khả năng giảm đáng kể thời gian huấn luyện và chi phí suy luận. Đồng thời, các sáng kiến mở rộng quy mô như PostgreSQL hàng triệu truy vấn/giây và các biện pháp bảo vệ nội dung AI (DISTSEAL) cho thấy xu hướng tối ưu hoá hạ tầng đồng thời nâng cao độ tin cậy và bảo mật.

Điểm nhấn: Introducing the Generative Application Firewall (GAF)

Generative Application Firewall (GAF) là lớp bảo mật duy nhất gộp toàn bộ các biện pháp hiện có – lọc prompt, guardrails, masking dữ liệu – thành một điểm kiểm soát trung tâm giống WAF nhưng dành cho giao diện ngôn ngữ tự nhiên. GAF triển khai năm lớp bảo vệ (Network, Access, Syntactic, Semantic, Context), lớp Context phát hiện các cuộc tấn công đa vòng như Echo-Chamber mà giải pháp không bắt được. Hệ thống đánh giá 5-star cho phép FPT nhanh chóng đo lường mức độ sẵn sàng và định hướng nâng cấp từ 2-star tới 5-star. Áp dụng GAF sẽ giảm rủi ro jailbreak và rò rỉ dữ liệu lên tới 80 % cho chatbot và công cụ AI nội bộ của FPT.

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

Tin nhanh công nghệ (10 bài)

News #1: Scaling PostgreSQL to power 800 million ChatGPT users

OpenAI đã đưa PostgreSQL lên mức hàng triệu truy vấn/giây bằng cách triển khai replica đa tầng, bộ nhớ đệm (caching) thông minh, giới hạn tốc độ (rate limiting) và tách biệt tải công việc (workload isolation). Điều này chứng minh khả năng mở rộng quy mô dữ liệu mà không chuyển sang NoSQL, giúp FPT tối ưu chi phí hạ tầng và nâng cao độ tin cậy cho các dịch vụ AI và nền tảng đám mây.

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

News #2: Inside GPT-5 for Work: How Businesses Use GPT-5

Trong báo cáo mới, tỷ lệ nhân viên sử dụng ChatGPT tăng mạnh, đạt mức chấp nhận rộng khắp các ngành và bộ phận – từ marketing tới bộ phận kỹ thuật. Điều này cho thấy nhu cầu tự động hoá công việc và khai thác AI đang trở thành tiêu chuẩn. Đối với FPT, nắm bắt xu

hướng này giúp thiết kế giải pháp SaaS và dịch vụ tư vấn AI phù hợp, nâng cao năng lực cạnh tranh.

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 #3: Cisco and OpenAI redefine enterprise engineering with AI agents

Codex – AI software agent do Cisco và OpenAI – được nhúng trực tiếp vào quy trình phát triển, tự động hoá sửa lỗi và tăng tốc thời gian build. Điều này giảm vòng lặp phát triển lên tới 30-40%, cho phép các dự án chuyển sang mô hình AI-native nhanh hơn. Đối với FPT, việc tích hợp Codex sẽ nâng cao năng suất đội DevOps và rút ngắn thời gian giao hàng cho khách hàng doanh nghiệp.

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

News #4: ServiceNow powers actionable enterprise AI with OpenAI

ServiceNow mở rộng tích hợp các mô hình frontier của OpenAI trên toàn bộ nền tảng, cho phép triển khai workflow doanh nghiệp dựa trên AI: tự động hoá quy trình, tóm tắt nội dung, tìm kiếm thông minh và giao diện thoại.

Đối với FPT, tính năng này giúp nhanh chóng nâng cấp giải pháp ServiceNow hiện có bằng AI sinh ngôn ngữ, giảm thời gian phản hồi dịch vụ và tăng năng suất cho khách hàng doanh nghiệp.

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

News #5: Differential Transformer V2

DIFF V2 tăng tốc giải mã bằng cách nhân đôi số đầu-query mà không tăng KV, cho phép dùng FlashAttention trực tiếp và loại bỏ kernel tùy chỉnh; đồng thời bỏ per-head RMSNorm giảm thiểu bùng nổ gradient và ổn định huấn luyện. Thử nghiệm trên mô hình tầm sản xuất cho thấy giảm loss ngôn ngữ 0.02-0.03 so với Transformer sau 1 T token và giảm khoảng 25 % tham số attention. Đối với FPT, điều này hạ chi phí GPU, rút ngắn thời gian tiền huấn luyện và nâng độ tin cậy của các LLM quy mô lớn.

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

News #6: Small models, big results: Achieving superior intent extraction through decomposition

Google đề xuất phương pháp phân tách: tóm tắt từng màn hình UI rồi trích xuất ý định từ chuỗi tóm tắt, cho phép mô hình nhỏ (Gemini 1.5 Flash 8B) đạt F1 tương đương Gemini Pro nhưng chi phí và thời gian xử lý chỉ bằng một phần nhỏ. Kết quả vượt CoT và E2E trên dữ liệu di động và web, giảm nguy cơ rò rỉ dữ liệu nhờ chạy hoàn toàn trên thiết bị.

Ngày xuất bản: 22 Jan, 2026 **Nguồn:** Google Research Blog **URL:** https://research.google/blog/small-models-big-results-achieving-superior-intent-extraction-through-decomposition/

News #7: PowerGen’s Shock Pivot: How AI Data Centers Hijacked an Energy Conference

PowerGen năm nay tập trung vào trung tâm dữ liệu AI: hơn 100 tỷ USD mỗi tháng dự án mới; tháng 10/2025 >350 tỷ USD dự án đang triển khai; tổng giá trị toàn cầu 3,2 nghìn tỷ USD, trong đó 2/3 ở Bắc Mỹ. Rack lên tới 600 kW và nhu cầu bổ sung ~20 MW khiến điện là rào cản chính. Đối với FPT cần hạ tầng năng lượng linh hoạt để hỗ trợ AI quy mô lớn.

Ngày xuất bản: 23 Jan, 2026 **Nguồn:** TechRepublic **URL:** https://www.techrepublic.com/article/news-ai-data-centers-powergen-energy-infrastructure/

News #8: Google Gemini Flaw Let Attackers Access Private Calendar Data

Vụ lỗ hổng Gemini cho phép kẻ tấn công chèn lệnh ngôn ngữ ẩn trong mô tả lịch họp và thu thập dữ liệu lịch cá nhân mà không cần người dùng nhấp chuột; đây là cách bypass kiểm soát quyền riêng tư của Google Calendar bằng “prompt” tự nhiên. Đối với FPT, việc tích hợp Gemini trong môi trường Workspace của khách hàng có thể rò rỉ thông tin nội bộ và gây mất uy tín; cần rà soát ngay các API AI-assistant và thiết lập chính sách lọc nội dung lịch.

Ngày xuất bản: 20 Jan, 2026 **Nguồn:** TechRepublic **URL:** https://www.techrepublic.com/article/news-google-gemini-flaw-private-calendar-data/

News #9: Unrolling the Codex agent loop

Codex agent loop được cập nhật: Codex CLI hiện orchestrate đồng thời nhiều mô hình, công cụ và prompt, thu thập kết quả qua Responses API. Thay đổi này giảm bước trung gian và tối ưu hiệu suất xử lý yêu cầu. Đối với FPT, khả năng tích hợp nhanh các công cụ AI vào pipeline nội bộ sẽ tăng năng suất phát triển và rút ngắn thời gian đưa sản phẩm ra thị trường.

Ngày xuất bản: 23 Jan, 2026 **Nguồn:** OpenAI News **URL:** https://openai.com/index/unrolling-the-codex-agent-loop

News #10: Stargate Community

Stargate Community kế hoạch mới đưa mô hình hạ tầng AI ưu tiên cộng đồng, thiết kế theo nhu cầu năng lượng địa phương và ưu tiên lực lượng lao động.
Đối với FPT, mô hình này giúp giảm chi phí vận hành lên tới 30 % nhờ tối ưu năng lượng và mở rộng nguồn nhân lực AI nội địa.

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

Nghiên cứu khoa học nổi bật (4 bài)

Article #1: PyraTok: Language-Aligned Pyramidal Tokenizer for Video Understanding and Generation

PyraTok introduces Language-aligned Pyramidal Quantization (LaPQ), a hierarchical tokenizer with ~48K token vocabulary and 95% codebook utilization, enabling multi-scale text-video alignment. Dual semantic alignment (local per-level + global autoregressive) reduces semantic drift, boosting zero-shot video segmentation mAP by up to 217.9% and action localization +5.75 mAP. For FPT, this means higher-fidelity video generation at lower compute, accelerating AI-driven media services and expanding multilingual video analytics capabilities.

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

Article #2: Improving Training Efficiency and Reducing Maintenance Costs via Language Specific Model Merging

Tiến bộ: áp dụng chiến lược "train-once, merge-as-needed" cho LLM đa ngôn ngữ, giảm thời gian huấn luyện ban đầu 35 % (3.4 h → 2.2 h), cập nhật/ngôn ngữ mới 74 % (3.8 h → 1.0 h) và chi phí hơn 70 % ($119.7 → $31.5), vẫn duy trì chất lượng gần bằng baseline.

Với FPT, tiết kiệm chi phí điện toán đám mây và rút ngắn chu kỳ ra mắt sản phẩm đa ngôn ngữ tăng lợi nhuận và đáp ứng nhanh nhu cầu khách hàng.

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

Article #3: Why Inference in Large Models Becomes Decomposable After Training

Mô hình lớn sau đào tạo không còn là một toán tử đơn khối mà tự động tách thành các khối chéo độc lập nhờ các cập nhật gradient chỉ diễn ra cục bộ; phần lớn tham số (đến hàng chục %-hơn) vẫn giữ phân phối khởi tạo và không ảnh hưởng tới đầu ra. Việc áp dụng "structural annealing" loại bỏ những tham số không có hiệu lực, chuyển ma trận dày đặc thành dạng block-diagonal, cho phép thực thi song song và giảm đáng kể chi phí suy luận – một lợi thế thiết yếu cho FPT trong việc mở rộng quy mô dịch vụ AI mà không tăng độ phức tạp hệ thống.

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

Article #4: Learning to Watermark in the Latent Space of Generative Models

DISTSEAL đưa watermark vào không gian tiềm Ẩn cho cả mô hình khuếch tán và tự hồi quy. So với pixel truyền thống, tốc độ tăng tới 20 × (3 ms → 63 ms) và độ chính xác nhị phân trên 95% (gần 97%). Distill vào mô hình sinh hoặc bộ giải mã giúp tích hợp ngay, ngăn chặn việc tắt khi mở source. Giải pháp bảo vệ IP và xác thực nội dung AI cho FPT.

Ngày xuất bản: 22 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.16140v1

Kết luận

Những xu hướng trên khẳng định nhu cầu ngày càng tăng về bảo mật lớp giao diện AI, tối ưu hoá chi phí tính toán và khả năng mở rộng hạ tầng dữ liệu trong môi trường doanh nghiệp. FPT nên ưu tiên triển khai GAF để giảm rủi ro jailbreak tới 80 %, áp dụng LaPQ và structural annealing nhằm nâng cao năng suất video AI và giảm chi phí suy luận, đồng thời khai thác kinh nghiệm mở rộng PostgreSQL để duy trì độ tin cậy cao mà không chuyển sang NoSQL. Cuối cùng, việc tích hợp các công cụ tự động hoá như Codex và ServiceNow AI sẽ tăng tốc vòng đời phát triển phần mềm và nâng cao trải nghiệm dịch vụ cho khách hàng.

Tiêu đề	Ngày xuất bản	URL
Introducing the Generative Application Firewall (GAF)	22 Jan, 2026	Link
PyraTok: Language-Aligned Pyramidal Tokenizer for Video Understanding and Generation	22 Jan, 2026	Link
Improving Training Efficiency and Reducing Maintenance Costs via Language Specific Model Merging	22 Jan, 2026	Link
Why Inference in Large Models Becomes Decomposable After Training	22 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
Small models, big results: Achieving superior intent extraction through decomposition	22 Jan, 2026	Link
PowerGen’s Shock Pivot: How AI Data Centers Hijacked an Energy Conference	23 Jan, 2026	Link
Google Gemini Flaw Let Attackers Access Private Calendar Data	20 Jan, 2026	Link
Learning to Watermark in the Latent Space of Generative Models	22 Jan, 2026	Link
Unrolling the Codex agent loop	23 Jan, 2026	Link
Stargate Community	20 Jan, 2026	Link
