Bản tin công nghệ - 23/01/2026

Mở đầu

Trong tuần 3 tháng 1 năm 2026, FPT Smart Cloud chứng kiến một loạt tiến bộ công nghệ nhằm củng cố an ninh LLM và tối ưu hoá hiệu suất triển khai AI ở mọi cấp độ – từ trung tâm dữ liệu tới thiết bị đầu cuối. Generative Application Firewall (GAF) giới thiệu mô hình bảo mật "5-star" có khả năng hiểu ngôn ngữ tự nhiên và giảm tới 70 % các cuộc tấn công prompt injection mà không làm tăng latency trên 15 ms. Đồng thời, các giải pháp như Codex và AgentOS rút ngắn chu kỳ phát triển và vận hành dự án AI lên đến 40 %, trong khi các cải tiến về mô hình đa phương tiện nhỏ (Gemini 1.5 Flash 8B) và đa ngôn ngữ gộp mô hình cho phép đạt độ chính xác tương đương mô hình lớn với chi phí tài nguyên giảm một nửa.

Điểm nhấn: Introducing the Generative Application Firewall (GAF)

Generative Application Firewall (GAF) – lớp bảo mật mới cho LLM – cho phép FPT tích hợp kiểm soát thống nhất từ mạng tới ngữ cảnh hội thoại. Khác với WAF truyền thống, GAF hiểu ngôn ngữ tự nhiên, theo dõi lịch sử đa vòng và có thể chặn hoặc redact nội dung, giảm nguy cơ jailbreak và rò rỉ dữ liệu nhạy cảm. Mô hình 5-star mỗi sao bổ sung một lớp; đạt 4 sao đã giảm 70 % các cuộc tấn công prompt injection trong thử nghiệm nội bộ. Đối với FPT Smart Cloud, GAF chuẩn hoá an ninh AI, hỗ trợ triển khai chatbot doanh nghiệp và dịch vụ AI-gateway mà không tăng latency trên 15 ms.

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: 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 sửa lỗi, tăng tốc thời gian build và hỗ trợ lập trình AI-native. Thay vì thực hiện thủ công các vòng lặp kiểm thử, Codex giảm thời gian build tới 30-40 % và tự động khắc phục lỗi phổ biến. Đối với FPT, việc rút ngắn chu kỳ phát triển giúp nâng cao năng suất dự án đám mây và tăng khả năng cạnh tranh trong các giải pháp AI doanh nghiệp.

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

News #2: Small models, big results: Achieving superior intent extraction through decomposition

Công nghệ mới: quy trình phân tách hai giai đoạn – tóm tắt từng màn hình UI rồi trích xuất ý định từ chuỗi tóm tắt – cho phép **mô hình đa phương tiện nhỏ** (ví dụ Gemini 1.5 Flash 8B) đạt **độ chính xác F1 tương đương** mô hình lớn Gemini Pro, giảm chi phí và tăng tốc độ xử lý trên thiết bị. Đối với FPT, khả năng triển khai AI hiểu ý định người

dùng **trực tiếp trên thiết bị** giúp nâng cao trải nghiệm di động, giảm độ trễ và bảo mật dữ liệu, mở đường cho các dịch vụ trợ lý thông minh nội bộ.

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 #3: GLM4-MoE Inference with SGLang

Novita AI đưa vào GLM4-MoE trên SGLang bốn tối ưu: Shared Experts Fusion (gộp chuyên gia chung), Qknorm Fusion, Async Transfer và Suffix Decoding cho mã hoá agentic. Các cải tiến giảm TTFT tới 65 % và TPOT 22 %; Shared Experts Fusion riêng còn cắt TTFT 23.7 % và ITL 20.8 %. Đối với FPT, tốc độ suy luận nhanh hơn giảm chi phí hạ tầng đám mây và cải thiện trải nghiệm khách hàng.

Ngày tổng hợp: 23 Jan, 2026 **Nguồn:** TLDR News **URL:** https://lmsys.org/blog/2026-01-21-novita-glm4/?utm_source=tldrai

News #4: The Runway for AI Engineering

AgentOS ra mắt runtime sẵn sàng sản xuất cho hệ thống AI đa-đại lý, loại bỏ nhu cầu tự xây dựng hạ tầng phân tán và quản lý vòng đời mô hình. Điều này giúp FPT rút ngắn thời gian triển khai dự án AI phức tạp, giảm chi phí vận hành lên tới 40 % và tăng khả năng mở rộng dịch vụ đám mây nội bộ.

Ngày tổng hợp: 23 Jan, 2026 **Nguồn:** TLDR News **URL:** https://x.com/ashpreetbedi/status/2014037319892762778?utm_source=tldrai

News #5: Unrolling the Codex agent loop

Codex agent loop mới trong Codex CLI tự động hoà trộn mô hình ngôn ngữ, công cụ và prompt qua Responses API. Thời gian phản hồi giảm 30 % và khả năng mở rộng tăng gấp đôi so với quy trình thủ công. Đối với FPT, tích hợp nhanh các mô hình AI vào dịch vụ cloud nâng cao năng suất phát triển và giảm chi phí vận hành.

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

News #6: Scaling PostgreSQL to power 800 million ChatGPT users

OpenAI mở rộng PostgreSQL lên hàng triệu truy vấn/giây bằng replica đa lớp, caching thông minh, rate limiting và workload isolation. Nhờ đó độ trễ trung bình < 5 ms và khả năng chịu tải tăng gấp 10 lần. Với FPT, cách tiếp cận này giúp triển khai dịch vụ AI quy mô lớn trên hạ tầng đám mây nội bộ, giảm chi phí và đáp ứng SLA nghiêm ngặt.

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

News #7: ServiceNow powers actionable enterprise AI with OpenAI

ServiceNow mở rộng truy cập mô hình tiên tiến của OpenAI (GPT-4o) trên nền tảng, cho phép AI trong quy trình doanh nghiệp, tóm tắt tài liệu, tìm kiếm và giọng nói.

Đối với FPT, điều này tạo cơ hội nhanh nhúng khả năng sinh ngôn ngữ mạnh vào giải pháp ServiceNow khách hàng, nâng cao tự động hoá ITSM và tăng sức cạnh tranh tại Việt Nam.

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

News #8: Differential Transformer V2

DIFF V2 bổ sung các đầu truy vấn (query heads) mà không tăng KV heads, cho phép giải mã nhanh bằng FlashAttention mà không cần kernel tùy chỉnh. Loại bỏ per-head RMSNorm giảm thiểu bùng nổ gradient; λ riêng token/đầu giúp kiểm soát RMS ngữ cảnh. Tiết kiệm $\approx 25\,\%$ tham số attention và đạt giảm loss 0.02-0.03 so với Transformer sau 1 T token, đồng thời giảm độ lệch kích hoạt.

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

News #9: PowerGen's Shock Pivot: How AI Data Centers Hijacked an Energy Conference

PowerGen 2024 chuyển trọng tâm sang AI data center: hơn $100 tỷ/tháng dự án mới được công bố trong năm qua; tháng 10/2025 ghi $350 tỷ dự án đang triển khai; tổng giá trị toàn cầu $3.2 nghìn tỷ, trong đó Bắc Mỹ chiếm 2/3. Rack lên tới 600 kW gây thiếu điện, kéo dài nhu cầu tự cung cấp bằng động cơ gas, pin hay thậm chí lò phản ứng hạt nhân. Đối với FPT, xu hướng này mở ra cơ hội cung cấp dịch vụ cloud năng lượng-hiệu quả và hợp tác xây dựng hạ tầng điện cho khách hàng doanh nghiệp.

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

News #10: Israel Activates National AI Supercomputer

Israel khởi động siÊU maY tÍnh Aİ qUỐC GIA vỚI HƠN ^{1000} bỏ tăNG tÓC NvIdIa B_{200} vaÀ khả năng $\approx^{15600}$ PETAFLOP - vƯỢT NHU CÂU HIỆN TẠI.
DưA an dỰ áNH thu HÚT đẦU TƯ $\approx_{500}$ TRIỆU NI S ($\sim_{158}$ TRIỆU USĐ) + HỖ TRỢ CHÍNH PHỦ $\approx_{150}$ TRIỆU NI S.
TÀI NGUYÊN ƯU TIÊN CHO DOANH NGHIỆP CÔNG NGHỆ VÀ PHẦN CÒN DÀNH CHO NGHIÊN CỨU HỌC THUẬT.
VỚI HƠN ^{22000} côNG TY ISRAEL ÁP DỤNG Aİ , MÔ HÌNH NAỲ GIÚP FPT XÂY DỰNG HẠ TẦNG NƠI ĐI A MẠNH MẼ VÀ GIẢM PHỤ THÕC VAO CLOUD TOÀN CẦU.)

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

Áp dụng gộp mô hình theo ngôn tiếng cho phép huấn luyện riêng từng ngôn rồi hợp nhất, rút thời gian huấn luyện ban đầu tới 50 % và giảm chi phí cập nhật/ngón mới hơn 70 %. Độ chính xác và BertScore của tóm tắt, suy luận giữ nguyên so với mô hình đa-ngôn; chỉ hơi giảm trong phân tích cảm xúc. Với FPT, tiết kiệm GPU và rút nhanh vòng đời sản phẩm đa-ngân giúp mở rộng dịch vụ hiệu quả

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

Article #2: PyraTok: Language-Aligned Pyramidal Tokenizer for Video Understanding and Generation

PyraTok đưa ra tokenizer video đa cấp với **Language-aligned Pyramidal Quantization (LaPQ)** và mã 48 K token, cùng chiến lược **dual semantic alignment** (căn chỉnh cục bộ + tự hồi quy toàn cục). Kết quả: PSNR ≈ 35.7 (trước ≈ 30), $+5.75$ mAP cho hành động thời gian và lên tới $+9.16$ % cho phân loại video . Đối với FPT, công nghệ này giảm chi phí tính toán khi tạo/hiểu video đa phương tiện và hỗ trợ mở rộng các dịch vụ AI video-ngôn ngữ .

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

Article #3: Learning to Watermark in the Latent Space of Generative Models

DISTSEAL cho phép chèn dấu vết trong không gian tiềm ẩn của cả mô hình khuếch tán và tự hồi quy, đồng thời giảm tới **20**$\times$ thời gian xử lý so với phương pháp pixel-space (**3 ms** $\rightarrow$ **63 ms**). Độ chính xác nhị phân đạt **95-96 %** sau các tấn công, gần bằng **97 %** của pixel. Việc distill vào trọng số mô hình hoặc decoder tạo watermark không thể tắt, bảo vệ tài sản trí tuệ và giảm chi phí tính toán cho các dịch vụ AI của FPT.

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

Article #4: Multimodal Climate Disinformation Detection: Integrating Vision-Language Models with External Knowledge Sources

Đã tích hợp mô hình đa phương tiện GPT-4o với các nguồn kiến thức bên ngoài (reverse image search, tìm kiếm web, trang fact-check và GPT-search). Khi kết hợp 4 nguồn, độ chính xác lên tới 69.6 % và F1 71.9 %, tỷ lệ từ chối giảm còn 0 %. Điều này cho phép FPT nâng cao khả năng phát hiện tin giả đa phương tiện trong dịch vụ AI/đám mây, bảo vệ uy tín và hỗ trợ khách hàng kiểm soát nội dung.

Ngày xuất bản: 22 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.16108v1

Kết luận

Những bước tiến vừa nêu khẳng định xu hướng ba trụ cột của ngành: an ninh LLM mạnh mẽ qua GAF; tốc độ phát triển nhanh chóng nhờ Codex và môi trường runtime AgentOS; và khả năng đưa trí tuệ nhân tạo vào biên với các mô hình nhẹ nhưng chính xác. Khi kết hợp

cùng hạ tầng mở rộng PostgreSQL và tích hợp GPT-4o trên ServiceNow, FPT có thể cung cấp dịch vụ AI quy mô lớn đáp ứng SLA chặt chẽ và tự động hoá quy trình doanh nghiệp. Đối với đội ngũ kỹ thuật, việc đánh giá pilot GAF cho AI-gateway, thử nghiệm AgentOS trong các dự án đa đại lý và khai thác các tối ưu hoá của Novita AI và PyraTok sẽ là những hành động thiết yếu để duy trì lợi thế cạnh tranh trong kỷ nguyên AI năng lượng-hiệu quả.

Tiêu đề	Ngày xuất bản	URL
Introducing the Generative Application Firewall (GAF)	22 Jan, 2026	Link
Cisco and OpenAI redefine enterprise engineering with AI agents	20 Jan, 2026	Link
Small models, big results: Achieving superior intent extraction through decomposition	22 Jan, 2026	Link
Improving Training Efficiency and Reducing Maintenance Costs via Language Specific Model Merging	22 Jan, 2026	Link
GLM4-MoE Inference with SGLang	23 Jan, 2026	Link
The Runway for AI Engineering	23 Jan, 2026	Link
Unrolling the Codex agent loop	23 Jan, 2026	Link
Scaling PostgreSQL to power 800 million ChatGPT users	22 Jan, 2026	Link
ServiceNow powers actionable enterprise AI with OpenAI	20 Jan, 2026	Link
Differential Transformer V2	20 Jan, 2026	Link
PowerGen’s Shock Pivot: How AI Data Centers Hijacked an Energy Conference	23 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
Learning to Watermark in the Latent Space of Generative Models	22 Jan, 2026	Link
Multimodal Climate Disinformation Detection: Integrating Vision-Language Models with External Knowledge Sources	22 Jan, 2026	Link

---