# Bản tin công nghệ - 22/01/2026

## Mở đầu

Tuần này FPT Smart Cloud ghi nhận một loạt tiến bộ công nghệ quan trọng từ các giải pháp AI và hạ tầng đám mây. Đặc biệt, kiến trúc MEC-oriented cho hệ thống phát hiện bất thường trên mạng 5G đã giảm độ trễ xuống dưới 1 ms và hỗ trợ tới 10 triệu flow/s; đồng thời mô hình GLM-4.7-Flash (30 tỷ tham số) cho phép đa nhiệm trong một mô hình duy nhất mà không cần kết nối đám mây. Các tích hợp mới như ServiceNow mở rộng truy cập tới các mô hình OpenAI và FastMCP 3.0 với các primitive Component/Provider/Transform cũng đang tạo tiền đề cho việc triển khai nhanh các quy trình doanh nghiệp dựa trên AI.

## Điểm nhấn: Dynamic Management of a Deep Learning-Based Anomaly Detection System for 5G Networks

Giải pháp quản lý động cho hệ thống phát hiện bất thường dựa trên deep learning tại mạng 5G mang lại ba điểm đột phá cho FPT Smart Cloud: (1) Kiến trúc MEC-oriented cho phép triển khai nhanh các mô-đun AI ngay tại biên mạng, giảm độ trễ xuống $< 1$ ms và hỗ trợ tới 10 triệu flow/s; (2) Chính sách tự động hoá (VI, ADF, MEApp) giúp mở rộng RAM, chuyển từ CPU sang GPU và cập nhật mô hình trong thời gian thực, giảm thời gian phản hồi từ 0,019 s lên 0,006 s khi lưu lượng vượt 4,3 triệu flow/s; (3) Đánh giá mô hình DNN đạt precision 95.37 % và recall 99.54 % trên dataset CTU, đồng thời duy trì precision 68.6 % cho botnet chưa biết. Những khả năng này cho phép FPT cung cấp dịch vụ an ninh mạng 5G linh hoạt, tiết kiệm tài nguyên và tăng giá trị dịch vụ đám mây vi mô cho khách hàng doanh nghiệp.

**Ngày xuất bản:** 21 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.15177v1

## Tin nhanh công nghệ (10 bài)

## News #1: GLM-4.7-Flash

GLM-4.7-Flash là mô hình MoE 30 tỷ tham số, hỗ trợ triển khai nội bộ không cần đám mây. Điểm mới so với phiên bản trước là đa nhiệm: lập trình, tác vụ đại lý AI, dịch thuật và sáng tạo văn bản trong một mô hình duy nhất. Đối với FPT, mô hình lớn chạy offline giảm chi phí băng thông, tăng bảo mật dữ liệu khách hàng và rút ngắn thời gian phản hồi AI.

**Ngày tổng hợp:** 22 Jan, 2026 **Nguồn:** TLDR News **URL:** https://x.com/Zai_org/status/2013261304060866758?utm_source=tldrai

## News #2: ServiceNow powers actionable enterprise AI with OpenAI

ServiceNow mở rộng truy cập tới các mô hình tiên tiến của OpenAI trên nền tảng của mình, cho phép triển khai quy trình doanh nghiệp dựa trên AI như tóm tắt nội dung, tìm kiếm thông minh và giao diện thoại.

Đối với FPT, tích hợp sẵn các mô hình này rút ngắn thời gian phát triển AI trên ServiceNow, tăng tính cạnh tranh trong dự án chuyển đổi số và giảm chi phí duy trì hạ tầng.

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

## News #3: Agentic AI on the rise: Keys to unlocking value (Sponsor)

Agentic AI đang trở thành dịch vụ chuẩn trong AWS Marketplace; ebook mới cung cấp 19 trang hướng dẫn thực tiễn về khung phát triển, tích hợp và vận hành các agent (native và open-source). Đồng thời webinar của SANS + AWS đề ra lộ trình bảo mật cho Generative AI trên Amazon Bedrock, nhấn mạnh quản trị dữ liệu và tính toàn vẹn mô hình. Đối với FPT, đây là tài nguyên nhanh để triển khai giải pháp AI tự động hoá quy trình và nâng cao an ninh khi mở rộng dịch vụ đám mây.

**Ngày tổng hợp:** 22 Jan, 2026 **Nguồn:** TLDR News **URL:** https://pages.awscloud.com/awsmp-gim-qmj2-adhoc-aim-agentic-applications-agentic-ai-ebook.html?trk=c539baab-ec1c-422b-8dc6-d22e2a28df26&sc_channel=el

## News #4: Inworld releases new TTS model designed for the next wave of consumer AI applications (Sponsor)

Inworld TTS-1.5 mang độ trễ thời gian thực và chất lượng tối ưu cho tương tác, hỗ trợ quy mô người dùng tiêu dùng. Tính năng clone giọng “zero-shot” tạo giọng tùy chỉnh chỉ với 15 giây âm thanh; phiên bản chuyên nghiệp yêu cầu ≥30 phút audio (khuyến nghị >20 phút) cho giọng trẻ em hoặc accent đặc biệt. Cho phép điều chỉnh temperature và tốc độ 0.5-1.5×, xuất timestamps và dữ liệu phoneme-level để đồng bộ môi môi trong các avatar AI của FPT.

**Ngày tổng hợp:** 22 Jan, 2026 **Nguồn:** TLDR News **URL:** https://inworld.ai/tts?utm_source=tldrai&utm_medium=paidemail&utm_campaign=tldr-ai-tts-1.5

## News #5: Introducing FastMCP 3.0 🚀

FastMCP 3.0 ra mắt ba primitive cơ bản – Component, Provider và Transform – cho phép nguồn dữ liệu linh hoạt, phiên bản công cụ song song và kiểm soát truy cập chi tiết. Phiên bản beta 3.0.0b1 hỗ trợ OpenTelemetry, background tasks và hot-reload. Với hơn 1 triệu lượt tải mỗi ngày và chiếm 70 % máy chủ MCP, nền tảng này giúp FPT triển khai Context Applications ổn định, bảo mật và quan sát tốt hơn.

**Ngày tổng hợp:** 22 Jan, 2026 **Nguồn:** TLDR News **URL:** https://www.jlowin.dev/blog/fastmcp-3?utm_source=tldrai

## News #6: Cisco and OpenAI redefine enterprise engineering with AI agents

Codex – AI software agent của Cisco và OpenAI – được nhúng trực tiếp vào quy trình phát triển, cho phép tăng tốc độ build lên tới 30 %, tự động sửa lỗi và hỗ trợ lập trình AI-native. Đối với FPT, việc tích hợp Codex giúp rút ngắn thời gian giao hàng dự án phần mềm, giảm

chi phí bảo trì và nâng cao năng lực cạnh tranh trong các giải pháp đám mây doanh nghiệp.

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

# News #7: Small models, big results: Achieving superior intent extraction through decomposition

Google giới thiệu quy trình phân đoạn: mỗi màn hình được mô hình nhỏ Gemini 1.5 Flash 8B tóm tắt riêng rồi tổng hợp thành câu khai báo ý định người dùng. Kết quả đạt F-score bằng Gemini 1.5 Pro nhưng chi phí và thời gian xử lý chỉ là một phần so với mô hình lớn. Điều này hỗ trợ FPT triển khai tính năng dự đoán hành vi trên thiết bị, giảm độ trễ và tăng bảo mật dữ liệu

**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 #8: Israel Activates National AI Supercomputer

Israel khai trương siêu máy tính AI quốc gia với 1000 bộ tăng tốc Nvidia B200 (~16000 PFLOPS), đầu tư >500 triệu NIS (~158 tr USD) và hỗ trợ nhà nước 160 triệu NIS. Tài nguyên chia: doanh nghiệp nhận-70%, tối thiểu 16 B200; viện nghiên cứu-30%, tối thiểu 8 B200. Với >90% nhân lực công nghệ dùng AI và 22k công ty đã tích hợp AI, mô hình này giúp FPT xây dựng hạ tầng nội địa giảm chi phí cloud.

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

# News #9: Differential Transformer V2

DIFF V2 tăng tốc giải mã bằng cách nhân đôi số query heads mà không tăng KV heads, cho phép dùng FlashAttention và loại bỏ kernel tùy chỉnh, đạt tốc độ tương đương Transformer. Loại bỏ per-head RMSNorm giảm mạnh các spike gradient và cải thiện ổn định khi huấn luyện với LR lớn (6e-4–1e-3). Tham số λ riêng token/đầu giảm Context RMS, xóa "attention sinks". Tiết kiệm ≈25 % tham số attention, đồng thời giảm loss LM 0.02-0.03 so với baseline tại 1 T token.

**Ngày tổng hợp:** 22 Jan, 2026 **Nguồn:** TLDR News **URL:** https://huggingface.co/blog/microsoft/diff-attn-v2?utm_source=tldrai

# News #10: Scaling PostgreSQL to power 800 million ChatGPT users

OpenAI tái cấu trúc PostgreSQL để xử lý hàng triệu truy vấn/giây bằng replica đọc đa cấp, cache nội bộ, rate-limiting và cô lập tải công việc; đạt >1 triệu QPS với độ trễ <5 ms. Đối với FPT, áp dụng kiến trúc này sẽ nâng cao khả năng phục vụ dịch vụ AI quy mô lớn, giảm chi phí hạ tầng và tăng tính ổn định của nền tảng đám mây.

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

# Nghiên cứu khoa học nổi bật (4 bài)

## Article #1: How to Build AI Agents by Augmenting LLMs with Codified Human Expert Domain Knowledge? A Software Engineering Framework

Triển khai khung phần mềm mã hoá kiến thức chuyên môn vào AI agent dựa trên LLM, kết hợp bộ phân loại yêu cầu, RAG và quy tắc codified. Kết quả: chất lượng đầu ra tăng 206 % (2.60 → 0.85) và độ ổn định mã cải thiện tới 267 %. Giúp FPT giảm phụ thuộc vào chuyên gia hiếm, tăng tốc triển khai giải pháp trực quan dữ liệu và mở rộng dịch vụ AI trên đám mây.

**Ngày xuất bản:** 21 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.15153v1

## Article #2: Emerging from Ground: Addressing Intent Deviation in Tool-Using Agents via Deriving Real Calls into Virtual Trajectories

RISE giải quyết độ lệch ý định trong đại lý dùng công cụ bằng cách chuyển lời gọi thực tế thành chuỗi hành động ảo và tạo mẫu tiêu cực qua đột biến tham số ICP. Phương pháp nâng cao hoàn thành nhiệm vụ +35.28% và phù hợp ý định +23.27%, vượt SOTA 1.2-42% (task) và 1.17-54% (intent). Đối với FPT, RISE giảm lỗi AI trên đám mây, tăng tin cậy và cải thiện trải nghiệm khách hàng.

**Ngày xuất bản:** 21 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.15120v1

## Article #3: Auditing Language Model Unlearning via Information Decomposition

Paper introduces PID-based audit cho việc “unlearning” LLM, tách rõ **knowledge đã xóa** (unique) và **knowledge còn lại** (redundant). Thử nghiệm cho thấy các phương pháp gradient giữ lại tới 0.41 bit “residual” (GA) trong khi RMU chỉ 0.08 bit; Forget Quality cao ($\approx 0.6$) nhưng không phản ánh rủi ro thực tế. Đánh giá này giúp FPT đáp ứng GDPR/AI Act và triển khai mô hình an toàn hơn.

**Ngày xuất bản:** 21 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.15111v1

## Article #4: The Why Behind the Action: Unveiling Internal Drivers via Agentic Attribution

Đã đề xuất khung phân tích nguyên nhân hành động (agentic attribution) cho các LLM-agent, thay vì chỉ xác định lỗi thất bại. Khung gồm cấp độ thành phần dựa trên động lực xác suất thời gian và cấp độ câu dựa trên phép loại bỏ/giữ lại, đạt Hit@1 = 0.944, Hit@5 = 1.0 trên 9 kịch bản. Đối với FPT, công cụ này giúp nhanh chóng phát hiện yếu tố nội tại gây ra quyết định không mong muốn của bot dịch vụ khách hàng, giảm rủi ro tài chính và nâng cao tính minh bạch khi triển khai AI quy mô lớn.

**Ngày xuất bản:** 21 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.15075v1

# Kết luận

Những cải tiến vừa nêu không chỉ nâng cao hiệu năng và độ tin cậy của các dịch vụ AI – từ giảm thời gian phản hồi của mô hình DNN (precision 95.37 %, recall 99.54 %) đến tối ưu hoá chi phí hạ tầng qua PostgreSQL xử lý $>1$ triệu QPS – mà còn củng cố khả năng tuân thủ (PID-based audit cho “unlearning”) và minh bạch (khung agentic attribution). Khi kết hợp các công cụ như RISE để giảm lỗi AI trên đám mây, Inworld TTS-1.5 cho trải nghiệm giọng nói thời gian thực và Codex tăng tốc quy trình phát triển phần mềm lên tới 30 %, FPT Smart Cloud sẵn sàng cung cấp dịch vụ an ninh mạng 5G linh hoạt, giải pháp AI nội bộ bảo mật dữ liệu khách hàng và nền tảng đám mây mạnh mẽ đáp ứng nhu cầu chuyển đổi số của doanh nghiệp.

| Tiêu đề | Ngày xuất bản | URL |
|---|---|---|
| Dynamic Management of a Deep Learning-Based Anomaly Detection System for 5G Networks | 21 Jan, 2026 | Link |
| GLM-4.7-Flash | 22 Jan, 2026 | Link |
| ServiceNow powers actionable enterprise AI with OpenAI | 20 Jan, 2026 | Link |
| How to Build AI Agents by Augmenting LLMs with Codified Human Expert Domain Knowledge? A Software Engineering Framework | 21 Jan, 2026 | Link |
| Emerging from Ground: Addressing Intent Deviation in Tool-Using Agents via Deriving Real Calls into Virtual Trajectories | 21 Jan, 2026 | Link |
| Auditing Language Model Unlearning via Information Decomposition | 21 Jan, 2026 | Link |
| The Why Behind the Action: Unveiling Internal Drivers via Agentic Attribution | 21 Jan, 2026 | Link |
| Agentic AI on the rise: Keys to unlocking value (Sponsor) | 22 Jan, 2026 | Link |
| Inworld releases new TTS model designed for the next wave of consumer AI applications (Sponsor) | 22 Jan, 2026 | Link |
| Introducing FastMCP 3.0 | 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 |
| Israel Activates National AI Supercomputer | 21 Jan, 2026 | Link |
| Differential Transformer V2 | 22 Jan, 2026 | Link |
| Scaling PostgreSQL to power 800 million ChatGPT users | 22 Jan, 2026 | Link |

---