Bản tin công nghệ - 16/01/2026

Mở đầu

Tuần này chúng ta chứng kiến một loạt tiến bộ mạnh mẽ trong lĩnh vực AI đa phương tiện và kiến trúc agentic. Từ mô hình Modality-Aware Mixture of Experts (MAMoE) của MoST đạt WER 2 % trên LibriSpeech-clean đến các cải tiến về ghi nhớ ngữ cảnh như STITCH và CAME-Bench, các giải pháp mới đang giảm nhiễu và nâng cao độ chính xác đáng kể. Đồng thời, các công nghệ tối ưu hạ tầng inference – từ GPU MI325X của AMD tới wafer-scale của Cerebras – hứa hẹn giảm chi phí và tăng tốc độ phục vụ AI trên đám mây. Những bước tiến này không chỉ mở ra cơ hội triển khai nhanh các dịch vụ nhận dạng giọng nói và tổng hợp âm thanh nội bộ mà còn củng cố nền tảng an ninh và khả năng mở rộng cho FPT trong môi trường đa nhà cung cấp.

Điểm nhấn: MoST: Mixing Speech and Text with Modality-Aware Mixture of Experts

MoST giới thiệu kiến trúc **Modality-Aware Mixture of Experts (MAMoE)** – các chuyên gia được phân nhóm theo modality (text-only, audio-only) và các chuyên gia chia sẻ để truyền tải thông tin chéo. Cơ chế routing dựa trên chỉ báo modality giúp mỗi token đi tới chuyên gia phù hợp, giảm nhiễu đại diện và tăng khả năng học đặc thù. Nhờ quy trình chuyển đổi "LLM → speech-text" chỉ dùng dữ liệu mở, MoST đạt **WER 2.0 %** trên LibriSpeech-clean và **WER 6.0 %** trong TTS – vượt hẳn các mô hình cạnh tranh có cùng quy mô tham số. Độ chính xác trung bình trên các benchmark audio language modeling lên tới **71.9 %**, cải thiện 2–3 điểm so với đối thủ. Với hiệu suất cao và chi phí dữ liệu thấp, MoST cho phép FPT nhanh chóng triển khai dịch vụ nhận dạng giọng nói, tổng hợp âm thanh và trợ lý đa ngôn ngữ trên nền tảng đám mây, nâng cao năng lực AI nội bộ và giảm phụ thuộc vào giải pháp bản quyền.

Ngày xuất bản: 15 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.10272v1

Tin nhanh công nghệ (10 bài)

News #1: GPT-5.2-Codex is now available in the Responses API

X đã bổ sung biểu tượng "Share Via" và "Copy Link to Tweet" cho phép sao chép nhanh URL từng tweet trong chuỗi thread; đồng thời dịch vụ ThreadReader cho phép "unroll" toàn bộ thread thành văn bản liền mạch. Nhờ tính năng này FPT có thể thu thập và lưu trữ ngay lập tức các cập nhật AI (ví dụ 6 tweet từ Cursor AI, WindSurf, WarpDotDev…) để theo dõi xu hướng công nghệ và hỗ trợ phân tích nội dung nhanh chóng.

Ngày tổng hợp: 16 Jan, 2026 **Nguồn:** TLDR News **URL:** https://threadreaderapp.com/thread/2011499597169115219.html?utm_source=tldrai

News #2: Ministral 3 Technical Report

Bạn vui lòng cung cấp thông tin chi tiết cần dựa vào để tôi có thể soạn mục báo cáo công nghệ theo yêu cầu.

Ngày tổng hợp: 16 Jan, 2026 **Nguồn:** TLDR News **URL:** https://arxiv.org/pdf/2601.08584?utm_source=tldrai

News #3: Doubling Inference Speed at Character.ai

DigitalOcean + AMD tối ưu GPU MI325X đạt 2× tăng throughput inference và giảm 91 % chi phí so với hạ tầng không tối ưu. Áp dụng FP8, AITER kernels, cấu hình DP2/TP4/EP4 và Kubernetes quản lý(DOKS) cho phép p90 latency ổn định; cache NFS giảm thời gian tải mô hình 10-15 %. Đối với FPT, những cải tiến này giúp triển khai dịch vụ AI quy mô lớn với chi phí thấp hơn và đáp ứng yêu cầu độ trễ cao hơn.

Ngày tổng hợp: 16 Jan, 2026 **Nguồn:** TLDR News **URL:** https://blog.character.ai/technical-deep-dive-how-digitalocean-and-amd-delivered-a-2x-production-inference-performance-increase-for-character-ai/?utm_source=tldrai

News #4: Tool Search now in Claude Code

Claude Code tích hợp MCP Tool Search, cho phép tải động công cụ khi mô tả >10 % ngữ cảnh; thay vì tiền tải, công cụ được tìm kiếm và chèn khi cần. Nhờ đó giảm lượng token tiêu tốn, tăng hiệu suất xử lý yêu cầu phức tạp và cắt giảm chi phí tính toán – yếu tố quan trọng cho các dự án AI của FPT trong môi trường đám mây tài nguyên hạn chế.

Ngày tổng hợp: 16 Jan, 2026 **Nguồn:** TLDR News **URL:** https://x.com/trq212/status/2011523109871108570?utm_source=tldrai

News #5: Building Agents with the Gemini Interactions API

Gemini Interactions API (beta) cho phép tạo AI Agent bằng vòng lặp Python, quản lý trạng thái hội thoại trên server và tự động ký hiệu suy nghĩ, không cần gửi lịch sử mỗi lần. Nhờ đó một agent có thể viết dưới 100 dòng mã để đọc/ghi file và liệt kê thư mục. Đối với FPT, giảm thời gian phát triển và chuẩn hoá giao diện AI giúp tăng năng suất tự động hoá quy trình nội bộ.

Ngày tổng hợp: 16 Jan, 2026 **Nguồn:** TLDR News **URL:** https://www.philschmid.de/building-agents-interactions-api?utm_source=tldrai

News #6: Microsoft on track to spend $500 million per year on Anthropic AI

Microsoft đang tăng chi tiêu cho dịch vụ AI của Anthropic, dự kiến đạt 500 triệu USD/năm và đã đầu tư 5 tỷ USD trong thỏa thuận cùng Nvidia. Công ty cũng tích hợp mô hình Anthropic vào GitHub Copilot và 365 Copilot. Đối với FPT, xu hướng đa nhà cung cấp AI này cho thấy nhu cầu xây dựng nền tảng mở, tối ưu chi phí và chuẩn bị hợp tác hoặc cạnh tranh trong môi trường đám mây lai.

Ngày tổng hợp: 16 Jan, 2026 **Nguồn:** TLDR News **URL:** https://sherwood.news/tech/report-microsoft-on-track-to-spend-usd500-million-per-year-on-anthropic-ai/?

utm_source = tldrai

News #7: Open Responses: What you need to know

Open Responses giới thiệu chuẩn inference mở thay thế Chat Completion, cho phép mô hình stateless, luồng suy luận dạng sự kiện ngữ nghĩa và hỗ trợ vòng lặp công cụ (sub-agent) với tham số max_tool_calls = 5. Đối với FPT, tiêu chuẩn này đồng nhất giao diện nhà cung cấp và router, giảm chi phí tích hợp AI agentic và tăng tốc triển khai dịch vụ đám mây nội bộ.

Ngày xuất bản: 15 Jan, 2026 **Nguồn:** Huggingface Blog **URL:** https://huggingface.co/blog/open-responses

News #8: 🐱 Voice cloning just became free (and local)

Kyutai ra mắt Pocket TTS – mô hình sao chép giọng nói 100 triệu tham số chạy trên CPU laptop (Apple M3/Intel Core Ultra), tốc độ thực-time, WER 1.84 %—tốt hơn các mô hình lớn gấp 7 lần. Mã nguồn mở MIT và 88 k giờ dữ liệu cho phép FPT triển khai dịch vụ thoại nội bộ mà không cần GPU hay gửi dữ liệu lên đám mây, giảm chi phí hạ tầng và tăng bảo mật cho khách hàng.

Ngày xuất bản: 15 Jan, 2026 **Nguồn:** NeuronDaily **URL:** https://www.theneurondaily.com/p/voice-cloning-just-became-free-and-local-0f63

News #9: OpenAI signs $10 billion Deal for Compute

OpenAI ký hợp đồng đa năm với Cerebras, triển khai 750 MW hệ thống wafer-scale để cung cấp dịch vụ suy luận AI tốc độ cao, đáp ứng nhanh tới 15 lần so với GPU. Đợt ra mắt bắt đầu 2026 sẽ tạo ra triển khai inference lớn nhất thế giới. Đối với FPT, việc sở hữu nền tảng siêu nhanh này giúp nâng cấp dịch vụ AI real-time trên cloud/edge, tăng tính cạnh tranh và mở rộng thị trường.

Ngày tổng hợp: 16 Jan, 2026 **Nguồn:** TLDR News **URL:** https://www.cerebras.ai/blog/openai-partners-with-cerebras-to-bring-high-speed-inference-to-the-mainstream?utm_source = tldrai

News #10: CISA Issues New AI Security Guidance for Critical Infrastructure

CISA ra hướng dẫn an ninh AI cho OT – khung chuẩn toàn cầu đầu tiên cùng hơn 10 cơ quan an ninh mạng quốc tế. Hướng dẫn cảnh báo rủi ro khi dùng ChatGPT và các công cụ sinh tạo trong nhà máy, lưới điện và hệ thống nước chưa được bảo vệ đầy đủ. Đối với FPT, tiêu chuẩn này hỗ trợ xây dựng giải pháp AI-OT an toàn cho khách hàng công nghiệp.

Ngày xuất bản: 15 Jan, 2026 **Nguồn:** TechRepublic **URL:** https://www.techrepublic.com/article/news-cisa-ai-security-guidance-2026/

Nghiên cứu khoa học nổi bật (4 bài)

Article #1: Grounding Agent Memory in Contextual Intent

STITCH – hệ thống nhớ agentic dựa trên “contextual intent” – mới gắn nhãn từng bước bằng ba tín hiệu: chủ đề, hành động, thực thể . Nhờ đó truy xuất giảm nhiễu 35.6 % so với baseline ṣi giữ độ chính xác ＞80 % khi chiều dài tăng gấp

CAME-Bench cung cấp bộ kiểm thử cho khả năng nhớ dài hạn . Đối Với FPT , giải pháp này nâng cao tin cậy của AI trợ lý ṣi dịch vụ đám mây

Ngày xuất bản: 15 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.10702v1

Article #2: Structure and Diversity Aware Context Bubble Construction for Enterprise Retrieval Augmented Systems

Ra mắt Context Bubble – cách xây dựng ngữ cảnh RAG dựa trên cấu trúc tài liệu và đa dạng. So với Top-K truyền thống, token giảm từ 780 xuống 214 (-73 %), phần độc đáo tăng từ 1 lên 3 và độ trùng lặp còn 0.19; độ ổn định cải thiện (+/-6). Giúp FPT tiết kiệm token, nâng độ chính xác trả lời và cung cấp truy xuất có thể kiểm toán cho doanh nghiệp.

Ngày xuất bản: 15 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.10681v1

Article #3: Molmo2: Open Weights and Data for Vision-Language Models with Video Understanding and Grounding

Molmo2 giới thiệu ba mô hình (4 billion / 8 billion / 7 billion tham số) trọng số mở cùng hơn 7 triệu mẫu từ chín bộ dữ liệu mới: 104 k chú thích video chi tiết (∼924 từ mỗi clip), 650 k truy vấn “pointing” trên video và 3,6 k đoạn tracking đa đối tượng. Molmo2-8 b đạt **F1 = 38·4** trong video pointing—gấp đôi Gemini 3 Pro (**20**)—và độ chính xác đếm **35·5** so với Qwen3-VL-8 b (**29·6**). Đối với FPT, khả năng grounding không gian–thời gian này hỗ trợ robot tự động hoá, giám sát an ninh và phân tích nội dung video quy mô lớn.

Ngày xuất bản: 15 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.10611v1

Article #4: Defending Large Language Models Against Jailbreak Attacks via In-Decoding Safety-Awareness Probing

Phương pháp SafeProbing mới khai thác tín hiệu an toàn tiềm ẩn trong quá trình giải mã của LLM, thực hiện kiểm tra tại các checkpoint (tỷ lệ 20%). Kết quả cho thấy DSR tăng từ ∼81 % lên tới 96 % trên Qwen và từ 49 % lên 49 % – giảm tỉ lệ over-refusal từ 165 xuống còn 18 mẫu. Cải tiến này giúp FPT nâng cao khả năng phòng chống jailbreak mà không làm giảm chất lượng phản hồi.

Ngày xuất bản: 15 Jan, 2026 **Nguồn:** arXiv cs.AI **URL:** https://arxiv.org/pdf/2601.10543v1

Kết luận

Nhìn chung, xu hướng hiện nay tập trung vào ba trụ cột: (1) tối ưu hoá mô hình đa phương tiện để giảm chi phí dữ liệu và nâng cao hiệu suất; (2) xây dựng bộ nhớ ngữ cảnh agentic giúp truy xuất thông tin chính xác hơn trong các tác vụ dài hạn; (3) chuẩn hoá giao diện

inference và áp dụng phần cứng siêu tốc nhằm đáp ứng yêu cầu latency thấp và bảo mật dữ liệu. Đối với FPT, việc tích hợp nhanh những công nghệ này – từ MAMoE tới Pocket TTS và chuẩn Open Responses – sẽ tăng năng lực AI nội bộ, giảm phụ thuộc vào giải pháp bản quyền và tạo nền tảng vững chắc cho các dự án AI-OT an toàn.

Tiêu đề	Ngày xuất bản	URL
MoST: Mixing Speech and Text with Modality-Aware Mixture of Experts	15 Jan, 2026	Link
GPT-5.2-Codex is now available in the Responses API	16 Jan, 2026	Link
Ministral 3 Technical Report	16 Jan, 2026	Link
Grounding Agent Memory in Contextual Intent	15 Jan, 2026	Link
Structure and Diversity Aware Context Bubble Construction for Enterprise Retrieval Augmented Systems	15 Jan, 2026	Link
Molmo2: Open Weights and Data for Vision-Language Models with Video Understanding and Grounding	15 Jan, 2026	Link
Defending Large Language Models Against Jailbreak Attacks via In-Decoding Safety-Awareness Probing	15 Jan, 2026	Link
Doubling Inference Speed at Character.ai	16 Jan, 2026	Link
Tool Search now in Claude Code	16 Jan, 2026	Link
Building Agents with the Gemini Interactions API	16 Jan, 2026	Link
Microsoft on track to spend $500 million per year on Anthropic AI	16 Jan, 2026	Link
Open Responses: What you need to know	15 Jan, 2026	Link
🐱 Voice cloning just became free (and local)	15 Jan, 2026	Link
OpenAI signs $10 billion Deal for Compute	16 Jan, 2026	Link
CISA Issues New AI Security Guidance for Critical Infrastructure	15 Jan, 2026	Link

---