Bản tin công nghệ - 20/01/2026

Mở đầu

Trong thời kỳ mà các mô hình ngôn ngữ lớn (LLM) đang mở rộng tới hàng nghìn tỷ tham số và yêu cầu ngữ cảnh lên tới hàng trăm nghìn token, việc tối ưu hoá kiến trúc tính toán và quản trị AI trở nên cấp bách hơn bao giờ hết. Những tiến bộ gần đây – từ **Chunked Pipeline Parallelism** của SGLang giảm thiểu bubble trong pipeline cho đến các trung tâm dữ liệu AI đa gigawatt của SoftBank – đang vẽ ra một bức tranh mới về khả năng phục vụ các ứng dụng "infinite context" và dịch vụ AI doanh nghiệp với chi phí hạ tầng được rút gọn đáng kể.

Điểm nhấn: Scaling Trillion-Parameter Models with PP

SGLang triển khai Pipeline Parallelism (PP) tối ưu cho siêu-dài context lên tới triệu token bằng ba kỹ thuật: Chunked Pipeline Parallelism, giao tiếp bất đồng bộ P2P và Dynamic Chunking, giúp giảm truyền dữ liệu và loại bỏ pipeline bubbles. Benchmark trên cụm H20 cho thấy PP4 × TP8 tăng 3.31 lần tốc độ Prefill so với TP8 và vượt TP32 tới 30.5%, giảm TTFT tới 67.9% đồng thời duy trì hiệu suất mở rộng mạnh 82.8%. Đối với FPT, khả năng phục vụ mô hình hàng nghìn tỷ tham số với ngữ cảnh hàng trăm nghìn token sẽ giảm chi phí hạ tầng, rút ngắn thời gian phản hồi cho các dịch vụ AI doanh nghiệp và mở đường cho các ứng dụng “infinite context” như trợ lý ảo hay phân tích tài liệu quy mô lớn

Ngày tổng hợp: 20 Jan, 2026 **Nguồn:** TLDR News **URL:** https://lmsys.org/blog/2026-01-15-chunked-pipeline/?utm_source=tldrai

Tin nhanh công nghệ

News #1: Netomi’s lessons for scaling agentic systems into the enterprise

Netomi triển khai AI doanh nghiệp dựa trên GPT-4.1 và GPT-5.2, kết hợp xử lý đồng thời, khung quản trị và suy luận đa bước để đạt quy trình sản xuất ổn định ở quy mô lớn. Áp dụng tương tự giúp FPT mở rộng dịch vụ AI cho khách hàng doanh nghiệp, giảm độ trễ và tăng độ tin cậy trong môi trường thực tế.

Ngày xuất bản: 08 Jan, 2026 **Nguồn:** OpenAI News **URL:** https://openai.com/index/netomi

News #2: OpenAI and SoftBank Group partner with SB Energy

OpenAI hợp tác SoftBank Group và SB Energy triển khai các khu trung tâm dữ liệu AI quy mô đa gigawatt, trong đó có dự án Texas 1,2 GW hỗ trợ sáng kiến Stargate. Việc sở hữu năng lực điện năng lớn cho AI giúp giảm chi phí vận hành và tăng tốc độ triển khai mô hình LLM, tạo cơ hội cho FPT nâng cấp hạ tầng đám mây nội địa và cung cấp dịch vụ AI quy mô doanh nghiệp.

Ngày xuất bản: 09 Jan, 2026 **Nguồn:** OpenAI News **URL:** https://openai.com/index/stargate-sb-energy-partnership

News #3: How Tolan builds voice-first AI with GPT-5.1

Tolan giới thiệu trợ lý AI dạng giọng nói dựa trên GPT-5.1, tích hợp phản hồi siêu nhanh, tái tạo ngữ cảnh theo thời gian thực và cá tính dựa trên bộ nhớ, cho phép hội thoại tự nhiên. Điều này giúp FPT nâng cao khả năng cung cấp dịch vụ AI thoại trên nền tảng đám mây, rút ngắn độ trễ giao tiếp và tạo trải nghiệm cá nhân hoá cho khách hàng doanh nghiệp.

Ngày xuất bản: 07 Jan, 2026 **Nguồn:** OpenAI News **URL:** https://openai.com/index/tolan

News #4: Datadog uses Codex for system-level code review

RỖNG

Ngày xuất bản: 09 Jan, 2026 **Nguồn:** OpenAI News **URL:** https://openai.com/index/datadog

News #5: Elon Musk's xAI brings 1GW Colossus 2 AI training cluster online

XAI đã đưa siêu máy tính Colossus 2 (công suất 1 GW) vào hoạt động; sẽ nâng lên 1,5 GW vào tháng 4 và đạt ~2 GW. Vòng Series E $20 tỷ cho phép triển khai hơn 1 triệu GPU H100 tương đương. Đối với FCI, sức mạnh tính toán quy mô thành phố này giúp rút ngắn thời gian huấn luyện mô hình AI, tăng năng lực dịch vụ đám mây và tạo lợi thế cạnh tranh.

Ngày tổng hợp: 20 Jan, 2026 **Nguồn:** TLDR News **URL:** https://www.teslarati.com/elon-musk-xai-brings-1gw-colossus-2-ai-training-cluster-online/?utm_source=tldrai

News #6: Uniqueness-Aware RL for LLM Diversity

Mô hình “Uniqueness-Aware Reinforcement Learning” (RL) mới cho phép các LLM (Large Language Model) ưu tiên các giải pháp hiếm và sáng tạo hơn trong quá trình giải quyết vấn đề. Đây là bước tiến quan trọng giúp FPT nâng cao khả năng tạo nội dung độc đáo, tăng sức cạnh tranh của các sản phẩm AI và mở rộng tiềm năng ứng dụng trong tự động hoá sáng tạo nội dung và thiết kế giải pháp thông minh.

Ngày tổng hợp: 20 Jan, 2026 **Nguồn:** TLDR News **URL:** https://arxiv.org/abs/2601.08763?utm_source=tldrai

News #7: 😼 AI needs independent auditors now

AVERI – phi lợi nhuận mới nhận 7,5 triệu USD – đề xuất chuẩn 'AI Assurance Levels' và thúc đẩy kiểm toán độc lập cho mô hình AI tiên tiến. Đối với FPT, áp dụng audit bên thứ ba sẽ giảm rủi ro pháp lý và nâng độ tin cậy khi cung cấp dịch vụ đám mây AI. Micron cảnh báo thiếu hụt chip nhớ do nhu cầu AI “không có tiền lệ”, kéo dài tới sau 2026, ảnh hưởng tới khả năng mở rộng hạ tầng tính toán.

Ngày xuất bản: 19 Jan, 2026 **Nguồn:** NeuronDaily **URL:** https://www.theneurondaily.com/p/ai-needs-independent-auditors-now

News #8: Benchmarking AI Agent Memory: Is a Filesystem All You Need?

Nền tảng mới ra mắt bộ mã Remember-First – Lettta Code – để agent lưu trữ&học lâu dài; kèm Letta Filesystem hỗ trợ gắn file với grep&search_files. Agent đạt 74 % độ chính xác trên LoCoMo chỉ bằng việc giữ lịch sử hội thoại trong file, vượt 68,5 % của Mem0. Kết quả

chứng minh quản lý ngữ cảnh tốt hơn công cụ nhớ chuyên biệt, giúp FPT phát triển AI ghi nhớ hiệu quả&giảm chi phí triển khai.

Ngày tổng hợp: 20 Jan, 2026 **Nguồn:** TLDR News **URL:** https://www.letta.com/blog/benchmarking-ai-agent-memory?utm_source=tldrai

News #9: The Code-Only Agent

Thay đổi: mô hình 'Code-Only Agent' chỉ cho phép một công cụ duy nhất – execute_code – thay thế mọi tool (bash, ls, grep). Tất cả hành động đều được biểu diễn bằng mã chạy được, tạo ra "code witness" có semantics xác định. Quan trọng với FPT vì giảm độ phức tạp tích hợp công cụ, tăng tính kiểm chứng và tái sử dụng script trong các dự án AI/đám mây, đồng thời hỗ trợ xây dựng pipeline tự động đáng tin cậy.

Ngày tổng hợp: 20 Jan, 2026 **Nguồn:** TLDR News **URL:** https://rijnard.com/blog/the-code-only-agent?utm_source=tldrai

Kết luận

Những xu hướng vừa nêu cho thấy ba trụ cột chiến lược cho FPT trong năm tới: (1) khai thác sâu hơn các kỹ thuật song song như Pipeline Parallelism và Dynamic Chunking để tăng tốc độ Prefill và giảm thời gian phản hồi; (2) củng cố nền tảng quản trị AI qua chuẩn "AI Assurance Levels" và mô hình Code-Only Agent nhằm nâng cao độ tin cậy và khả năng kiểm chứng; (3) đẩy mạnh đầu tư vào năng lực tính toán quy mô đô thị – siêu máy tính Colossus 2 và các trung tâm dữ liệu đa gigawatt – để đáp ứng nhu cầu đào tạo và triển khai LLM siêu dài. Khi những yếu tố này được tích hợp đồng bộ vào hạ tầng đám mây nội địa của FCI, chúng sẽ không chỉ rút ngắn thời gian phản hồi cho trợ lý ảo hay phân tích tài liệu quy mô lớn mà còn mở ra cơ hội thực thi các giải pháp sáng tạo nội dung và dịch vụ AI thoại cá nhân hoá cho khách hàng doanh nghiệp.

Tiêu đề	Ngày xuất bản	URL
Scaling Trillion-Parameter Models with PP	20 Jan, 2026	Link
Netomi's lessons for scaling agentic systems into the enterprise	08 Jan, 2026	Link
OpenAI and SoftBank Group partner with SB Energy	09 Jan, 2026	Link
How Tolan builds voice-first AI with GPT-5.1	07 Jan, 2026	Link
Datadog uses Codex for system-level code review	09 Jan, 2026	Link
Elon Musk's xAI brings 1GW Colossus 2 AI training cluster online	20 Jan, 2026	Link
Uniqueness-Aware RL for LLM Diversity	20 Jan, 2026	Link
😼 AI needs independent auditors now	19 Jan, 2026	Link
Benchmarking AI Agent Memory: Is a Filesystem All You Need?	20 Jan, 2026	Link
The Code-Only Agent	20 Jan, 2026	Link

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