Bản tin công nghệ - 20/01/2026

Mở đầu

Trong bối cảnh nhu cầu mở rộng quy mô mô hình ngôn ngữ lớn và tăng cường khả năng sáng tạo AI đang bùng nổ, tuần này chúng ta chứng kiến một loạt tiến bộ công nghệ quan trọng: kiến trúc Pipeline Parallelism (PP) của SGLang cho phép xử lý siêu-dài context lên tới 1 triệu token với tăng tốc Prefill Throughput gấp 3.31 lần; siêu máy tính Colossus-2 của X-AI đạt gần 2 GW công suất và hơn một triệu GPU H100; phương pháp Reinforcement Learning "Rewarding the Rare" hướng tới các đáp án hiếm và sáng tạo; mô hình Code-Only Agent cung cấp "code witness" cho mọi hành động; Letta API và Letta Code khẳng định sức mạnh của quản lý ngữ cảnh nhớ-first; và chế độ Extended Thinking mở rộng giới hạn token lên tới 63 999 cho các mô hình 64K output. Tất cả đều mở ra cơ hội giảm chi phí hạ tầng, rút ngắn thời gian phản hồi và nâng cao tính kiểm chứng cho các giải pháp AI doanh nghiệp.

Điểm nhấn: Scaling Trillion-Parameter Models with PP

SGLang đã ra mắt kiến trúc Pipeline Parallelism (PP) tối ưu cho siêu-dài context, kết hợp Chunked PP, giao tiếp P2P bất đồng bộ và cơ chế Dynamic Chunking. Trong môi trường đa-node H20, PP4 × TP8 mang lại **3.31 lần tăng tốc Prefill Throughput** so với TP8 và vượt **TP32 tới 30.5 %**, đồng thời giảm TTFT lên tới **67.9 %** và duy trì **82.8 % strong scaling efficiency**. Đặc biệt, giải pháp hỗ trợ context lên tới **1 triệu token**, tích hợp PD Disaggregation và HiCache giúp chuyển KVCache nhanh chóng giữa node tiền-prefill và decode. Đối với FPT, công nghệ này cho phép triển khai mô hình hàng nghìn tỷ tham số với chi phí phần cứng thấp hơn, rút ngắn thời gian phản hồi AI doanh nghiệp và mở rộng dịch vụ đám mây thông minh một cách bền vững.

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 (6 bài)

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

X-AI đã đưa siêu máy tính Colossus-2 vận hành ở công suất ban đầu **1 GW**—cao hơn nhu cầu điện tối đa của San Francisco—sẽ nâng lên **1.5 GW** vào tháng 4 để đạt gần **2 GW** và hơn một triệu GPU **H100** tương đương. Vòng Series-E thu hút **US$20 billion**, định giá xAI ở **US$250 billion**. Đối với FPT việc nắm bắt kiến trúc siêu máy tính quy mô thành phố giúp thiết kế hạ tầng đám mây/edge mạnh mẽ hơn và giảm chi phí cho các dự án AI lớn.

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 #2: TLDR is hiring a Senior Software Engineer, Applied AI ($200k-$300k, Fully Remote)

Xin lỗi, tôi cần nội dung thông tin cụ thể để có thể soạn mục báo cáo công nghệ theo yêu cầu. Vui lòng cung cấp chi tiết cần trình bày.

Ngày tổng hợp: 20 Jan, 2026 **Nguồn:** TLDR News **URL:** https://jobs.ashbyhq.com/tldr.tech/3b21aaf8-dea5-4127-be71-602d30e5001e?utm_source=tldrai

News #3: Uniqueness-Aware RL for LLM Diversity

Bài "Rewarding the Rare: Uniqueness-Aware RL for Creative Problem Solving in LLMs" giới thiệu phương pháp Reinforcement Learning mới, ưu tiên các đáp án hiếm và sáng tạo thay vì chỉ tối ưu độ chính xác truyền thống. Điều này cho phép các mô hình ngôn ngữ lớn tạo ra giải pháp độc đáo, hỗ trợ FPT trong việc phát triển sản phẩm AI sáng tạo và tăng sức cạnh tranh trên thị trường dịch vụ trí tuệ nhân tạo.

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

News #4: The Code-Only Agent

Thay đổi: giới thiệu mô hình Code-Only Agent chỉ dùng một công cụ duy nhất — execute_code—toàn bộ tác vụ bằng mã chạy trực tiếp, thay vì chuỗi lệnh bash, ls, grep. Kết quả là mỗi hành động đều có "code witness" với ngữ nghĩa xác định (trả về <1 KB hoặc ghi JSON nếu >1 KB). Tầm quan trọng: giúp FPT giảm lỗi hallucination, tăng tính tái sử dụng và kiểm chứng trong các dịch vụ AI/đám mây.

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

News #5: Benchmarking AI Agent Memory: Is a Filesystem All You Need?

Letta API cho phép xây dựng agent nhớ và học người dùng lâu dài; Letta Code là khung mã nhớ-first, hỗ trợ đa mô hình. Đánh giá LoCoMo, Letta Filesystem đạt 74 % độ chính xác chỉ bằng công cụ file (grep, search_files), vượt Mem0 68.5 %. Điều này chứng minh rằng quản lý ngữ cảnh hiệu quả quan trọng hơn công cụ nhớ phức tạp, giúp FPT triển khai AI nhanh, chi phí thấp và mở rộng dễ dàng.

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

News #6: UltraThink is Dead. Long Live Extended Thinking

Ultrathink đã bị loại bỏ; chế độ Extended Thinking bật mặc định cho các model hỗ trợ với ngân sách 31 999 token. Đối với mô hình 64K output (Opus 4.5, Sonnet 4/4.5, Haiku 4.5) có thể tăng lên 63 999 token bằng biến môi trường MAX_THINKING_TOKENS = 63999; cũng có thể tắt bằng MAX_THINKING_TOKENS = 0 hoặc alwaysThinkingEnabled:false. Điều này cho phép FPT khai thác suy luận sâu hơn cho thiết kế hệ thống phức tạp và tối ưu chi phí token

Ngày tổng hợp: 20 Jan, 2026 **Nguồn:** TLDR News **URL:** https://decodeclaude.com/ultrathink-deprecated/?utm_source=tldrai

Nghiên cứu khoa học nổi bật (0 bài)

Kết luận

Những đột phá vừa nêu không chỉ nâng cao hiệu suất tính toán và khả năng sáng tạo của LLM mà còn cung cấp nền tảng hạ tầng mạnh mẽ, linh hoạt cho FPT trong việc triển khai mô hình hàng nghìn tỷ tham số, phát triển dịch vụ AI sáng tạo và mở rộng đám mây/edge một cách bền vững. Việc áp dụng Pipeline Parallelism, tận dụng sức mạnh siêu máy tính quy mô thành phố, tích hợp RL ưu tiên độ hiếm, sử dụng Code-Only Agent và Letta API sẽ giúp chúng ta giảm chi phí phần cứng, tối ưu token và tăng độ tin cậy của hệ thống—điều kiện tiên quyết để duy trì lợi thế cạnh tranh trên thị trường trí tuệ nhân tạo.

Tiêu đề	Ngày xuất bản	URL
Scaling Trillion-Parameter Models with PP	20 Jan, 2026	Link
Elon Musk's xAI brings 1GW Colossus 2 AI training cluster online	20 Jan, 2026	Link
TLDR is hiring a Senior Software Engineer, Applied AI ($200k-$300k, Fully Remote)	20 Jan, 2026	Link
Uniqueness-Aware RL for LLM Diversity	20 Jan, 2026	Link
The Code-Only Agent	20 Jan, 2026	Link
Benchmarking AI Agent Memory: Is a Filesystem All You Need?	20 Jan, 2026	Link
UltraThink is Dead. Long Live Extended Thinking	20 Jan, 2026	Link

---

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