# Bản tin công nghệ - 20/01/2026

## Mở đầu

Trong bối cảnh các mô hình ngôn ngữ lớn (LLM) đang bứt phá lên tới hàng triệu token và yêu cầu tính toán siêu tốc, việc tối ưu hoá pipeline và giảm chi phí hạ tầng trở thành yếu tố quyết định cho các dịch vụ AI doanh nghiệp. Những tiến bộ gần đây – từ Pipeline Parallelism (PP) của SGLang với Chunked PP và Dynamic Chunking, đến kiến trúc agent đa phiên của Netomi dựa trên GPT-4.1/5.2, cùng các sáng kiến về năng lượng tái tạo và chuẩn kiểm toán AI – đang mở ra một kỷ nguyên mới cho khả năng mở rộng và độ tin cậy của trí tuệ nhân tạo.

## Điểm nhấn: Scaling Trillion-Parameter Models with PP

Pipeline Parallelism (PP) của SGLang tạo bước nhảy vọt cho triển khai mô hình LLM siêu lớn với ngữ cảnh lên tới triệu token. Ba cải tiến cốt lõi – Chunked PP, giao tiếp bất đồng bộ P2P và Dynamic Chunking – giảm truyền dữ liệu và tối ưu thời gian chờ trong pipeline. Benchmark trên cụm H20 cho thấy PP4 TP8 tăng tốc tiền xử lý (prefill) 3.31× so với TP8 và vượt TP32 tới 30.5%, đồng thời rút TTFT xuống 67.9% và duy trì hiệu suất mở rộng mạnh trên 82%. Đối với FPT, công nghệ này giúp cung cấp dịch vụ AI ngữ cảnh dài, giảm chi phí hạ tầng và mở rộng nhanh các giải pháp doanh nghiệp thông minh.

**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 agent AI doanh nghiệp dựa trên GPT-4.1 và GPT-5.2, hỗ trợ chạy đồng thời hàng nghìn phiên, tích hợp khung quản trị và khả năng suy luận đa bước để đảm bảo quy trình sản xuất ổn định. Đối với FPT, kiến trúc này giúp mở rộng dịch vụ AI quy mô doanh nghiệp, rút ngắn thời gian triển khai và nâng độ tin cậy của chatbot và trợ lý ảo.

**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 và SoftBank Group hợp tác với SB Energy xây dựng trung tâm dữ liệu AI đa gigawatt; dự án 1,2 GW ở Texas hỗ trợ Stargate. Quy mô điện tính toán này sẽ mang lại tài nguyên AI siêu tốc cho dịch vụ đám mây. Đối với FPT, mẫu quy mô lớn kết hợp năng lượng tái tạo là tham chiếu để mở rộng hạ tầng AI nội bộ, giảm chi phí và tăng độ tin cậy.

**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

Đổi mới: Tolan ra mắt trợ lý AI dạng giọng nói dựa trên GPT-5.1, tích hợp phản hồi siêu

nhanh, tái tạo ngữ cảnh thời gian thực và cá tính nhớ lâu cho hội thoại tự nhiên. Ý nghĩa với FPT: mở rộng dịch vụ AI thoại trên nền tảng đám mây/edge, nâng cao trải nghiệm người dùng và tăng khả năng cạnh tranh trong thị trường AI tương tác.

**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

"

**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 vừa đưa siêu máy tính Colossus 2 vào hoạt động, đạt công suất 1 GW – vượt mức nhu cầu điện tối đa của San Francisco – và sẽ nâng lên 1,5 GW vào tháng 4, hướng tới khoảng 2 GW. Đồng thời vòng gọi vốn Series E thu hút 20 tỷ USD, định giá công ty 250 tỷ USD, tăng tài sản ròng Musk 62 tỷ USD. Đối với FPT Smart Cloud, khả năng mở rộng quy mô tính toán ở mức đô thị mở ra cơ hội hợp tác cung cấp hạ tầng đám mây siêu tốc và dịch vụ AI quy mô 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 #6: 😼 AI needs independent auditors now

AVERI – tổ chức phi lợi nhuận mới do cựu trưởng bộ phận chính sách OpenAI thành lập – nhận 7,5 triệu USD để thúc đẩy kiểm toán độc lập cho các mô hình AI tiên tiến. Sáng kiến này đề xuất “AI Assurance Levels”, tạo chuẩn đánh giá an toàn và áp lực buộc các nhà cung cấp AI chịu kiểm định bên thứ ba. Đối với FPT, việc chuẩn bị đáp ứng tiêu chuẩn này sẽ hỗ trợ tuân thủ EU AI Act, giảm rủi ro pháp lý và nâng cao uy tín trong lĩnh vực AI tại Việt Nam.

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

## News #7: Benchmarking AI Agent Memory: Is a Filesystem All You Need?

Letatta ra mắt API và Lettatta Code – khung memory-first cho phép tạo agent lưu trữ và học người dùng. Với Lettatta Filesystem chỉ đưa lịch sử hội thoại vào file, agent đạt 74 % độ chính xác trên LoCoMo, vượt mức 68.5 % của Mem0. Điều này giúp FPT phát triển AI agents với trí nhớ dài hạn mà không cần công cụ phức tạp; giảm chi phí & tăng tự động hoá.

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

## News #8: Uniqueness-Aware RL for LLM Diversity

"

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

utm_source=tldrai

### News #9: The Code-Only Agent

Thay đổi: Áp dụng mô hình Code-Only Agent, chỉ cho phép một công cụ duy nhất – execute_code – để thực thi mã Turing-complete thay vì nhiều công cụ bash, ls, grep. Quan trọng vì FPT có thể giảm độ phức tạp hệ thống agent, tăng tính kiểm chứng và tái sử dụng mã (code witness), đồng thời giảm rủi ro hallucination và lỗi công cụ. Nhờ chạy trực tiếp Python/TypeScript, thời gian triển khai dự án có thể rút ngắn tới 30 %.

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

## Kết luận

Tổng hợp lại, các cải tiến về song song hoá pipeline, kiến trúc agent đa bước và hạ tầng tính toán năng lượng xanh không chỉ nâng cao hiệu suất xử lý ngữ cảnh dài mà còn giảm đáng kể chi phí vận hành cho FPT Smart Cloud. Đối với chúng ta, việc tích hợp Pipeline Parallelism vào nền tảng đám mây sẽ giúp cung cấp dịch vụ AI ngữ cảnh dài với thời gian phản hồi nhanh hơn 30 %, trong khi áp dụng mô hình Code-Only Agent sẽ rút ngắn thời gian triển khai dự án tới 30 % và tăng tính kiểm chứng. Song song đó, chuẩn "AI Assurance Levels" của AVERI và mô hình dữ liệu trung tâm đa gigawatt sẽ định hướng chiến lược tuân thủ EU AI Act và mở rộng hạ tầng AI quy mô đô thị. Như vậy, FPT nên ưu tiên thí nghiệm PP trên cụm nội bộ, triển khai framework agent mới cho chatbot doanh nghiệp và xây dựng lộ trình hợp tác với các nhà cung cấp năng lượng tái tạo để chuẩn bị cho một nền tảng AI siêu tốc và an toàn trong tương lai.

| 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 |
| AI needs independent auditors now | 19 Jan, 2026 | Link |
| Benchmarking AI Agent Memory: Is a Filesystem All You Need? | 20 Jan, 2026 | Link |
| Uniqueness-Aware RL for LLM Diversity | 20 Jan, 2026 | Link |
| The Code-Only Agent | 20 Jan, 2026 | Link |

---

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