# Bản tin công nghệ - 20/01/2026

## Mở đầu

Trong thời kỳ nhu cầu mở rộng siêu-dài context và triển khai mô hình hàng nghìn tỷ tham số đang tăng vọt, các đột phá vừa được công bố – từ Pipeline Parallelism của SGLang cho phép xử lý tới một triệu token không gặp OOM đến các giải pháp AI-driven như Codex của Cisco-OpenAI và tích hợp OpenAI trên ServiceNow – đang định hình lại cách chúng ta xây dựng và vận hành dịch vụ AI quy mô lớn. Những tiến bộ này không chỉ hứa hẹn giảm chi phí hạ tầng và thời gian phản hồi mà còn mở ra cơ hội mới cho FPT Smart Cloud trong việc cung cấp các giải pháp đám mây AI mạnh mẽ và an toàn.

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

SGLang ra mắt Pipeline Parallelism (PP) tối ưu cho siêu-dài context, kết hợp Chunked PP, giao tiếp P2P bất đồng bộ và Dynamic Chunking. Trong môi trường đa nút H20, cấu hình PP4 + TP8 đạt tăng tốc Prefill Throughput 3.31× so với TP8 và vượt TP32 tới 30.5%. Thời gian Time-to-First-Token giảm tới 67.9% và hiệu suất mở rộng mạnh giữ mức 82.8%. Kiến trúc cho phép xử lý lên tới một triệu token mà không gặp OOM, mở đường cho các mô hình hàng nghìn tỷ tham số. Đối với FPT, công nghệ này hạ chi phí hạ tầng AI, rút ngắn độ trễ dịch vụ và tạo lợi thế cạnh tranh trong các giải pháp đám mây AI 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 (9 bài)

## News #1: Cisco and OpenAI redefine enterprise engineering with AI agents

Mới: Cisco hợp tác OpenAI ra mắt Codex – một agent AI được nhúng trực tiếp vào quy trình phát triển phần mềm, tự động hoá sửa lỗi, tăng tốc thời gian build và hỗ trợ lập trình AI-native. Đối với FPT, Codex giúp rút ngắn chu kỳ dự án khách hàng, giảm chi phí bảo trì và nâng cao năng lực cung cấp giải pháp AI-driven cho dịch vụ đám mây và phần mềm doanh nghiệp.

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

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

ServiceNow tích hợp các mô hình tiên tiến của OpenAI vào nền tảng, hỗ trợ quy trình AI như tóm tắt tài liệu, tìm kiếm ngữ nghĩa và giao diện thoại. Nhờ đó tự động hoá nhanh hơn và giảm chi phí vận hành cho khách hàng. Đối với FPT, việc kết hợp ServiceNow + OpenAI mở ra cơ hội cung cấp giải pháp SaaS AI tùy chỉnh cho doanh nghiệp Việt Nam, tăng cạnh tranh trong chuyển đổi số.

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

## News #3: Differential Transformer V2

DIFF V2 tăng gấp đôi head truy vấn mà không mở rộng KV, bỏ RMSNorm per-head và áp dụng λ riêng cho mỗi token/head. Nhờ vậy giải mã đạt tốc độ FlashAttention ngang Transformer chuẩn, giảm loss LM 0.02-0.03 ở 1 T token và hạ đỉnh gradient/ngoại lệ; đồng thời tiết kiệm ≈25 % tham số attention để tái phân bổ. Đối với FPT, giúp tăng tốc dịch vụ đám mây và cắt giảm chi phí huấn luyện LLM lớn.

**Ngày xuất bản:** 20 Jan, 2026 **Nguồn:** Huggingface Blog **URL:** https://huggingface.co/blog/microsoft/diff-attn-v2

## News #4: Elon Musk's xAI brings 1GW Colossus 2 AI training cluster online

XAI đưa Colossus 2 vào hoạt động – siêu máy tính đào tạo AI công suất 1 GW, sẽ tăng lên 1.5 GW vào tháng 4 và mục tiêu ~2 GW; chứa hơn một triệu GPU H100 và nhận vốn Series E 20 tỷ USD. Đối với FPT Smart Cloud, năng lực quy mô thành phố này cho phép cung cấp dịch vụ AI đám mây nhanh hơn, rút ngắn thời gian huấn luyện và tăng sức 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 #5: Google Gemini Flaw Let Attackers Access Private Calendar Data

Google Gemini trong Calendar có lỗ hổng cho phép kẻ tấn công nhúng lệnh ngôn ngữ ẩn vào mô tả lời mời và lấy dữ liệu lịch riêng mà không cần người dùng tương tác (zero-click). Với FPT đang dùng Google Workspace + AI trợ lý, nguy cơ rò rỉ thông tin nội bộ ảnh hưởng tới bảo mật khách hàng và tuân thủ quy định, cần rà soát ngay quyền truy cập của mô hình.

**Ngày xuất bản:** 20 Jan, 2026 **Nguồn:** TechRepublic **URL:** https://www.techrepublic.com/article/news-google-gemini-flaw-private-calendar-data/

## News #6: The Code-Only Agent

Thay đổi: Giới thiệu mô hình 'Code-Only Agent' chỉ dùng một công cụ duy nhất – execute_code – thay cho chuỗi tool như bash, ls, grep. Quan trọng vì LLM tạo mã thực thi, cung cấp bằng chứng code có semantics xác định, giảm rủi ro hallucination và cho phép tái sử dụng script kiểm tra lại. Đối với FPT, mô hình này giúp xây dựng dịch vụ AI đáng tin cậy và giảm chi phí phát triển công cụ tới 50 %.

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

## News #7: Introducing Waypoint-1: Real-time interactive

## video diffusion from Overworld

Waypoint-1 mang lại khả năng khuếch tán video thời gian thực, cho phép người dùng điều khiển môi trường bằng văn bản, chuột và bàn phím mà không gặp độ trễ. Mô hình được huấn luyện trên 10 000 giờ footage trò chơi đa dạng và chạy trên RTX 5090 đạt ~30 000 token-passes/giây, cho 30 FPS ở 4 bước hoặc 60 FPS ở 2 bước. Đối với FPT, công nghệ này mở đường cho dịch vụ game-cloud và AR/VR tương tác ngay trên hạ tầng đám mây nội bộ, giảm chi phí phát triển nội dung và tăng trải nghiệm người dùng.

**Ngày xuất bản:** 20 Jan, 2026 **Nguồn:** Huggingface Blog **URL:** https://huggingface.co/blog/waypoint-1

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

Mô hình "Uniqueness-Aware RL" mới được đề xuất cho các LLM nhằm thưởng cho các đáp án hiếm và sáng tạo, thay vì chỉ tối ưu độ chính xác truyền thống. Đây là bước tiến quan trọng giúp FPT nâng cao khả năng tạo nội dung độc đáo trong chatbot, công cụ hỗ trợ thiết kế và giải quyết vấn đề phức tạp, đồng thời tạo lợi thế cạnh tranh trên thị trường AI sáng tạo. Phiên bản v2 (arXiv:2601.08763v2) đã cập nhật thuật toán để tăng tỷ lệ phát sinh ý tưởng mới lên 30 % so với mô hình gốc.

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

## News #9: Bandcamp and Sweden Hit Back Against AI-Generated Music

Bandcamp cấm mọi bản nhạc được tạo 'toàn bộ hoặc phần lớn' bằng AI; IFPI Thụy Điển loại bỏ ca khúc "Jag vet, du är inte min" – hơn 5 triệu lượt stream – vì chủ yếu do AI sản xuất. Điều này cho thấy rủi ro vi phạm bản quyền khi nội dung tự động lan tràng. FPT nên áp dụng chính sách bảo vệ sáng tác và duy trì uy tín dịch vụ đám mây âm thanh.

**Ngày xuất bản:** 20 Jan, 2026 **Nguồn:** TechRepublic **URL:** https://www.techrepublic.com/article/news-bandcamp-sweden-ai-music/

## Nghiên cứu khoa học nổi bật (0 bài)

## Kết luận

Những xu hướng tuần này cho thấy ba trục chính đang thúc đẩy chuyển đổi số tại FPT: (1) Tối ưu hoá kiến trúc mô hình lớn – Pipeline Parallelism, DIFF V2 và Uniqueness-Aware RL – giúp tăng tốc huấn luyện và tạo nội dung sáng tạo đồng thời cắt giảm chi phí tài nguyên; (2) Tự động hoá quy trình phát triển phần mềm và dịch vụ SaaS qua Codex và ServiceNow + OpenAI, mang lại lợi thế cạnh tranh trong cung cấp giải pháp AI-driven cho khách hàng doanh nghiệp; (3) An ninh và tuân thủ – từ lỗ hổng Gemini trong Google Workspace tới rủi ro bản quyền âm nhạc do AI – yêu cầu chúng ta ngay lập tức rà soát quyền truy cập mô hình và thiết lập chính sách bảo vệ nội dung. Đề xuất hành động ngay bao gồm thử nghiệm PP trên môi trường đa nút nội bộ để phục vụ các LLM dịch vụ khách hàng, triển khai pilot Codex cho dự án phần mềm khách hàng chiến lược, tích hợp ServiceNow + OpenAI vào danh mục SaaS của FPT và thiết lập quy trình giám sát bảo mật cũng như quản lý bản

quyền nội dung AI.

| Tiêu đề | Ngày xuất bản | URL |
|---|---|---|
| Scaling Trillion-Parameter Models with PP | 20 Jan, 2026 | Link |
| Cisco and OpenAI redefine enterprise engineering with AI agents | 20 Jan, 2026 | Link |
| ServiceNow powers actionable enterprise AI with OpenAI | 20 Jan, 2026 | Link |
| Differential Transformer V2 | 20 Jan, 2026 | Link |
| Elon Musk's xAI brings 1GW Colossus 2 AI training cluster online | 20 Jan, 2026 | Link |
| Google Gemini Flaw Let Attackers Access Private Calendar Data | 20 Jan, 2026 | Link |
| The Code-Only Agent | 20 Jan, 2026 | Link |
| Introducing Waypoint-1: Real-time interactive video diffusion from Overworld | 20 Jan, 2026 | Link |
| Uniqueness-Aware RL for LLM Diversity | 20 Jan, 2026 | Link |
| Bandcamp and Sweden Hit Back Against AI-Generated Music | 20 Jan, 2026 | Link |

---

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