Gmail Julian Andrade <jgrandrade@gmail.com>

AI Week in Review 26.01.17

1 message

Patrick McGuinness from AI Changes Everything <patmcguinness@substack.com> Sun, Jan 18, 2026 at 12:14 AM
Reply-To: Patrick McGuinness from AI Changes Everything
<reply+323d2o&3a84o1&&d81b917b5bf030d3e1271da3364f48d0118a113d3d90e30f48fc2d4c95449582@mg1.substack.com>
To: jgrandrade@gmail.com

AI insights, news, research, how-tos, thoughts and wisdom.

AI Week in Review 26.01.17

Ralph Loop, Anthropic Cowork, GLM-Image, Gemini Personal Intelligence, LongCat-Flash-Thinking-2601, GPT-5.2 Codex in API, TranslateGemma, ChatGPT Go, MedGemma 1.5 MedASR, DeepSeek's Engram, OptiMind.

PATRICK MCGUINNESS
JAN 18

READ IN APP

Figure 1. Ralph Wiggum has recently been the talk of the AI coding developer community, as Vivek on X explains. The "Ralph loop" in Claude Code adds a "verify and try again" loop to agentic coding tasks. Many users found it remarkably useful for completing some agentic coding tasks, so the method has gone viral.

Top Tools

Cowork is designed to make using Claude for new work as simple as possible. You don't need to keep manually providing context or converting Claude's outputs into the right format, nor do you have to wait for Claude to finish before offering further ideas or feedback: you can queue up tasks and let Claude work through them in parallel. - Anthropic

Anthropic launched Claude Cowork, a desktop agent based on Claude Code designed to make agentic AI workflows accessible to non-developers for general computer-based knowledge work. Claude Cowork wraps Claude Code in a user-friendly interface and allows Claude to read and write files on a user's Mac. It also can navigate websites if you install Claude's Chrome plugin and can use Anthropic's Connectors framework to access third-party apps like Canva.

Available as a research preview, Claude Cowork is available for now only on macOS desktop, but **Anthropic opened up Claude Cowork to anyone with**

a $20 subscription. The product was built entirely using Claude Code in about a week and a half, showing the power of Claude Code's agentic capabilities, so expect enhancements including wider accessibility soon.

Anthropic introduced Labs, an expansion of their team and efforts focused on incubating experimental products building on Claude's capabilities. Anthropic is realizing great success in building the AI ecosystem beyond the frontier AI model, with Claude Code and standardization around Model Context Protocol (MCP) and Claude Skills, and this effort is intended to leverage that success and grow their strength in AI ecosystem products.

AI Tech and Product Releases

Z.AI introduced GLM-Image, the first open auto-regressive image generation model, sharing model architecture features of Nano Banana Pro and GPT-image models. GLM-Image adopts a hybrid architecture combining an auto-regressive module with a diffusion decoder, and this gives it advantages in text rendering and precise instruction-following, with benchmarks on par with leading image models. This model is available on Hugging Face and through an API for developers.

Figure 2. GLM-Image is open and adopts the autoregressive architecture of leading image models like GPT-Image. They are the first, but other Chinese AI image model makers may follow.

Google debuted Personal Intelligence for Gemini, a new opt-in feature that lets users securely connect Gmail, Photos, YouTube and Search to Gemini to answer personal, contextual queries better. It's launching to Google AI Pro and Ultra subscribers with stated privacy limits, promising that data will not be used to train models directly. Google says it will expand to more countries and bring it to AI Mode in Search.

Google's Flow tool, based on Veo 3.1, is now available to Business, Enterprise, and Education Workspace plans. Flow with Veo 3.1 generates 8-second clips from prompts and supports vertical video and scene stitching.

Meituan introduced the agentic reasoning model LongCat-Flash-Thinking-2601, an open (MIT license) 560B parameter mixture-of-experts (MoE) model with 27B active parameters. It is optimized for agentic reasoning tasks and features a "Heavy Thinking" mode that achieved perfect scores on AIME-25 benchmarks. The model is positioned for advanced tool use and reasoning in complex workflows.

OpenAI made GPT-5.2 Codex model available via its API after previously exposing it only through the Codex product, marking an expansion of access for developers. The model achieves state-of-the-art results on benchmarks like SWE-Bench Pro and Terminal-Bench 2.0 and introduces native context compaction to support longer, agentic tasks. It has been integrated into tools such as Cursor, GitHub Copilot, VS Code / RooCode, and **OpenCode**.

Cursor used GPT-5.2 Codex in a long-running autonomous agent system to build *FastRenderer*, a complete web browser of over a million lines of code written from scratch in Rust. Scaling up GPT-5.2 Codex-based AI coding agents, they were able to accomplish this feat in just a week, pushing new limits in rapid agentic software development.

Google introduced TranslateGemma, a suite of open-source translation models available in 4B, 12B, and 27B parameter sizes built on its Gemma 3 architecture. Using SFT and RL to fine-tune these models for translation across 55 languages, they produced efficient high-performance models for building multilingual applications. The 4B variant is capable of running entirely on-device, broadening options for mobile device translation. Google has a demo cookbook useful for developers and a Technical Report.

NotebookLM has launched a Data Tables feature that allows users to organize, view, and manipulate information in NotebookLM more effectively. An example prompt that could be used:

> ***"Make a table of clinical trial outcomes, with columns: Study Name / Year, Intervention Method, Sample Size (N), Primary Outcome Statistics (including p-values)."***

OpenAI launched ChatGPT Go globally, expanding their low-cost $8/month Go tier from India to all nations, including USA. The $8/month Go plan has

higher limits than free ChatGPT and access to newer models, but with lower limits than Plus and Pro.

In a companion post to the Go plan, **OpenAI laid out their plans for rolling out ads in ChatGPT**, as well as their principles (answer independence, privacy, choice) on how they will be implemented. OpenAI will begin testing ads on free and Go tiers to help monetize these users further, while Plus, Pro, and Enterprise tiers will remain ad-free. OpenAI's goal is broader access without degrading trust in answers.

Google Research has released MedGemma 1.5 and MedASR, iterating on their open AI models for medical diagnostics and clinical documentation. **MedGemma 1.5 4b** is an update on their **MedGemma model** for medical image and text comprehension that brings native 3D imaging support for modalities such as CT and MRI. The newly released *MedASR* is an open speech recognition model fine-tuned for medical dictation; it reportedly outperforms Whisper v3 by a significant margin on clinical dictation error rates. MedASR is available on Hugging Face and Vertex AI.

Baichuan AI announced **Baichuan-M3, an open-source 235B medical LLM.** Baichuan-M3 235B is fine-tuned from Qwen3 and is trained to explicitly model the **clinical decision-making process**. It gets SOTA 65.1% on HealthBench, outperforming GPT-5.2, with a low hallucination rate (3.5%) and supports full clinical consultation workflows. The model is designed to reason through questions, ask follow-ups, and perform differential diagnoses. **Baichuan-M3 is available on HuggingFace**.

Raspberry Pi has released a new 8GB add-on board to run AI models. Combining a Hailo 10H chip with 40 TOPS of AI performance and 8GB RAM, this upgrade enables hobbyists to run more demanding AI models locally on Raspberry Pi devices.

AI Research News

DeepSeek research published an important paper "**Conditional Memory via Scalable Lookup: A New Axis of Sparsity for Large Language Models**" that improves how AI models handle memory. The concept is conditional memory, selecting static patterns to remember (or not), and is implemented in the Engram architecture in the transformer-based model.

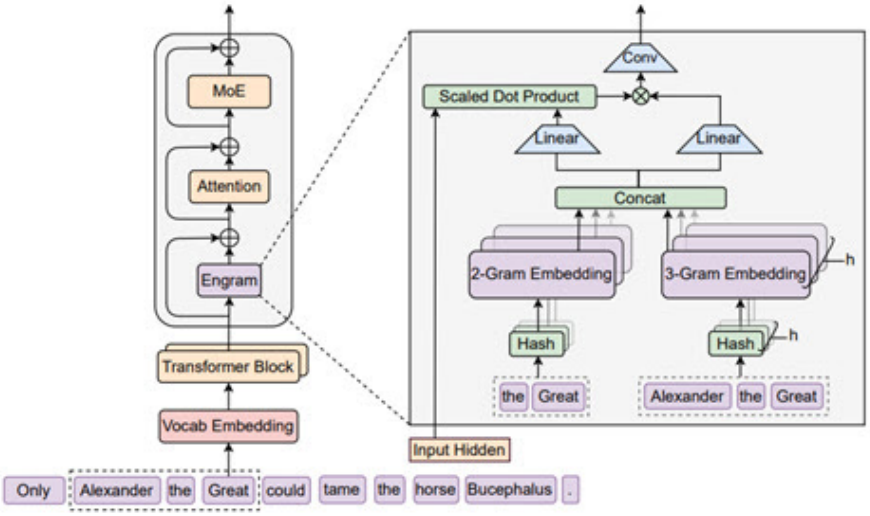

Figure 3. The Engram module augments the backbone by retrieving static N- gram memory and fusing it with dynamic hidden states via context-aware gating. This module is applied only to specific layers to decouple memory from compute.

DeepSeek showed that Engram relieves the transformer backbone from reconstructing static knowledge in early layers, thereby increasing effective depth available for complex reasoning. The result is this architecture improves performance relative to same-sized MoE architecture baselines on both knowledge-intensive tasks (improved MMLU by 3.0) but also general-reasoning tasks (improved BBH by 5.0).

DeepSeek is set to release DeepSeek V4 soon, which is rumored to be extremely strong on long context coding tasks, likely due in part to the Engram architecture innovation. DeepSeek has been the most open and innovative AI lab recently, publishing more papers on practical and fundamental AI model innovations than any other AI lab.

Microsoft Research introduced OptiMind, a 20B-parameter optimization-savvy model designed to solve business optimization problems. OptiMind translates natural-language business problems into solver-ready mathematical formulations (e.g., MILP), claiming performance comparable to larger systems and the ability to run locally for data sensitivity. Hugging Face is hosting the model for open experiments.

AI models are starting to crack high-level math problems, with **three open Erdos problems** solved with help of GPT-5.2 Pro in the last 10 days. Famed mathematician Terence Tao started a wiki page **AI contributions to Erdős problems** to cover this. It's not just GPT-5.2. **Grok 4.20 found a new**

Bellman function, **according to researchers at UCI** improving on a 30-year-old math result.

Anthropic shared a post on **how scientists are using Claude to accelerate research and discovery,** highlighting several novel uses of AI in life sciences. One is Bionmi, a biomedical agent that uses hundreds of tools and databases to design experimental protocols, formulate hypotheses, and perform analyses across many biological subfields.

AI Business and Policy

OpenAI signed a $10B compute deal with Cerebras, partnering with Cerebras to add 750MW of the AI chipmaker's ultra-low-latency compute to its AI infrastructure, with capacity coming online in tranches through 2028. **OpenAI's partnership deal** reflects OpenAI's critical need for specialized hardware to support their most advanced AI models and diversification beyond Nvidia GPUs. Reuters notes the pact could aid Cerebras' IPO bid.

To fund the AI infrastructure build-out, analysts project that **AI hyperscalers will drive up U.S. corporate bond issuance in 2026 as companies seek capital to fund AI data center expansions** and specialized hardware scaling. Barclays forecasts nearly $2.46 trillion in total issuance for the year, with significant contributions from major AI platform builders.

In a joint statement, **Apple and Google have partnered to put Gemini AI into the new Siri**. Apple's long-anticipated Siri AI update will be powered by Google's Gemini AI models and cloud tech, aiming to enhance intelligence and personalization while preserving privacy. The move is "much needed" by most observers, impatient with Apple's slow page of AI development.

Reuters reports parts suppliers halted H200 production following a Chinese customs directive to block Nvidia H200 shipments. It's unclear whether the move is temporary or a prelude to a formal ban.

The Federal Trade Commission will conduct deeper reviews of "acqui-hiring" licensing-plus-talent deals that resemble acquisitions without formal M&A, scrutinizing deals that are increasingly common in AI and made to avoid merger review complications. The FTC may require more disclosure and potential remedies even when no outright merger occurs.

Anthropic has launched an Economic Index to provide data-driven insights into how its AI systems are being utilized across various industries. The report aims to inform businesses and regulators about the practical

adoption trends of advanced AI, helping to enhance transparency regarding the reliability and economic impact of deploying Claude models in the enterprise.

Wikimedia Foundation announced **Microsoft, Meta, Amazon, Perplexity, and Mistral AI agreed to pay for enterprise access to Wikipedia content** to support Wikipedia's AI services and model accuracy. This partnership creates a new revenue stream for the Wikimedia Foundation while providing AI companies with high-volume, real-time data for their systems.

OpenAI has invested in Merge Labs, a research lab focused on bridging AI and brain-computer interfaces (BCIs). The partnership focuses on creating technology to advance the development of BCIs that facilitate human-AI collaboration, using direct neural communication with AI systems to improve accessibility and interaction speed.

Bandcamp has officially banned AI-generated music, becoming the first major audio platform to prohibit synthetic content. The policy update reinforces the platform's commitment to supporting human artistry and addresses copyright concerns within the independent music community.

The global scrutiny over xAI's Grok deepfake imagery continues, as California's attorney general sent a cease-and-desist over non-consensual sexual images allegedly generated by Grok, and regulators in the UK and elsewhere are also probing. xAI has limited Grok image features and says it's tightening safeguards.

AI Opinions and Articles

How Nano Banana got its name, from a Google PM called Naina.

AI Changes Everything *is here to help you survive and thrive the AI revolution. It is free and will remain so. If you enjoyed this post, tell others by sharing and leaving a comment. Let me know what your thoughts are. Thank you for helping grow our community!*

Share

Leave a comment

SHARE

LIKE COMMENT RESTACK

2370 Cedar Hollow Rd,, Georgetown TX, 78628