State of AI: July 31 - August 7, 2025

Grok 3

August 7, 2025

Abstract

This report provides a comprehensive overview of the significant advancements in artificial intelligence (AI) from July 31 to August 7, 2025. It covers new model releases, notable architectural improvements, and key research presented at the Association for Computational Linguistics (ACL) 2025 conference. The past week has seen a flurry of activity, with major organizations like Deep Cogito, OpenAI, DeepMind, Anthropic, and Ollama launching innovative models and services, alongside groundbreaking research in natural language processing (NLP) that enhances AI's capabilities in reasoning, world modeling, and culturally sensitive translations.

1 Introduction

The field of artificial intelligence continues to evolve rapidly, with the past week of July 31 to August 7, 2025, marking significant milestones in model development and research. This period coincides with the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025), held in Vienna, Austria, from July 27 to August 1, 2025, where researchers showcased advancements in NLP. This report synthesizes the latest model releases and architectural improvements, drawing from various sources including company announcements, news articles, and conference highlights. The developments reflect a trend toward more accessible, powerful, and culturally aware AI systems, with implications for industries ranging from education to gaming and software development.

2 New Model Releases

Several organizations have introduced new AI models or services, each contributing to the advancement of AI capabilities. Below is a detailed summary of these releases, presented in chronological order of their announcement within the specified week.

2.1 Cogito v2 by Deep Cogito

On July 31, 2025, Deep Cogito, a San Francisco-based AI startup, released Cogito v2, a suite of four hybrid reasoning models: 70B dense, 109B MoE, 405B dense, and 671B MoE. The largest model, a 671-billion parameter Mixture-of-Experts (MoE), matches or exceeds the performance of DeepSeek v3 and DeepSeek R1, approaching the capabilities of closed frontier models like OpenAI's o3 and Anthropic's Claude 4 Opus. A key innovation is the use of Iterated Distillation and Amplification (IDA), which internalizes the reasoning

Table 1: New AI Model Releases: July 31 - August 7, 2025

Model/Service	Release Date	Organization	Key Features
Cogito v2	July 31, 2025	Deep Cogito	Four hybrid reasoning models (70B dense, 109B MoE, 405B dense, 671B MoE); 671B MoE rivals DeepSeek v3, uses IDA for enhanced intuition.
gpt-oss-120b, gpt-oss-20b	August 5, 2025	OpenAI	Open-weight reasoning models under Apache 2.0; flexible, safety-evaluated.
Genie 3	August 6, 2025	DeepMind	Generates interactive 3D environments from text prompts with consistent physics.
Claude Opus 4.1	August 6, 2025	Anthropic	Improved coding and reasoning, 74.5% on SWE-bench Verified.
Ollama Turbo	August 6, 2025	Ollama	Runs large open-source models on cloud hardware for faster inference.
Gemini Storybook	August 6, 2025	Google	Creates 10-page illustrated storybooks with narration from text prompts.

process through iterative policy improvement, resulting in 60% shorter reasoning chains compared to DeepSeek R1. These models are open-source, available on platforms like Huggingface, Together AI, Baseten, and RunPod, and can be run locally with Unsloth. The combined training cost for all models was under $3.5 million, demonstrating cost efficiency (https://www.deepcogito.com/research/cogito-v2-preview).

2.2 gpt-oss-120b and gpt-oss-20b by OpenAI

On August 5, 2025, OpenAI released two open-weight reasoning models, gpt-oss-120b and gpt-oss-20b, under the Apache 2.0 license. These models mark OpenAI's first open-source release in five years, designed to be lightweight and flexible for various tasks, including integration into agentic workflows. They have undergone adversarial fine-tuning and safety evaluations using OpenAI's Preparedness Framework, ensuring robustness and ethical considerations. The models can be run locally on sufficiently powerful hardware, with training parameters disclosed for transparency (https://gizmodo.com/openai-finally-lives-up-to-its-name-drops-two-new-open-source-ai-models-2000639136).

2.3 Genie 3 by DeepMind

DeepMind unveiled Genie 3 on August 6, 2025, a foundation world model that generates interactive 3D environments from text prompts or images. Unlike traditional video generators, Genie 3 supports real-time navigation at 720p resolution and 24 frames per second, maintaining environmental consistency for several minutes. It features ”promptable world events,” allowing users to modify elements like weather or characters dynamically. This model builds on DeepMind’s previous work with Genie 2 and Veo 3, offering applications in gaming, AI agent training, and simulations. Currently in a limited research preview, Genie 3 is seen as a step toward artificial general intelligence (AGI) (https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/).

2.4 Claude Opus 4.1 by Anthropic

Anthropic released Claude Opus 4.1 on August 6, 2025, as an upgrade to Claude Opus 4. This model excels in agentic tasks, real-world coding, and reasoning, achieving a 74.5% score on the SWE-bench Verified coding benchmark, a 2% improvement over its predecessor. It offers enhanced precision in code refactoring and data analysis, with a 200,000-token context window. Available to paid Claude users, Claude Code, and through APIs on Amazon Bedrock and Google Cloud’s Vertex AI, it maintains the same pricing as Opus 4 (https://www.anthropic.com/news/claude-opus-4-1).

2.5 Ollama Turbo

On August 6, 2025, Ollama launched Ollama Turbo, a cloud-based service that enhances the performance of open-source models by leveraging datacenter-grade hardware. It supports models like OpenAI’s gpt-oss-20b and gpt-oss-120b, offering faster inference and reduced local resource demands. Compatible with Ollama’s CLI and API, Turbo mode emphasizes privacy by not retaining user data, with all hardware located in the United States. This service addresses the challenge of running large models on consumer GPUs (https://ollama.com/turbo).

2.6 Gemini Storybook by Google

Google introduced the Gemini Storybook feature on August 6, 2025, within the Gemini app. This tool generates 10-page illustrated storybooks with read-aloud narration from text prompts, supporting over 45 languages and various art styles like pixel art and comics. Users can upload photos or drawings to personalize stories, making it ideal for educational and family-oriented applications. The feature likely leverages Google’s Veo 3 for illustrations and text-to-speech models for narration (https://blog.google/products/gemini/storybooks/).

3 Notable Architectural Improvements and Research

The ACL 2025 conference, held from July 27 to August 1, 2025, in Vienna, Austria, featured significant research in NLP, with a focus on improving AI’s generalization and applicability to real-world scenarios. Below are key highlights, particularly from Sony AI, which presented innovative approaches to translation challenges.

Table 2: Notable Research from ACL 2025

Paper Title	Organization	Description
IdiomCE	Sony AI	Uses graph neural networks to improve idiomatic translation across Indian languages, mapping idioms culturally.
In-Domain African Languages Translation Using LLMs and Multi-armed Bandits	Sony AI	Employs a multi-armed bandit approach to select optimal translation models for African languages in specific domains.
Native Sparse Attention: Hardware-Aligned and Natively Trainable Sparse Attention	DeepSeek	Introduces a sparse attention mechanism for faster and cheaper model performance, winning a best paper award.

3.1 Sony AI's Contributions at ACL 2025

Sony AI presented two papers at ACL 2025, addressing translation challenges for underrepresented languages: - **IdiomCE**: This system uses graph neural networks (GNNs) to enhance idiomatic translation across Indian languages like Hindi, Tamil, Telugu, and Bengali. Unlike traditional word-by-word translation, IdiomCE maps idioms across cultures, ensuring translations retain cultural and contextual significance. For example, it translates phrases like "bury the hatchet" into culturally appropriate equivalents (https://ai.sony/blog/New%2520Research-at-ACL-2025-Tackles-Real-World-Translation-Challenges/). - **In-Domain African Languages Translation Using LLMs and Multi-armed Bandits**: This research tackles the challenge of selecting the best translation model for specific domains (e.g., news, religion) in African languages. By treating model selection as a multi-armed bandit problem from reinforcement learning, the approach dynamically chooses the most effective model, improving translation accuracy (https://ai.sony/blog/New%2520Research-at-ACL-2025-Tackles-Real-World-Translation-Challenges/).

3.2 Other Notable ACL 2025 Research

- **Native Sparse Attention by DeepSeek**: This paper, which won a best paper award, introduces a hardware-aligned sparse attention mechanism that enhances model efficiency, making AI models faster and more cost-effective. It highlights a trend toward optimizing large language models for practical deployment (https://eu.36kr.com/en/p/3401632759482502). - **Generalization in NLP Models**: The ACL 2025 theme track focused on generalization, exploring how models can perform robustly on diverse data. Papers like those from Capital One on scaling laws and USC's Information Sciences Institute on pedagogical utility of LLMs underscore efforts to make AI more reliable and applicable in real-world settings (https://www.capitalone.com/tech/ai/acl-2025/, https://www.isi.edu/news/79180/isi-at-acl-2025/).

4 Conclusion

The week of July 31 to August 7, 2025, has been pivotal for AI, with new model releases from Deep Cogito, OpenAI, DeepMind, Anthropic, Ollama, and Google, alongside significant research at ACL 2025. These developments enhance AI's capabilities in reasoning, interactive simulations, coding, and culturally sensitive translations, making it more accessible and impactful. The open-source nature of models like Cogito v2 and gpt-oss, combined with innovative tools like Gemini Storybook, democratizes AI, while research at ACL 2025 pushes for more robust and inclusive language processing. As the AI landscape continues to evolve, these advancements set the stage for further breakthroughs in accessibility, efficiency, and ethical AI development.

5 References

- Deep Cogito. (2025). Introducing Cogito v2 Preview. https://www.deepcogito.com/research/cogito-v2-preview
- Gizmodo. (2025). OpenAI Finally Lives Up to Its Name, Drops Two New Open Source AI Models. https://gizmodo.com/openai-finally-lives-up-to-its-name-drops-two-new-oper
- Google DeepMind. (2025). Genie 3: A New Frontier for World Models. https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/
- Anthropic. (2025). Claude Opus 4.1. https://www.anthropic.com/news/claude-opus-4-1
- Ollama. (2025). Ollama Turbo. https://ollama.com/turbo
- Google. (2025). Create Personal Illustrated Storybooks in the Gemini App. https://blog.google/products/gemini/storybooks/
- Sony AI. (2025). New Research at ACL 2025 Tackles Real-World Translation Challenges. https://ai.sony/blog/New%2520Research-at-ACL-2025-Tackles-Real-World-Translation-
- 36kr. (2025). DeepSeek's Liang Wenfeng & Peking University's Yang Yaodong's Team Win Best Paper Award at ACL 2025. https://eu.36kr.com/en/p/3401632759482502