NEWMIND AI JOURNAL WEEKLY CHRONICLES

01.3.2025 - 10.3.2025

This first week of March 2025 was marked by significant advancements in AI models, applications, and governance. Key highlights include Amazon's launch of the AI-powered Alexa+, Meta's unveiling of the voice-centric LLAMA 4, and OpenAI's release of GPT-4.5. Microsoft challenged OpenAI with new reasoning AI models, while Mistral AI introduced Mistral OCR, a groundbreaking document understanding tool.

In AI hardware, Chinese scientists developed the world's first carbon-based AI microchip using a ternary logic system, and a cross-border investigation unfolded around Nvidia's high-end AI chips being illegally routed to unauthorized destinations.

The European Commission withdrew its proposed AI Liability Directive, revealing a divide in global AI governance strategies.

I. MODELS					
#	Highlights	Summary	Author	Source	Date
1.1	**Amazon Launches AI-Powered Alexa+**	Amazon announced Alexa Plus (Alexa+), an ambitious new version of its virtual assistant powered by generative AI. Alexa+ is designed to be more conversational, intuitive, and capable than its predecessor. It can handle complex tasks like making reservations, managing smart homes, summarizing topics, and providing personalized assistance across entertainment, learning, and shopping. By leveraging advanced AI to enable more natural dialogues and proactive help, Alexa+ aims to redefine user experience and firmly position Amazon ahead in the competitive virtual assistant market. By leveraging advanced AI to enable more natural dialogues and proactive help, Alexa+ aims to redefine user experience and firmly position Amazon ahead in the competitive virtual assistant mark.market.	By Panos Panay, SVP of Devices & Services		February 26, 2025
1.2	**Meta Unveils Voice-Centric LLaMA 4**	Meta Platforms is expanding into voice-driven AI with its upcoming LLaMA 4 model. Described as an "omni model" for native speech interaction, LLaMA 4 enables fluid, two-way voice conversations without first converting speech to text. Meta envisions voice as the future of AI agents, allowing users to talk to AI as naturally as to a person. The company has invested heavily (up to $65 billion in 2025) toward AI, aiming for LLaMA 4 to power more interactive AI assistants in	By PYMNTS		March 7, 2025

I. MODELS					
#	Highlights	Summary	Author	Source	Date
		devices like smart glasses and to enhance services such as customer support and real-time translation. By making AI responses interruptible and more context-aware, Meta's voice-centric model seeks to make AI-driven communication seamless and conversational			
1.3	**Evo 2: Largest AI Genome Model**	Researchers from the Arc Institute, Stanford, and NVIDIA unveiled Evo 2, the largest AI model for biology to date. Evo 2 was trained on 128,000 genomes (about 9.3 trillion nucleotides) spanning all domains of life, enabling it to generate entire chromosomes and small genomes and interpret DNA sequences.	By NVIDIA		February 19, 2025
1.4	**OpenAI's GPT-4.5 Release**	OpenAI has unveiled GPT-4.5, code-named "Orion," its largest and most advanced AI model to date. This release emphasizes enhanced anthropomorphic features, including a deeper understanding of emotions and more intuitive communication, aiming to provide users with a more natural conversational experience. Despite its increased capabilities and computational demands, OpenAI has provided limited details regarding the model's exact improvements. Initially, GPT-4.5 is available to ChatGPT Pro subscribers, with plans for broader access in the future. This development underscores OpenAI's ongoing efforts to balance its ambitions in artificial general intelligence with its commercial objectives.	By Reece Rogers		Mar 6, 2025
1.5	**Microsoft Challenges OpenAI with Reasoning-AI**	Microsoft is reportedly developing new AI reasoning models to compete with OpenAI's advanced systems. These models focus on improved logical reasoning and decision-making capabilities, positioning Microsoft as a strong competitor in the field of general AI development. This initiative highlights Microsoft's commitment to advancing AI technologies that can perform complex cognitive tasks, potentially reshaping the landscape of AI applications in business and technology sectors.	By Reuters		March 7, 2025
1.6	**Mistral OCR Redefines Document AI**	On March 6, 2025, Mistral AI launched Mistral OCR, a groundbreaking document understanding tool. It surpasses competitors like Google Document AI and GPT-4o by accurately interpreting complex elements: media, text, tables, and equations, especially in multilingual and mathematical contexts. Processing 2000	By Mistral AI Team		Mar 6, 2025

I. MODELS					
#	Highlights	Summary	Author	Source	Date
		pages per minute, it's exceptionally fast. Mistral OCR's "Doc-as-Prompt" feature outputs structured JSON, streamlining AI integrations. It natively supports thousands of scripts and fonts, ensuring broad language coverage. Available via API, Le Chat, and soon cloud partners, it also offers self-hosting for enhanced security. This innovation is transforming fields like scientific digitization and legal document indexing, enabling efficient, large-scale knowledge extraction.			
1.7	**QwQ-32B: Smaller, Smarter AI Model**	Qwen team introduced QwQ-32B, a 32 billion parameter AI model that rivals the much larger DeepSeek-R1 in performance. The key breakthrough lies in scaling Reinforcement Learning (RL) to enhance reasoning, tool use, and adaptability. The model excels in mathematical reasoning, coding, and problem-solving, as shown by benchmarks like AIME24, LiveCodeBench, and BFCL. QwQ-32B's success demonstrates that RL can bridge the gap between model size and performance, offering an efficient alternative to conventional AI training methods.	By Ryan Daws		
1.8	**Microsoft Unveils Compact Phi-4 Models**	Microsoft's Phi-4 series debuts with Phi-4-Mini, a 3.8B parameter language model excelling in math and coding. Trained on curated web and synthetic data, it rivals larger models in complex reasoning. Phi-4-Multimodal expands this, integrating text, vision, and speech. Utilizing LoRA adapters and modality routers, it enables diverse inference modes without interference. Notably, it leads the OpenASR leaderboard with a small speech component. It adeptly handles vision-language, vision-speech, and speech-audio tasks, surpassing larger models. These compact models demonstrate significant advancements in efficient and versatile AI processing.	By Microsoft		Feb 27, 2025
1.9	**AI21 Labs Launches Jamba, Maestro**	AI21 Labs unveiled Jamba 1.6 on March 6, an open model for private enterprise. Outperforming rivals like Llama 3.3 70B, it features a 256K context window and a hybrid Mamba-Transformer MoE architecture, balancing performance and cost. Accessible via AI21 Studio, Hugging Face, or private deployment, it offers flexibility and security. On March 10, they launched Maestro, an AI planning system optimizing LLM tasks. Maestro breaks complex prompts into substeps, ensuring	By AI21 Editorial Team		Mar 6, 2025

I. MODELS					
#	Highlights	Summary	Author	Source	Date
		accuracy through simulations and user-defined requirements. This enhances AI output reliability, benefiting document analysis and process automation.			
1.10	**Google Unveils Gemini Embedding Model**	On March 7, 2025, Google introduced a new text embedding model based on its Gemini AI architecture. This model is now available to developers through the Gemini API, under the identifier gemini-embedding-exp-03-07. It generates state-of-the-art embeddings for words, phrases, code, and sentences, enabling applications such as semantic search, text classification, and clustering. This release underscores Google's commitment to advancing natural language understanding and providing developers with robust tools to enhance their AI-driven applications.	By Logan Kilpatrick, Zach Gleicher, Parashar Shah		March 7, 2025,
1.11	**Vision-R1: Multimodal Reasoning Redefined**	Vision-R1 is a multimodal large language model (MLLM) designed to improve reasoning capabilities through reinforcement learning (RL). Inspired by DeepSeek-R1-Zero, it addresses reasoning challenges in MLLMs by integrating cold-start initialization with RL training. High-Quality Dataset: A 200K multimodal Chain-of-Thought (CoT) dataset was generated without human annotation, enhancing reasoning capabilities. Progressive Training: Introduces Progressive Thinking Suppression Training (PTST) to prevent overthinking and refine logical steps. Performance: Achieves 73.5% accuracy on MathVista, nearly matching OpenAI O1 and outperforming some models with 70B+ parameters despite being only 7B. Vision-R1 demonstrates strong reasoning skills by incorporating human-like cognitive processes, such as questioning and reflection.	By Wenxuan Huang and Bohan Jia and Zijie Zhai and Shaosheng Cao and Zheyu Ye and Fei Zhao and Zhe Xu and Yao Hu and Shaohui Lin		11 Mar 2025
1.12	**EuroBERT**	A collaboration between CentraleSupélec's MICS laboratory, Diabolocom, Artefact, and Unbabel, supported by AMD and CINES, introduced EuroBERT—a state-of-the-art multilingual encoder model. Trained on a vast dataset of 5 trillion tokens, EuroBERT supports 15 languages, including major European and widely spoken global languages, and excels in tasks related to mathematics and programming. Its architecture incorporates advanced features such as grouped query attention and rotary position embeddings, enabling it to handle sequences of up to 8,192 tokens. EuroBERT is available in three sizes—210 million, 610	By Nicolas-BZRD, Hippolyte Gisserot-Boukhlef, Duarte Alves, Manuel Faysse		March 10, 2025

I. MODELS					
#	Highlights	Summary	Author	Source	Date
		million, and 2.1 billion parameters—offering optimal performance across various natural language processing tasks like retrieval, classification, and regression. The model is open-sourced under the Apache 2.0 license and accessible on platforms like Hugging Face, promoting further research and application development in multilingual NLP.			
1.13	**AMD Launches Open-Source Instella LLM**	AMD has introduced Instella, a series of fully open-source language models featuring 3 billion parameters, trained on 128 Instinct MI300X GPUs. The Instella-3B model comprises 36 decoder layers with 32 attention heads each, supporting a sequence length of up to 4,096 tokens and utilizing a vocabulary of approximately 50,000 tokens. The training process incorporated advanced techniques such as FlashAttention-2, Torch Compile, and Fully Sharded Data Parallelism (FSDP) with hybrid sharding to enhance performance and resource efficiency. Notably, Instella-3B demonstrates superior performance compared to existing fully open models of similar sizes and achieves competitive results against state-of-the-art open-weight models like Llama-3.2-3B and Qwen-2.5-3B. AMD has made all related artifacts, including model weights, training configurations, datasets, and code, publicly available to encourage collaboration and innovation within the AI community.	By Jiang Liu, Jialian Wu, Xiaodong Yu, Prakamya Mishra, Sudhanshu Ranjan, Zicheng Liu, Chaitanya Manem, Yusheng Su, Pratik Prabhanjan Brahma, Gowtham Ramesh, Ximeng Sun, Ze Wang, Emad Barsoum		March 5, 2025
1.14	**STORM Revolutionizes Long Video Understanding**	STORM (Spatiotemporal Token Reduction for Multimodal LLMs), a novel architecture designed to enhance long video understanding in multimodal large language models (LLMs). STORM integrates a dedicated temporal encoder, specifically a Mamba-based temporal projector, between the image encoder and the LLM. This integration enriches visual tokens with temporal dynamics, significantly improving the model's video reasoning capabilities. Additionally, STORM employs effective token reduction strategies, including test-time sampling and training-based temporal and spatial pooling, which substantially reduce computational demands without sacrificing key temporal information. Extensive evaluations demonstrate that STORM achieves state-of-the-art results across	By Nvidia		6 Mar 2025

I. MODELS					
#	Highlights	Summary	Author	Source	Date
		various long video understanding benchmarks, with more than a 5% improvement on MLVU and LongVideoBench, while reducing computation costs by up to 8 times and decoding latency by 2.4 to 2.9 times for fixed numbers of input frames.			
1.15	**Cohere Launches Multilingual Aya Vision**	Cohere has unveiled Aya Vision, a state-of-the-art multimodal large language model that excels in both language and image understanding across 23 languages. This model supports tasks such as image captioning, visual question answering, text generation, and translations from both texts and images into coherent text. Aya Vision is available in two configurations: an 8-billion parameter variant optimized for low latency and performance, and a 32-billion parameter variant designed for state-of-the-art multilingual performance. Developers can access Aya Vision through the Cohere platform, enabling the creation of applications that seamlessly integrate multilingual and multimodal capabilities.	By Cohere For AI Team		Mar 04, 2025
1.16	**DINOv2 Transforms AI Pathology Analysis**	Meta’s AI blog highlights Dr. Faisal Mahmood's research on using DINOv2 to enhance pathology analysis through foundation models. The Mahmood Lab aligns with DINOv2’s hypothesis that data diversity is more important than sheer volume for effective AI training. By applying DINOv2 to pathology images, the lab has made significant progress in medical image analysis. Their approach demonstrates that models trained on diverse and representative datasets perform better in complex fields like pathology. This advancement paves the way for more accurate disease diagnosis and understanding, improving healthcare outcomes.	By Meta Team		March 6, 2025

II. AI CHIPS					
#	Highlights	Summary	Author	Source	Date
2.1	**World's First Ternary Carbon Microchip**	Chinese scientists unveiled a pioneering carbon-based AI microchip, using ternary logic, not binary. Built with carbon nanotubes (CNTs), it processes data in three states, boosting speed and reducing energy use. Peking University researchers showed perfect handwriting recognition accuracy, demonstrating AI potential. This CNT approach advances semiconductor tech, addressing silicon limitations. While CNT integration density trails silicon (e.g., NVIDIA GPUs), this is a significant step toward next-gen AI hardware. It underscores China's ambition in post-silicon chip research, potentially transforming AI processing with efficient, high-performance chips.	By Tribune Team		March 08, 2025
2.2	**Nvidia AI Chips: Cross-Border Scandal**	A cross-border probe investigates illegal routing of Nvidia AI chips. Singapore charged three men for fraud, involving servers with U.S.-made chips, allegedly destined for Chinese AI startup DeepSeek, violating export restrictions. Dell and SuperMicro supplied the servers, shipped via Malaysia, potentially to China. An anonymous tip prompted the investigation. Singapore and the U.S. are jointly probing if export-controlled components were involved. The U.S. also investigates DeepSeek's use of restricted chips. This scandal highlights geopolitical tensions over AI hardware, as demand for AI accelerators surges. It may lead to stricter export controls to prevent illicit chip transfers.	By Bing Hong Lok		March 4, 2025
2.3	**Singapore Investigates Nvidia Chip Diversion**	Singapore's **Home Affairs Minister, K. Shanmugam**, confirmed that **Dell** and **SuperMicro** supplied the servers under investigation, which were allegedly shipped to **Malaysia** as part of a **circuitous route** leading to **China**. The investigation was sparked by an **anonymous tip**, with authorities analyzing	By Gao Yuan and Mackenzie Hawkins		March 8, 2025

II. AI CHIPS					
#	Highlights	Summary	Author	Source	Date
		whether the servers contained **export-controlled components**. They are coordinating with U.S. counterparts for a joint investigation into potential **sanctions violations**. This development could have significant impacts on the **global technology supply chain**, leading to stricter **compliance requirements** for **hardware manufacturers** and **distributors** to prevent the diversion of restricted components through intermediary countries.			

III. LLM TECHNICS AND METRICS					
#	Highlights	Summary	Author	Source	Date
3.1	**SWE-Lancer Benchmark Highlights**	OpenAI's **SWE-Lancer** benchmark, announced in March, was developed to assess the real-world software engineering capabilities of advanced language models. Based on **over 1,400** freelance tasks sourced from **Upwork**, the benchmark represents **$1 million worth** of projects. Covering various domains such as coding and project management, these tasks are evaluated through **end-to-end testing verified by professional engineers**. Initial results indicate that LLMs struggle to complete complex, multi-step software projects effectively. SWE-Lancer serves as a crucial benchmark, highlighting the need for improvements in **reasoning, long-term planning, and reliability**, making it a key measure for future advancements in AI-driven engineering.	By Daniel Dominguez		08.03.2025
3.2	**Microsoft Develops In-House Reasoning LLMs (Project "MAI"):**	Microsoft is developing its own large language models (LLMs) to reduce reliance on OpenAI. Led by Mustafa Suleyman, its AI division has trained the MAI model family, which performs competitively with OpenAI and Anthropic on benchmarks. These models prioritize chain-of-thought reasoning, enhancing multi-step decision-making. Microsoft is also testing external models from xAI, Meta, and DeepSeek as potential OpenAI alternatives for 365 Copilot. The MAI models, more advanced than the Phi series, are being trialed as GPT-4 replacements internally and may launch via API this year. This move signals Microsoft's ambition to become a leading AI provider beyond OpenAI.	By Reuters		07.03.2025
3.3	**START(Self-taught Reasoner with Tools):**	In **March 2025**, researchers introduced **START (Self-Taught Reasoner with Tools)**, an advanced **LLM** designed for complex reasoning by integrating external tools, especially **Python execution**. While traditional LLMs like **OpenAI-o1** and **DeepSeek-R1** excel in **long chain-of-thought (CoT) reasoning**, they often struggle with **hallucinations** and **inefficiencies**. Built by fine-tuning **QwQ-32B-Preview**, START achieved high accuracy on **science QA** (63.6%), **math** (95.0%, 66.7%), and	By University of Science and Technology of China and Alibaba Group		07.03.2025

III. LLM TECHNICS AND METRICS					
#	Highlights	Summary	Author	Source	Date
		coding benchmarks (47.1%, 47.3%), surpassing its base model and competing with **state-of-the-art models** like **R1-Distill-Qwen-32B** and **o1-Preview**.			
3.4	**HoT (Highlighted Chain of Thought):**	In **March 2025**, researchers introduced **Highlighted Chain-of-Thought Prompting (HoT)**, a novel technique aimed at improving the **factual accuracy** and **transparency** of **Large Language Models (LLMs)**. HoT prompts LLMs to generate responses that include **XML tags** linking facts directly to those in the query. This method involves **reformatting** the input question by highlighting key facts with **XML tags** and then generating a response that includes similar tags. This approach allows users to **trace statements** in the answer back to the original input, improving the **traceability** and **reliability** of the response.	By Tin Nguyen Logan, Bolton Mohammad, Reza Taesiri, Anh Totti Nguyen		5.3.2025
3.5	**Industry-Scale LLMs**	Major tech firms debuted new large language models. Notably, Foxconn's research arm unveiled **FoxBrain**, Taiwan's first LLM with advanced reasoning optimized for Traditional Chinese. FoxBrain is built on Meta's Llama 3.1 architecture and was trained on 120 Nvidia H100 GPUs in about four weeks. The model is being applied to manufacturing and supply-chain tasks, and Foxconn plans to open-source it for collaboration. Its performance is close to state-of-the-art, with only a slight gap versus a distilled version of China's leading model (DeepSeek). This showcases a trend of organizations customizing LLMs for specific languages and domains.	By Reuters		10.03.2025
3.6	**Multimodal AI Systems:**	On **March 2–3**, researchers introduced **TaxaBind**, a groundbreaking **multimodal LLM** that integrates **six data modalities**—including **ground-level photos, satellite imagery, audio, and text**—into a unified model. Using an innovative technique called **"multimodal patching"**, TaxaBind merges features from each modality into a **common representation**, enabling **cross-modal reasoning**. This allows **zero-shot learning** for tasks like **species identification** from images (even for unseen species) and **cross-modal retrieval** (e.g., linking animal photos to **climate data**).	By Shawn Ballard		03.03.2025

III. LLM TECHNICS AND METRICS					
#	Highlights	Summary	Author	Source	Date
		TaxaBind outperforms traditional single-modality models, demonstrating how LLMs are evolving to incorporate **vision, audio**, and other inputs for a deeper understanding in **AI**.			
3.7	**New Benchmarks & Leaderboards:**	New benchmarks and leaderboards emerged to assess LLM performance on advanced tasks. **GPQA Diamond** now tests graduate-level reasoning with complex logic questions, while **SWE Bench (Verified)** evaluates coding skills on difficult programming challenges requiring correct, executable solutions. These benchmarks feed into leaderboards like **Vellum AI's**, ranking top models by task and updating as new results arrive. Early rankings show that different models excel in different areas—some leading in reasoning (GPQA), others in coding (SWE Bench). Traditional tests like **MMLU** and **Big-Bench Hard** remain key, while new domain-specific evaluations provide deeper insights into LLM capabilities.	By Vellum AI		
3.8	**Holistic Evaluation & Real-World Testing:**	Static QA tests alone are no longer enough to evaluate modern LLMs. Researchers are developing interactive benchmarks that mimic real-world use, testing models in multi-turn environments where they must use tools, call APIs, or complete complex tasks. **Berkeley Function Calling (BFCL)** and **SWE-Lancer** simulate scenarios like writing and executing code, with automatic verification of correctness (e.g., does the code run?). These evaluations assess reasoning, planning, and tool use, moving beyond simple question-answering. As LLMs take on roles with ethical and financial stakes, ensuring reliability and multi-step consistency is critical. "Agentic" evaluations are now complementing traditional benchmarks.	By Vishakha Agrawal, Archie Chaudhury, Shreya Agrawal		5.3.2025
3.9	**Built-in Safety Reasoning**	Ensuring LLMs behave safely is crucial, and a new approach, **"Rational"** (Reasoning-Enhanced Fine-Tuning for Interpretable LLM Safety), was introduced by Carnegie Mellon in March. Instead of relying on hard-coded filters, the model learns to generate a **step-by-step safety rationale** before answering. During fine-tuning, it analyzes prompts, considers intent, and explains why a query is harmful or safe. This method improves refusals to adversarial prompts, making them more **context-aware and interpretable**. By internalizing ethical reasoning, LLMs can	By Yuyou Zhang, Miao Li, William Han, Yihang Yao, Zhepeng Cen, Ding Zhao		06.03.2025

III. LLM TECHNICS AND METRICS					
#	Highlights	Summary	Author	Source	Date
		avoid inappropriate content while reducing unnecessary refusals, marking a shift toward **built-in, explainable safety mechanisms** in AI models.			
3.10	**Bias Mitigation via Activation Steering**	In **March 2025**, researchers introduced a **bias-mitigation** method using **Steering Vector Ensembles (SVE)** to reduce **social biases** in **LLMs** without retraining. They applied **steering vectors** to model activations, using **Bayesian optimization** to adjust responses along **nine bias axes** (e.g., gender, race). These vectors, optimized on the **BBQ benchmark**, were combined into an **ensemble**, achieving up to **12% bias reduction** in models like **Mistral, Llama, and Qwen**. The **modular and interpretable** approach allows dynamic adjustments without degrading performance. SVE demonstrates how **activation engineering** can offer an **efficient, adaptable** alternative to costly **fine-tuning** for fairness in AI systems.	By Zara Siddique, Irtaza Khalid, Liam D. Turner, Luis Espinosa-Anke		07.03.2025
3.11	**Chain-of-experts (CoE)**	The **Chain-of-Experts (CoE)** framework enhances **LLM efficiency** and **accuracy** by activating specialized experts sequentially instead of all at once. This approach improves **contextual understanding** and optimizes **resource usage**, providing a **cost-effective alternative** to traditional **dense models** and **Mixture-of-Experts (MoE)** architectures. CoE offers **improved reasoning**, as experts collaborate progressively for better accuracy, while also reducing **computation costs** by lowering redundant processing and memory usage. It outperforms **MoE models** with **17.6% less memory usage** while maintaining similar accuracy. CoE provides a **scalable, efficient AI solution**, making **LLMs more accessible and sustainable**.	By Ben Dickson		10.03.2025
3.12	**Sketch-of-Thought**	Recent advances in LLMs have improved reasoning via Chain of Thought (CoT) prompting but at the cost of excessive verbosity. Sketch-of-Thought (SoT) is a new prompting framework that reduces token usage while maintaining accuracy. Inspired by cognitive science, SoT integrates Conceptual Chaining, Chunked Symbolism, and Expert Lexicons, selecting the best approach dynamically via a lightweight routing model. Tested across 15 reasoning datasets in multiple languages and modalities, SoT reduces tokens by 76% with minimal accuracy loss. In math and multi-hop reasoning, it even improves accuracy while being more efficient, making it a scalable solution for AI reasoning tasks.	By Simon A. Aytes, Jinheon Baek, Sung Ju Hwang		10.03.2025

III. LLM TECHNICS AND METRICS					
#	Highlights	Summary	Author	Source	Date
3.13	**LLM-as-a-Judge: Evaluating AI with AI**	The "Awesome LLM-as-a-Judge" initiative explores leveraging Large Language Models (LLMs) as evaluators across various domains. This approach aims to harness LLMs' capabilities to provide scalable and flexible assessments, potentially reducing reliance on traditional expert-driven evaluations. However, ensuring the reliability of LLM-as-a-Judge systems presents challenges such as maintaining consistency, mitigating biases, and adapting to diverse assessment scenarios. To address these issues, the initiative proposes strategies to enhance reliability and introduces a benchmark designed for evaluating LLM-based judgments. By outlining practical applications, challenges, and future directions, this work serves as a foundational reference for advancing AI-driven evaluation systems across multiple domains.	By Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Kun Zhang, Saizhuo Wang, Yuanzhuo Wang, Wen Gao, Lionel Ni, Jian Guo		9.3.2025
3.14	**Multi Agent Bench**	**Multi Agent Bench** is a comprehensive benchmark designed to evaluate **multi-agent systems** powered by **Large Language Models (LLMs)**. It overcomes the limitations of traditional **single-agent** and **domain-specific benchmarks** by focusing on diverse, interactive scenarios that emphasize both **collaboration** and **competition**. The framework, **MARBLE**, incorporates innovative metrics such as **milestone-based KPIs** and supports **flexible coordination protocols**. Key findings show that **graph-based coordination** and **cognitive planning** outperform other methods, highlighting **emergent social behaviors** and advancing research toward **AGI-level multi-agent collaboration**.	By Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, Jiaxuan You		3.3.25
3.15	**Meta Plan Optimization (MPO)**	Recent LLM advancements enable interactive planning, but existing methods suffer from hallucinations and require retraining for new agents. Meta Plan Optimization (MPO) enhances agent planning by incorporating explicit guidance via meta plans, avoiding reliance on complex human-curated knowledge. Unlike traditional methods, MPO continuously optimizes meta plans using feedback from	By Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai,		4.3.25

III. LLM TECHNICS AND METRICS					
#	Highlights	Summary	Author	Source	Date
		task execution, improving efficiency and generalization. Experiments on two benchmark tasks show MPO outperforms existing approaches. Analysis confirms MPO as a plug-and-play solution, boosting task completion while adapting to unseen scenarios. This framework provides a scalable way to refine agent reasoning without retraining, making AI planning more effective and flexible.	Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni and Jian Guo		
3.16	**Process-based Self-Rewarding Language Models (PSRLM)**	**Process-based Self-Rewarding Language Models"** introduces a framework to improve LLMs' reasoning, especially in mathematics. Traditional self-rewarding methods, where LLMs evaluate their own outputs, often struggle with complex reasoning and may reduce accuracy. To solve this, a **process-based approach** is proposed, incorporating **step-by-step reasoning** where LLMs judge each intermediate step. This enables **fine-grained feedback** and iterative self-improvement. The method significantly enhances performance across mathematical benchmarks, showing that LLMs can refine their reasoning **without external rewards**. This approach suggests that LLMs could achieve, or even surpass, **human-level reasoning** through structured self-optimization.	By Shimao Zhang, Xiao Liu, Xin Zhang, Junxiao Liu, Zheheng Luo Shujian Huang, Yeyun Gong		5.3.2025

IV. AI USE CASES					
#	**Highlights**	**Summary**	**Author**	**Source**	**Date**
4.1	**Google's AI Mode for Complex Questions**	Google has introduced an **AI-powered search mode** that allows users to ask **multi-part, complex questions**. This update improves search by providing **contextual** and **conversational responses**, enhancing the experience for **in-depth queries**. Reflecting Google's continuous effort to refine **AI-driven search technologies**, this development has the potential to **transform how users interact with search engines** and access information. By offering **more nuanced and relevant answers**, it paves the way for a **more intuitive and dynamic search experience**.	By Aisha Malik		
4.2	**SOFYA**	**Sofya** integrates **Llama** models for **real-time healthcare solutions**, hosting them on **Oracle Cloud** and utilizing **Sglang** and **VLLM** for efficient model serving. To meet **low-latency demands**, the team fine-tuned smaller **Llama versions (8B, 3B, 70B)** using **knowledge distillation** and **self-reflection prompt engineering**. Llama automates **data structuring, entity recognition, and question answering**, reducing errors and boosting efficiency. This has led to **30% less time spent on documentation**, improved workflows, and an **average CSAT score of 90%**. Sofya plans to scale to **1 million consultations per month**, expanding its **AI-driven agent flow** with **Llama 70B** for **real-time healthcare applications**.	By Meta Team		5.3.254.3

IV. AI USE CASES					
#	**Highlights**	**Summary**	**Author**	**Source**	**Date**
4.3	**AGNTCY**	On **March 6, 2025**, **Cisco, LangChain, Galileo, Glean, and LlamaIndex** launched **AGNTCY**, an open-source collective for **AI agent interoperability**. AGNTCY aims to establish a **standardized communication framework**, like **TCP/IP**, to enable seamless interaction between AI platforms. This **"Internet of Agents"** promotes **collaboration, efficiency, and scalability** across AI ecosystems. **Cisco's Outshift** emphasizes its role in connecting AI systems from different vendors. By engaging the AI and infrastructure community, AGNTCY seeks to build an **open, interoperable foundation**, addressing **multi-agent collaboration challenges** and advancing next-generation AI applications globally.	By Emilia David		6.3.25
4.4	**Manus**	China introduced **Manus**, a **general AI agent** developed by **Monica**, gaining attention for its **autonomous capabilities**. Designed to **think, plan, and execute tasks independently**, Manus is compared to leading AI systems from **OpenAI, Google, and Anthropic**. Reports from **Newsweek and Business Standard** highlight its ability to **build websites, plan trips, and analyze stocks** without human supervision. This advancement has sparked **concerns in Silicon Valley** over **China's AI leadership**, raising **ethical and regulatory questions** regarding **accountability** in **autonomous AI systems**. Manus represents a **major step** in AI-powered industries, potentially giving China a **first-mover advantage** globally.	By Kyle Wiggers		9.3.2025

V. AI POLICIES, REGULATIONS & STRATEGIES					
#	Highlights	Summary	Author	Source	Date
5.1	**EU Halts AI Liability Law Amid Disagreements**	The European Commission withdrew its AI Liability Directive, citing a lack of stakeholder agreement. Intended to establish AI accountability, its withdrawal followed criticism that it stifled innovation. EU officials will now reassess AI accountability, considering alternative frameworks. The decision sparked debate: some see it as a consumer protection setback, others as pragmatic to avoid premature regulation. This highlights the innovation-oversight tension. Without the directive, questions remain on legal redress for AI-caused damages. The EU's next regulatory move is highly anticipated, shaping AI confidence across member states.	By Tim Wright, Nathan Evans		07/03/2025
5.2	**Global AI Summit Reveals Governance Divide**	The Paris AI Summit in March 2025 exposed a global AI governance split. 57 nations, including China and India, signed a declaration for "inclusive" AI, emphasizing ethics. The U.S. and UK declined, citing security concerns and lack of clarity. Yoshua Bengio's AI Safety Report 2025 offered risk mitigation blueprints. The UK rebranded its AI Safety Institute to AI Security Institute, prioritizing national security. This highlights a divide: collective governance versus sovereign control. Reconciling these views is crucial for global AI innovation and trust.	By Tim Wright, Nathan Evans		07/03/2025
5.3	**Anthropic Secures $3.5B, Expands Claude**	On March 3, 2025, Anthropic secured $3.5 billion in Series E funding, reaching a $61.5 billion valuation. Led by Lightspeed Venture Partners and supported by investors like Salesforce Ventures, Cisco Investments, and Fidelity, the funds will enhance Anthropic's next-generation AI development, computing capacity, interpretability research, and global expansion. Recent launches, Claude 3.7 Sonnet and Claude Code, significantly improved AI-driven coding capabilities. Claude is now integrated into platforms like Replit's Agent and Thomson Reuters' CoCounsel, boosting productivity across industries, including healthcare and finance. Novo Nordisk notably reduced clinical report-writing time using Claude.	By Antrophic Team		Mar 3, 2025
5.4	**Trump Shifts AI Policy Focus**	The Trump Administration has shifted U.S. AI policy to prioritize innovation and leadership over new regulation. A January 23 executive order, *"Removing Barriers to American Leadership in AI,"* revoked earlier AI directives and launched an AI	By Michael Charalambous		06 March, 2025

V. AI POLICIES, REGULATIONS & STRATEGIES					
#	**Highlights**	**Summary**	**Author**	**Source**	**Date**
		Action Plan aimed at sustaining U.S. dominance in AI. The White House is actively soliciting public input on this plan through mid-March. Meanwhile, state-level lawmakers are stepping in: Virginia recently **passed a comprehensive AI law** to curb algorithmic bias in high-risk systems and ensure transparency, making it one of the first U.S. states with broad AI legislation.			
5.5	**China Boosts Strategic AI Investment**	AI remains a strategic priority in China's latest government agenda. At the National People's Congress on March 5, Beijing announced plans to **boost support for AI R&D** – backing large-scale AI models, "industries of the future" (e.g. embodied AI and 6G), and venture funding – to drive tech breakthroughs and self-reliance. This marks the first time China's annual work report explicitly mentioned AI models, following the global buzz around a Chinese startup's advanced AI system. Regulators in China continue to enforce strict rules on deepfakes and generative AI content, but the government's message in 2025 underscores an *"innovation-friendly"* approach alongside security oversight.	By Reuters		March 5, 2025
5.6	**UK Delays AI Legislation, Debate Continues**	The **UK government** has paused plans for sweeping **AI-specific legislation**, delaying the anticipated **national AI bill** until at least **summer 2025**, potentially aligning with the U.S.'s **lighter regulatory approach**. Meanwhile, a **House of Lords** member reintroduced a **private AI regulation bill** on **March 4**, sparking debate on AI governance. Internationally, countries like **Canada** and **Japan** have joined the **Council of Europe's new AI Convention**, the first global AI treaty. **Switzerland** plans to ratify this convention, aiming to balance **innovation** with **human rights** and **AI transparency** safeguards.	By Michael Charalambous		06 March, 2025
5.7	**OpenAI Faces Criticism on Alignment**	In **March 2025**, OpenAI reaffirmed its commitment to **AI safety** and **ethics** in a blog post, outlining efforts to ensure future **frontier AI** systems are **controllable** and **beneficial**. However, **Miles Brundage**, a former **OpenAI policy lead**, criticized the post, accusing the company of **rewriting the history** of its **AI safety journey**. He argued that OpenAI downplayed past internal debates, such as concerns over releasing **GPT-2** in 2019, and painted an overly positive picture of	By ResearchBuzz		10March, 2025

V. AI POLICIES, REGULATIONS & STRATEGIES					
#	**Highlights**	**Summary**	**Author**	**Source**	**Date**
		its vigilance. This incident highlights ongoing tensions in the AI industry between **rapid development** and **transparency**.			
5.8	**Bias Mitigation & Fairness:**	Recent regulations, such as Virginia's High-Risk Artificial Intelligence Developer and Deployer Act (HB 2094), require developers to prevent "algorithmic discrimination" by auditing AI systems for bias. Similarly, global frameworks like the EU AI Act and ISO/IEC 42001 emphasize robust data governance, advocating for high-quality, representative datasets and comprehensive bias testing. Standardized datasets are emerging to evaluate facial recognition technologies across diverse demographics, aiding in quantifying ethical performance. Organizations increasingly recognize fairness audits as essential, adopting practices akin to security evaluations. Despite ongoing challenges, a consensus is building around routinely embedding fairness evaluations into AI development processes to mitigate algorithmic discrimination.	By Reena R. Bajowala, Wouter van Wengen of Greenberg Traurig		March 5, 2025