







NEWMIND AI JOURNAL WEEKLY CHRONICLES




17.3.2025 - 25.3.2025



The third week of March 2025 has underscored the accelerating pace of artificial intelligence innovation, marked by groundbreaking advancements across models, hardware, and real-world applications. Leading this charge is NVIDIA, which unveiled transformative tools like the Isaac GROOT N1 humanoid robot foundation model and the Rubin Ultra GPU architecture, signaling a leap in AI-driven robotics and hardware efficiency. Meanwhile, open-source initiatives dominated the landscape, with Roblox’s Cube 3D democratizing 3D object generation and AMD’s Gaia project enabling localized large language model execution on consumer hardware, challenging NVIDIA’s dominance. Multimodal AI took center stage, as Tencent’s 3D generation tools and Manycore Research’s SpatialLM-Llama-1B bridged unstructured 3D data with actionable insights, enhancing applications in robotics and spatial analysis. Concurrently, policy debates intensified, with Europol warning of AI-driven crime and the EU advancing its Chips Act 2.0 to secure semiconductor independence. From healthcare innovations like AI-powered physiotherapy to Disney’s collaborative robotics with NVIDIA, this week’s developments highlight AI’s expanding role in reshaping industries while navigating the dual imperatives of innovation and ethical governance.




I. MODELS					
#	Highlights	Summary	Author	Source	Date
1.1	Roblox Unveils Cube 3D: An Open-Source AI Model for 3D Object Generation	Roblox introduced Cube 3D, a foundational AI model capable of generating 3D objects and environments. This model was launched alongside the beta version of the Mesh Generation API, which allows users to create 3D objects using simple text commands. Cube 3D tokenizes 3D objects and predicts the next shape token, similar to language models. By releasing Cube 3D as open source, Roblox enables developers to fine-tune the model, create extensions, or train it with their own data. In the future, Cube 3D is expected to support image inputs and process multimodal data, including text, images, and videos.	By Anupam Singh, Vice President of Engineering, and Nick Tornow, Vice President of Creator Engineering		March 17, 2025
1.2	Cornstarch	Cornstarch is a distributed multimodal training framework designed to optimize the training of models that process multiple data types, such as text, images, and	By Insu Jang et al.		March 17, 2025



I. MODELS					
#	Highlights	Summary	Author	Source	Date
		audio. It addresses inefficiencies in existing systems by being "multimodality-aware," ensuring balanced resource allocation and synchronization across modalities. This approach minimizes bottlenecks and maximizes throughput during training, making it particularly effective for large-scale multimodal tasks like vision-language models. By intelligently managing data flow and computation, Cornstarch enhances scalability and efficiency, enabling faster and more cost-effective training of complex multimodal systems			
1.3	NVIDIA Announces Isaac GROOT N1: The World's First Open Humanoid Robot Foundation Model	NVIDIA has introduced the open-source Isaac GROOT N1 model to accelerate humanoid robot development. Inspired by human cognition, the model features a dual-system architecture, where "System 1" manages fast and intuitive actions, while "System 2" handles deliberate decision-making. GROOT N1 can learn to grasp, transport objects, and perform multi-step tasks. The model can be customized with real or synthetic data, offering flexible applications. NVIDIA is collaborating with Google DeepMind and Disney Research to develop Newton, an open-source physics engine based on the NVIDIA Warp framework, to enhance robotic simulations.	By Kristin Bryson		March 18, 2025
1.4	Cosmos World Foundation Models	The Cosmos AI model series has been developed to generate cost-effective and highly realistic videos for robot training, enabling more efficient and scalable AI development. These models operate within NVIDIA's Omniverse platform, which the company describes as "the operating system for physical AI." By leveraging Omniverse, developers can condition Cosmos to create systematically controlled	By Nvidia		March 18, 2025



I. MODELS					
#	Highlights	Summary	Author	Source	Date
		yet infinitely diverse "infinite environments", making it an ideal tool for training robotic systems in a wide range of scenarios. This approach allows AI models to learn and adapt in virtual spaces before being deployed in real-world applications, improving both accuracy and efficiency.			
1.5	LG AI Research Unveils 'EXAONE Deep' Reasoning AI Model	LG AI Research introduced ' EXAONE Deep ', an advanced reasoning AI model marking a significant milestone in AI innovation. EXAONE Deep-32B, with 32 billion parameters, demonstrated exceptional efficiency, achieving remarkable performance at just 5% of a very large competitor model's size. It excelled in solving complex mathematical and scientific problems, recording the highest score of 94.5 in the 2025 CSAT mathematics area and 95.7 points in MATH-500. Additionally, it achieved top grades in physics, chemistry, and biology evaluations. This open-source release underscores Korea's competitiveness in the global AI market.	By PR Newswire		March 18, 2025
1.6	Razer Introduces Wyvrn Developer Platform with AI Tools	Razer unveiled Wyvrn, a new developer platform featuring automated AI tools aimed at enhancing game development processes. The platform includes the AI QA Copilot, a cloud-based plugin compatible with engines like Unreal and Unity, designed to streamline quality assurance testing by identifying bugs and generating detailed reports. Razer claims this tool can enhance bug detection by up to 25% and reduce QA time by 50%, potentially impacting QA team sizes. Additionally, Wyvrn offers the AI Gamer Copilot, an AI voice assistant providing real-time gaming tips and strategies, marking a significant advancement in integrating AI into gaming experiences.	By Antonio G. Di Benedetto		March 19, 2025


I. MODELS					
#	Highlights	Summary	Author	Source	Date
1.7	Llama-3.3-Nemotron-Super-49B-v1	Llama-3.3-Nemotron-Super-49B-v1 is a reasoning-optimized LLM derived from Meta Llama-3.3-70B-Instruct. It is post-trained for reasoning, chat preferences, retrieval-augmented generation (RAG), and tool use, supporting a 128K token context. Designed for efficiency and accuracy, it balances high performance with cost savings. A Neural Architecture Search (NAS) approach reduces memory usage, enabling deployment on a single H200 GPU with high throughput. This optimization allows larger workloads while maintaining strong reasoning capabilities. The NAS method fine-tunes the accuracy-efficiency tradeoff, making it a cost-effective, high-performance alternative to larger models, ideal for various AI-driven applications.	By Nvidia		March 18, 2025
1.8	RF-DETR	RF-DETR is a state-of-the-art object detection model that enhances the DETection Transformer (DETR) architecture with advanced optimizations for better performance and efficiency. It tackles challenges like overlapping objects and complex scene accuracy through refined feature extraction and improved training techniques. Ideal for precise object localization and classification, RF-DETR excels in applications such as autonomous driving, surveillance, and robotics. By integrating cutting-edge transformer-based advancements, it provides a robust, scalable solution for modern object detection tasks, balancing accuracy and efficiency effectively.	By Peter Robicieux et al.		March 20, 2025
1.9	SpatialLM-Llama-1B: Advancing 3D Scene	Manycore Research has introduced SpatialLM-Llama-1B , a 3D large language model designed to process 3D point cloud data and generate structured 3D scene understanding outputs, including architectural elements like walls, doors, windows, and oriented object bounding boxes with their semantic categories.	By Manycore Research		March 20, 2025


I. MODELS					
#	Highlights	Summary	Author	Source	Date
	Understanding with Language Models	Unlike previous methods that require specialized equipment for data collection, SpatialLM can handle point clouds from diverse sources such as monocular video sequences, RGBD images, and LiDAR sensors. This multimodal architecture bridges the gap between unstructured 3D geometric data and structured 3D representations, enhancing spatial reasoning capabilities for applications in embodied robotics, autonomous navigation, and complex 3D scene analysis tasks.			
1.9	Orpheus TTS: Advancements in Text-to-Speech Synthesis	Canopy Labs has released Orpheus TTS , a state-of-the-art text-to-speech model based on Llama architecture, fine-tuned to deliver human-like speech synthesis with natural intonation and emotion. Key features include zero-shot voice cloning, guided emotion and intonation control through simple tags, and low-latency streaming suitable for real-time applications. The model is available in 3 billion parameter configurations, with quantized versions for efficient deployment. Developers can access the model and its codebase via Hugging Face and GitHub	By CanopyLabs		March 18,2025
1.10	Aardvark Weather	A new AI-driven forecasting model developed collaboratively by the University of Cambridge, the Alan Turing Institute, Microsoft Research, and the European Centre for Medium-Range Weather Forecasts, represents a significant advancement in meteorology. Unlike previous AI models that rely on traditional numerical weather prediction systems, Aardvark utilizes a purely machine-learning approach, processing observations from satellites, weather stations, and other sensors to produce high-resolution global and local forecasts. Notably, it operates efficiently on desktop computers, delivering results within minutes and outperforming the U.S. Global Forecast System in certain variables. This	By Ben Geman, Andrew Freedman		March 24, 2025


I. MODELS					
#	Highlights	Summary	Author	Source	Date
		innovation holds promise for tailored applications in sectors like renewable energy and agriculture.			
1.11	First Mamba-Powered Ultra-Large Model	The Hunyuan-T1 model, built on the TurboS Hybrid-Transformer-Mamba MoE architecture, represents a significant advancement in reasoning capabilities. With optimized long-text processing and efficient computation, it achieves double the decoding speed of its predecessor while excelling in mathematics, logic, and scientific reasoning. Trained with 96.7% reinforcement learning focus and curriculum learning strategies, T1 demonstrates superior alignment with human preferences. It achieves top-tier performance across benchmarks like MMLU-Pro (87.2) and MATH-500 (96.2), rivaling industry-leading models. T1's innovative architecture and training make it a cutting-edge solution for reasoning and alignment tasks.	By Tencent Hunyuan Research		March 21, 2025
1.12	AMD's Gaia Project	AMD launched Gaia, an open-source initiative aimed at enabling efficient local execution of large language models (LLMs) on standard PCs. Gaia optimizes hardware resources, particularly focusing on AMD's Ryzen CPUs and Radeon GPUs, enhancing performance even on less powerful consumer systems. This approach allows users to run sophisticated AI workloads without cloud dependencies, significantly improving privacy and accessibility. AMD's Gaia directly challenges Nvidia's dominance by providing viable alternatives for decentralized AI processing, fostering greater hardware democratization in the rapidly expanding AI computing landscape.	By Aaron Klotz		March 21,2025
1.13	DeepSeek V3 Update	DeepSeek unveiled version 3 (V3) of its AI model, featuring significant improvements in performance and efficiency alongside a new licensing structure.	By Ambar Warrick		March 25,2025



I. MODELS					
#	Highlights	Summary	Author	Source	Date
		V3 emphasizes enhanced multilingual capabilities, reasoning accuracy, and computational efficiency. Importantly, the updated license now allows commercial use with clearer terms, aiming to attract a broader range of developers and enterprises. This strategic shift positions DeepSeek to better compete in the commercial AI market, appealing especially to businesses seeking robust AI tools with transparent usage rights. Such licensing adjustments are increasingly critical in the competitive AI ecosystem dominated by both closed and open-source alternatives.			
1.14	Beijing-Backed AI Startup Manus Aims to Be China's Next DeepSeek	China is backing AI startup Manus, seen as a potential rival to domestic leaders like DeepSeek and Baidu. Supported by Beijing's strategic funding and talent pipelines, Manus is focused on developing cutting-edge large language models tailored to Chinese linguistic and regulatory needs. The move reflects China's ambition to cultivate a diverse and competitive AI ecosystem, reducing reliance on U.S. technologies. Manus' emergence is part of a broader national effort to scale homegrown AI innovation and secure leadership in foundational model development.	By Josh Ye		March 21, 2025
1.15	Qwen2.5-VL-32B: Smarter and Lighter	Alibaba's Qwen team has released Qwen-VL-Max , a new 32B vision-language model with enhanced capabilities in document understanding, complex reasoning, chart interpretation , and OCR . It supports multi-image and multi-round dialogues, surpassing Gemini Pro 1.5 and GPT-4V in multiple benchmarks like MathVista and AI2D. Qwen-VL-Max also introduces a lightweight API that simplifies region selection, coordinates extraction, and chart data reading. With strong performance on benchmarks such as ChartQA, DocVQA, and MMMU, it showcases its superiority across vision-language tasks. The model, weights, and	By Qwen Team		March 24, 2025




I. MODELS					
#	Highlights	Summary	Author	Source	Date
		demo are now open source , promoting transparency and broader accessibility for researchers and developers.			
1.16	Cosmos-Reason1: From Physical Common Sense to Embodied Reasoning	Cosmos-Reason1 explores how AI can develop embodied reasoning by integrating physical common sense with multimodal learning. The framework enhances an agent’s ability to predict and reason about physical interactions in the world, leveraging text, vision, and motion cues. The model is trained on large-scale datasets with physically grounded scenarios, improving zero-shot and few-shot reasoning capabilities. Experimental results show advancements in tasks like object dynamics prediction, intuitive physics, and embodied decision-making. This work contributes to bridging the gap between common sense understanding and real-world physical reasoning in AI systems.	By Nvidia		March 18, 2025
1.17	Nemotron-H: A Family of Accurate, Efficient Hybrid Mamba-Transformer Models	The Nemotron-H family introduces advanced hybrid Mamba-Transformer models, offering state-of-the-art accuracy with up to 3x faster inference compared to similar-sized Transformer models. Spanning 8B to 56B parameters, these models leverage FP8 precision for efficient training and support long-context inference, achieving breakthroughs in reasoning, math, and vision-language tasks. The 56B model, trained on 20 trillion tokens, demonstrates competitive performance against much larger models, while the 47B variant enables efficient deployment on commodity GPUs. Nemotron-H represents a significant leap in LLM scalability, accuracy, and efficiency, with applications in AI reasoning, code generation, and physical AI systems.	By NVIDIA ADLR		March 21, 2025



I. MODELS					
#	Highlights	Summary	Author	Source	Date
1.18	Cosmos-Transfer1: Conditional World Generation with Adaptive Multimodal Control	Cosmos-Transfer1 proposes a framework for generating conditional world representations using multimodal data, enabling adaptive control over generated environments. The model integrates textual, visual, and spatial inputs to create coherent world representations suitable for simulations and robotics. The authors introduce a novel transfer learning strategy that adapts knowledge from diverse datasets, improving generalization across tasks. The results demonstrate enhanced fidelity and consistency in generated environments. This approach is particularly relevant for AI-driven simulations, virtual reality, and autonomous system training, pushing the boundaries of controllable and context-aware world generation.	By Nvidia	https://arxiv.org/pdf/2503.14492	March 18, 2025
1.19	Tokenize Image as a Set	TokenSet, a groundbreaking paradigm for image generation that tokenizes images into unordered sets instead of fixed-position sequences, enabling dynamic coding allocation based on regional semantic complexity. To model these sets, the authors propose a dual transformation mechanism that converts sets into fixed-length integer sequences while retaining summation constraints. They further develop Fixed-Sum Discrete Diffusion, a novel framework tailored for discrete data modeling with fixed-sum properties. Experiments demonstrate TokenSet's superior robustness, semantic-aware representation, and generation quality, marking a significant advancement over traditional sequential tokenization methods in visual generation tasks.	By University of Science and Technology of China, Tencent Hunyuan Research		March 20, 2025



I. MODELS					
#	Highlights	Summary	Author	Source	Date
1.20	Lyra: An Efficient and Expressive Subquadratic Architecture for Modeling Biological Sequences	Lyra, a novel subquadratic architecture for biological sequence modeling that integrates Projected Gated Convolutions (PGCs) and diagonalized State Space Models (S4D). Designed to efficiently model epistasis (context-dependent interactions), Lyra achieves state-of-the-art (SOTA) performance across over 100 biological tasks, including protein fitness prediction, RNA function analysis, and CRISPR guide design. With up to 120,000-fold fewer parameters and 64x faster inference than Transformers, Lyra democratizes access to advanced sequence modeling. Its efficiency and scalability make it transformative for computational biology, enabling breakthroughs in therapeutic design, diagnostics, and gene editing.	By Krithik Ramesh, Sameed M. Siddiqui, Albert Gu, Michael D. Mitzenmacher, Pardis C. Sabeti		March 20, 2025




II. AI CHIPS					
#	Highlights	Summary	Author	Source	Date
2.1	NVIDIA Blackwell Ultra	NVIDIA has introduced the Blackwell Ultra platform, an advanced evolution of its Blackwell GPU architecture, tailored for AI applications. A major upgrade includes	By Nvidia		March 18, 2025



II. AI CHIPS					
#	Highlights	Summary	Author	Source	Date
		expanding High Bandwidth Memory (HBM) from 192GB to 288GB in the Blackwell Ultra GB300, coupled with a 50% performance increase over the B200 series. The GB300 AI chip is available in dual-chip (CoWoS-L) and single-chip (CoWoS-S) configurations. Trial production is set for Q2 2025, with mass production in Q3. These advancements highlight NVIDIA's commitment to scaling AI hardware performance to support increasingly complex AI workloads.			
2.2	NVIDIA Rubin Ultra GPU and NVIDIA Vera CPU	NVIDIA CEO Jensen Huang honored astronomer Vera Rubin by unveiling a roadmap for advancing data center performance. The plan features next-generation NVIDIA Rubin Ultra GPU and NVIDIA Vera CPU architectures, introducing major innovations. Huang stated, "basically everything is brand new except for the chassis." Systems built on the Rubin Ultra platform, including the Vera Rubin NVL 144, will launch in late 2026, with additional Rubin Ultra-based systems arriving in 2027. Huang emphasized that Rubin will significantly reduce costs, reinforcing NVIDIA's push for efficient AI computing.	By Nvidia		March 18, 2025
2.3	RTX PRO 6000 Blackwell	NVIDIA has unveiled the RTX PRO 6000 Blackwell series, designed for professional users, offering up to 96GB of GDDR7 memory and fifth-generation PCIe support. These GPUs are engineered to enhance performance in compute-intensive tasks, such as AI inference, data analytics, and content creation. Jensen Huang highlighted the series' innovative capabilities, emphasizing its role in advancing both AI applications and professional graphics processing. With cutting-edge architecture, the RTX PRO 6000 Blackwell delivers exceptional efficiency and power, making it an ideal choice for demanding workloads across industries like scientific computing, media production, and AI-driven research.	By Nvidia		March 18, 2025



II. AI CHIPS					
#	Highlights	Summary	Author	Source	Date
2.4	Micron Stock Rises as AI Memory Boosts Revenue	Micron Technology saw a strong financial performance in Q1 2025, with revenue rising 38% year-over-year to \$8.05 billion, driven by high demand for its high-bandwidth memory (HBM) chips used in AI servers. The company's net income nearly doubled compared to the previous year. CEO Sanjay Mehrotra highlighted the robust growth in data center revenues and projected sustained demand for AI memory through at least 2026. Micron expects another record quarter ahead, further boosting investor confidence. As a result, the company's stock gained 2% in after-hours trading, signaling optimism around its positioning in the rapidly growing AI hardware market.	By Mike Wheatley		March 20, 2025
2.5	Lisa Su Sets Sights on Nvidia as AMD Expands AI Chip Ambitions	Since becoming CEO of AMD in 2014, Lisa Su has led a major transformation of the company, enabling it to surpass Intel. Now, she faces a new challenge: competing with Nvidia, which currently holds about 90% of the AI chip market. Su aims to strengthen AMD's software capabilities to rival Nvidia's dominant CUDA platform. The company is investing in open-source software to support training and inference of large language models. Her leadership style—focused on consistency and precise execution—has earned respect from both partners and critics. This strategy is designed to position AMD more competitively in the AI space.	By Kif Leswing		March 20, 2025
2.6	AMD signs huge multi-billion dollar deal with Oracle to build a cluster of 30,000	AMD has secured a multi-billion dollar deal with Oracle to build a massive AI supercomputer cluster powered by 30,000 MI355X accelerators. These next-generation GPUs, based on AMD's CDNA 4 architecture and TSMC's advanced 3nm process, are set to rival Nvidia's Blackwell chips. The partnership underscores	By Wayne Williams		March 21, 2025




II. AI CHIPS					
#	Highlights	Summary	Author	Source	Date
	MI355X AI accelerators	Oracle’s ambition to expand its cloud AI infrastructure and support growing enterprise demand for generative AI workloads. With a projected 35-fold increase in AI server GPU demand by 2027, this strategic move strengthens AMD’s position in the AI hardware market and marks a significant leap forward in large-scale AI compute capabilities.			
2.7	European Lawmakers Advocate for Chips Act 2.0	European lawmakers are pressing the European Commission to rapidly advance the proposed Chips Act 2.0, designed to strengthen Europe's semiconductor independence amid global supply chain concerns. Driven by intensifying geopolitical tensions and technological competition, the legislation emphasizes increased investment, advanced chip manufacturing, and research innovation. Lawmakers warn delays could jeopardize Europe's competitiveness in AI and emerging technologies. With semiconductor demand surging, the Chips Act 2.0 seeks to bolster Europe's strategic autonomy and reduce reliance on non-European suppliers, ensuring stability and resilience in crucial tech sectors and enhancing regional economic security.	By Foo Yun Chee		March 24,2025
2.8	Micron Shares Dip Despite Strong AI Demand Due to Weak Margin Outlook	Micron Technology's shares fell after the company issued a weaker-than-expected margin forecast, overshadowing optimism about AI-driven demand for its memory chips. While Micron highlighted robust growth in high-bandwidth memory (HBM) used in AI workloads, investors were concerned about profitability pressures and elevated production costs. The company remains a key supplier for Nvidia and other AI players, but its forecast sparked worries about near-term financial performance. The stock decline reflects market sensitivity to both the opportunities and operational challenges facing AI-related semiconductor firms.	By Max A. Cherney,		March 21,2025



II. AI CHIPS					
#	Highlights	Summary	Author	Source	Date
2.9	Nvidia-Backed CoreWeave Targets Up to \$27 Billion Valuation in U.S. IPO	CoreWeave, a cloud computing startup backed by Nvidia, is planning a U.S. IPO that could value the company at up to \$27 billion. Specializing in GPU-accelerated infrastructure for AI workloads, CoreWeave has rapidly grown as demand for high-performance computing surges. The company provides scalable infrastructure for clients training large AI models, positioning itself as a key player in the AI chip ecosystem. This IPO underscores investor appetite for firms supporting the AI boom, particularly those aligned with industry leaders like Nvidia.	By Manya Saini, Niket Nishant		March 20, 2025
2.10	NVBleed: Covert and Side-Channel Attacks on NVIDIA Multi-GPU Interconnect	NVBleed exposes covert and side-channel vulnerabilities in NVIDIA's multi-GPU NVLink interconnect. Researchers demonstrate how malicious actors can exploit shared resources to leak sensitive data (e.g., encryption keys) across GPUs without explicit communication. The attack leverages contention in memory access patterns or power consumption to infer activity from a victim GPU, bypassing isolation guarantees. These flaws arise from hardware design choices prioritizing performance over security. Mitigations require architectural changes, as software patches may degrade performance. The work highlights risk in multi-GPU environments, urging a reevaluation of interconnect security for cloud and HPC systems relying on NVIDIA's technology.	Bt Yicheng Zhang et al.		March 22, 2025



III. LLM TECHNIQUES AND METRICS					
#	Highlights	Summary	Author	Source	Date
3.1	reWordBench: Evaluating and Enhancing Reward Model Robustness	reWordBench investigates how even small input transformations affect the performance of reward models, which play a crucial role in text evaluation and alignment processes. However, due to overfitting, these models may not fully reflect their true capabilities. The authors developed reWordBench, a tool that tests models using transformations that preserve meaning or ranking. Results indicate that state-of-the-art models are fragile. As a solution, they propose a training strategy that assigns similar scores to paraphrases, significantly improving robustness and achieving up to a 59% performance gain in alignment tasks.	By Zhaofeng Wu, Michihiro Yasunag, Andrew Cohen, Yoon Kim, Asli Celikyilmaz, Marjan Ghazvininejad		March 18, 2025
3.2	Creation-MMBench: Assessing Context-Aware Creative Intelligence in MLLM	Creation-MMBench is a specialized benchmark designed to evaluate the creative capabilities of Multimodal Large Language Models (MLLMs) in real-world, image-based tasks. It comprises 765 test cases across 51 fine-grained tasks, each with instance-specific evaluation criteria to assess both the general quality of responses and their factual consistency with visual inputs. Experimental results indicate that current open-source MLLMs significantly underperform compared to proprietary models in creative tasks. Additionally, visual fine-tuning has been observed to negatively impact the base LLM's creative abilities. Creation-MMBench offers valuable insights for advancing MLLM creativity and establishes a foundation for future improvements in multimodal generative intelligence.	By Xinyu Fang, Zhijian Chen, Kai Lan, Shengyuan Ding, Yingji Liang, Xiangyu Zhao, Farong Wen, Zicheng Zhang, Guofeng Zhang, Haodong Duan, Kai Chen, Dahua Lin		March 18, 2025
3.3	R1-VL: Step-wise Group Policy Optimization for Multimodal LLMs	R1-VL: Learning to Reason with Multimodal Large Language Models via Step-wise Group Relative Policy Optimization" aims to enhance the reasoning capabilities of Multimodal Large Language Models (MLLMs). Existing methods typically improve MLLMs' reasoning abilities by fine-tuning them on high-quality reasoning data	By Jingyi Zhang, Jiaying Huang, Huanjin Yao, Shunyu Liu, Xikun		March 17, 2025



III. LLM TECHNIQUES AND METRICS					
#	Highlights	Summary	Author	Source	Date
		generated by powerful models. However, these approaches may lead models to merely imitate successful reasoning paths, preventing them from understanding incorrect reasoning paths. To address this issue, the authors propose a novel online reinforcement learning framework called Step-wise Group Relative Policy Optimization (StepGRPO), which enables MLLMs to enhance their reasoning capabilities autonomously.	Zhang, Shijian Lu, Dacheng Tao		
3.4	Synopsys Introduces AgentEngineer for AI-Assisted Chip Design	Synopsys has introduced AgentEngineer , a groundbreaking technology that leverages AI agents to assist engineers in complex chip design tasks , such as testing circuit designs. As companies like Nvidia transition to designing AI server systems with thousands of interconnected chips, the complexity and speed of the design process have increased significantly. AgentEngineer enhances research and development capabilities without the need to expand engineering teams. It coordinates intricate system designs, ensuring timely product delivery , marking a significant step in the evolution of AI-assisted engineering for chip development.	By Stephen Nellis		March 19, 2025
3.5	ELTEX: A Framework for Domain-Driven Synthetic Data Generation	ELTEX (Efficient LLM Token Extraction) is a domain-driven framework for generating high-quality synthetic training data in specialized fields like cybersecurity. LLMs struggle in such domains due to limited domain-specific data. ELTEX tackles this by integrating domain indicator extraction and dynamic prompting to retain critical knowledge. Tested on blockchain cyberattack detection, ELTEX fine-tuned Gemma-2B, achieving GPT-4-level performance in classification and uncertainty calibration with fewer resources. A synthetic dataset of social media texts for cyberattack detection is released. This study	By Arina Razmyslovich, Kseniia Murasheva, Sofia Sedlova, Julien Capitaine, Eugene Dmitriev		March 19, 2025




III. LLM TECHNIQUES AND METRICS					
#	Highlights	Summary	Author	Source	Date
		shows that domain-driven synthetic data can help smaller models compete with larger ones.			
3.6	SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks	This paper, SWEET-RL is a reinforcement learning method designed to improve multi-turn collaboration between LLM agents and humans. Introduced alongside ColBench, a benchmark for backend programming and frontend design tasks, SWEET-RL uses a critic model trained with extra training-time data to deliver step-wise rewards. This enhances the agent’s decision-making across turns. Experiments show a 6% absolute gain in success and win rates over prior RL methods. Notably, SWEET-RL enables Llama-3.1-8B to perform on par with or better than GPT-4o in collaborative reasoning and content creation tasks.	By Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, Xian Li		March 19, 2025
3.7	Plug-and-Play 1.x-Bit KV Cache Quantization for Video Large Language Models	VideoLLMs can handle long video inputs but face memory and speed bottlenecks due to large KV caches from thousands of visual tokens. While 2-bit KV cache quantization has minimal impact on performance, lower-bit quantization remains underexplored. This paper introduces VidKV , a plug-and-play method compressing KV cache below 2 bits. It uses mixed-precision key quantization (2-bit for anomalous channels, 1-bit with FFT for others) and 1.58-bit value quantization with selective token filtering. Experiments on LLaVA-OV-7B and Qwen2.5-VL-7B show near-FP16 performance with 1.5-bit compression, outperforming prior per-token quantization by using per-channel strategies.	By Keda Tao, Haoxuan You, Yang Sui, Can Qin, Huan Wang		March 21, 2025




III. LLM TECHNIQUES AND METRICS					
#	Highlights	Summary	Author	Source	Date
3.8	When Less is Enough: Adaptive Token Reduction for Efficient Image Representation	Vision encoders generate many visual tokens, offering rich representations but increasing computational costs. This paper introduces a method to assess and select only the most informative tokens, based on the idea that less useful features can be reconstructed from more important ones. Using an autoencoder with a Gumbel-Softmax mechanism, the approach retains key tokens while discarding redundant ones. Applied to the LLaVA-NeXT model, the method shows that over 50% of visual context can be removed in OCR tasks with minimal performance loss. It enables efficient, adaptive multimodal pruning for scalable inference without compromising quality.	By Eduard Allakhverdov, Elizaveta Goncharova, Andrey Kuznetsov		March 20, 2025
3.9	Modifying Large Language Model Post-Training for Diverse Creative Writing	This paper introduces a method to enhance both output diversity and quality in creative writing tasks using large language models (LLMs). The approach incorporates a "deviation"-based objective during training, prioritizing rare and high-quality samples by measuring how different each example is from others. Applied to Direct Preference Optimization (DPO) and Odds Ratio Preference Optimization (ORPO), the method enables an 8B-parameter model to generate outputs as diverse as human-curated datasets and match the quality of top models like GPT-4o and DeepSeek-R1. The approach is validated through human evaluations, ablation studies, and comparisons with the existing diversification method DivPO.	By John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, Max Kreminski		March 21, 2025
3.10	Judge Anything: MLLM as a Judge Across Any Modality	Multimodal Large Language Models (MLLMs) are gaining traction for their ability to process visual and textual data, but evaluating their judgment remains difficult due to limited benchmarks aligned with human preferences. The "MLLM-as-a-Judge" benchmark addresses this with tasks like Scoring Evaluation, Pair Comparison, and Batch Ranking. Results show MLLMs perform well in Pair	By Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao		March 21, 2025



III. LLM TECHNIQUES AND METRICS					
#	Highlights	Summary	Author	Source	Date
		Comparison but struggle with scoring and ranking, revealing issues like bias, hallucinations, and inconsistency—even in top models like GPT-4V. To improve reliability, the study introduces OmniArena, an automated platform designed to support better evaluation of multimodal models and reward systems through more rigorous, scalable testing.	Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, Yi Gui, Yao Wan, Philip S. Yu		
3.11	Typed-RAG: Type-aware Multi-Aspect Decomposition for Non-Factoid Question Answering	Non-factoid question-answering (NFQA) is challenging due to its open-ended nature, varied intents, and need for multi-aspect reasoning, making traditional factoid approaches like RAG insufficient. To address this, Typed-RAG is introduced—a type-aware framework that enhances RAG by classifying NFQs into categories such as debate, experience, and comparison. It decomposes complex questions into single-aspect sub-queries, improving both retrieval and response quality. By aggregating these focused results, Typed-RAG delivers more informative and relevant answers. A new benchmark, Wiki-NFQA, was developed to evaluate performance, showing that Typed-RAG outperforms existing baselines through its structured, type-driven approach to NFQA.	By DongGeon Lee, Ahjeong Park, Hyeri Lee, Hyeonseo Nam, Yunho Maeng		March 21, 2025
3.12	Modifying Large Language Model Post-Training for Diverse Creative Writing	Creative writing tasks require large language models (LLMs) to produce diverse, high-quality responses. However, post-training methods often prioritize quality while neglecting diversity. Researchers introduce a novel approach that incorporates “deviation”—how different a sample is from others with the same prompt—into the training objective. Applied to Direct Preference Optimization (DPO) and Odds Ratio Preference Optimization (ORPO), this method improves output diversity with minimal quality loss. An 8B-parameter model achieved	By John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqia		March 21, 2025

III. LLM TECHNIQUES AND METRICS					
#	Highlights	Summary	Author	Source	Date
		human-level diversity and matched the quality of top instruction-tuned models like GPT-4o and DeepSeek-R1. Human evaluations and comparisons to DivPO further validate the effectiveness of the approach.	n Sun, Max Kreminski		
3.13	SuperBPE: Space Travel for Language Models	SuperBPE, an innovative tokenization algorithm that extends the traditional Byte-Pair Encoding (BPE) by incorporating "superword" tokens that bridge multiple words. SuperBPE improves encoding efficiency by up to 33% compared to BPE, enabling shorter token sequences and reducing inference compute by 27%. Pretraining experiments on 8B-scale language models show that SuperBPE achieves a +4.0% average improvement across 30 downstream tasks, including a +8.2% gain on MMLU. This straightforward modification enhances both performance and efficiency without altering model architectures, making it a transformative advancement in tokenization for language modeling.	By NVIDIA, Allen Institute for AI		March 17, 2025
3.14	TULIP: Towards Unified Language-Image Pretraining	TULIP (Towards Unified Language-Image Pretraining) is a novel framework designed to overcome the limitations of existing contrastive image-text models like CLIP and SigLIP, which struggle with fine-grained visual understanding and spatial reasoning. By integrating generative data augmentation, patch-level contrastive learning, and reconstruction-based feature regularization, TULIP achieves a balance between fine-grained visual detail and global semantic alignment. Scaling to over 1B parameters, TULIP establishes state-of-the-art performance across benchmarks, excelling in zero-shot classification, vision-language tasks, and fine-grained recognition, marking a significant advancement in multimodal representation learning. Its open-source availability further amplifies its impact.	By University of California		March 19, 2025

IV. AI USE CASES					
#	Highlights	Summary	Author	Source	Date
4.1	NetApp Announce Agentic AI Reasoning Solutions with NVIDIA	NetApp has partnered with NVIDIA to launch a new AI infrastructure designed for agentic AI solutions. This system integrates NetApp's ONTAP storage operating system with NVIDIA's accelerated computing, networking, and AI software to enhance enterprise data management and AI workloads. Built on the NVIDIA AI Data Platform reference architecture, it optimizes large-scale data processing, ensuring AI applications run efficiently, scalably, and securely. By accelerating AI-driven decision-making and automation, businesses can better leverage data-driven AI models. This innovation enables enterprises to process vast datasets, improve AI inference speeds, and unlock new possibilities in real-time analytics, automation, and competitive intelligence.			March 18, 2025
4.2	Innodata Unveils Generative AI Test and Evaluation Platform, Built with NVIDIA Technology	Innodata has introduced the beta version of the Generative AI Test & Evaluation Platform to enhance the security, reliability, and performance of AI models. Powered by NVIDIA's advanced inference technology, this platform conducts automated attack tests on AI models to identify potential security vulnerabilities and performs model comparisons. Early adopters, such as MasterClass, plan to use the platform to ensure their AI investments deliver reliable and secure customer experiences. The full version of the platform is expected to be released in the second quarter of 2025.	By Innodata		March 18, 2025


IV. AI USE CASES					
#	Highlights	Summary	Author	Source	Date
4.3	Tencent expands AI push with open-source 3D generation tools	Tencent has announced the release of a suite of new AI tools that convert text and images into 3D visuals, enhancing China's advancements in generative AI. The company introduced five open-source models leveraging its Hunyuan3D-2.0 technology, with "turbo" versions capable of generating high-quality 3D visuals in just 30 seconds. This launch signifies Chinese firms' increasing competition with the U.S. by offering high-performing AI models at reduced costs. Tencent's initiative builds on its first introduction of 3D AI models in November 2024, focusing on designers and game developers.	By Reuters		March 18, 2025
4.4	AI-Powered Robot Developed to Autonomously Make Coffee	Researchers from the University of Edinburgh have developed an AI-powered robot capable of autonomously making coffee. Equipped with a seven-jointed robotic arm, the robot integrates AI, sensors, and motor skills to navigate dynamic environments like kitchens. It can interpret verbal instructions, locate and retrieve a mug, and mix precise ratios of coffee and water. Unlike traditional robots reliant on pre-programmed actions, this robot adapts to unforeseen events, such as unexpected object movements. This breakthrough signifies a leap in integrating reasoning, movement, and perception in robotics, potentially enhancing automation in daily tasks.	By Elizabeth Hunter		March 19, 2025
4.5	Disney Collaborates with Nvidia and Google DeepMind on Advanced Robotics	Disney is partnering with Nvidia and Google DeepMind to develop a new physics engine, Newton , to enhance its advanced robotic characters. At Nvidia's GTC AI Conference, CEO Jensen Huang introduced an endearing robot named Blue , created by Disney Research. Newton, built on Nvidia's Warp framework, will aid robots in completing complex tasks more efficiently. Disney plans to deploy these upgraded, lifelike robots in its parks, including Walt Disney World, Tokyo	By Maxwell Zeff		March 18, 2025



IV. AI USE CASES					
#	Highlights	Summary	Author	Source	Date
		Disneyland, and Disneyland Paris, marking a significant leap in the development of expressive and functional robots for entertainment.			
4.6	Digital Physiotherapy for Back Pain Management	With around one in six people globally suffering from chronic back pain, access to physiotherapy remains limited—especially in rural or underserved areas. AI-powered digital physiotherapists are emerging as a scalable solution. These systems guide patients through tailored exercise routines, monitor movements using computer vision, and provide real-time feedback. Such tools not only extend access to care but also reduce the burden on healthcare systems. While they don't replace human therapists entirely, AI-driven platforms are proving effective in supporting rehabilitation and encouraging consistent, at-home treatment for back pain patients.	By Scott Nover		March 24, 2025
4.7	Coffee-making robot breaks new ground for AI machines	Researchers at the University of Edinburgh have developed a robotic arm that combines advanced motor skills with AI, enabling it to perform complex tasks in unpredictable settings like kitchens. Unlike traditional robots limited to pre-programmed actions, this robot interprets verbal instructions, analyzes its surroundings, and adapts in real time. It can open unfamiliar drawers, locate a mug, and prepare coffee—even adjusting if objects are moved mid-task. Published in Nature Machine Intelligence, the study highlights the integration of reasoning, perception, and movement as key to building robots capable of operating alongside humans.	By University Of Edinburgh		March 19, 2025
4.8	NatWest Partners with OpenAI in	NatWest has entered a milestone partnership with OpenAI, becoming the first UK bank to formally integrate the company's generative AI tools into its operations. The collaboration will support customer service, automate internal workflows,	By Iain Withers,		March 20, 2025




IV. AI USE CASES					
#	Highlights	Summary	Author	Source	Date
	Landmark UK Banking Collaboration	and assist staff with complex decision-making tasks. NatWest aims to leverage OpenAI's capabilities to enhance productivity and client engagement while ensuring responsible AI governance. This deal marks a significant step in AI adoption within the financial sector and sets a precedent for broader industry use of generative models.			
4.9	Micron forecasts upbeat quarterly revenue on strong AI memory chip demand	Micron Technology has issued an upbeat revenue forecast for the upcoming quarter, driven by soaring demand for high-bandwidth memory (HBM) chips used in AI applications. These advanced memory chips are essential for powering large language models and AI data centers. Micron is ramping up production to compete with industry leaders like Samsung and supply key players such as NVIDIA. The company's bullish outlook signals the growing strategic importance of AI hardware, as demand accelerates across cloud, enterprise, and AI sectors. Investors responded positively, pushing Micron shares higher following the announcement.	By Juby Babu		March 21, 2025
4.10	Training Video Foundation Models with NVIDIA NeMo	This paper introduces NVIDIA NeMo's framework for training large-scale video foundation models. It highlights a modular pipeline optimized for multimodal learning, leveraging transformer-based architectures and large-scale datasets. The authors discuss pretraining techniques, dataset curation, and efficiency improvements, including memory optimization and distributed training strategies. They also showcase benchmark results, demonstrating state-of-the-art performance in video understanding tasks. The paper emphasizes NeMo's	By Nvidia		March 17, 2025



IV. AI USE CASES					
#	Highlights	Summary	Author	Source	Date
		adaptability for various applications, such as video captioning and action recognition, providing insights into the future of scalable video model training.			



V. AI POLICIES, REGULATIONS & STRATEGIES					
#	Highlights	Summary	Author	Source	Date
5.1	Europol Warns of AI-Driven Crime Threats	Europol has issued a warning about the growing threats posed by AI-driven crime, highlighting how organized crime gangs are leveraging advanced technologies to enhance their global operations. The agency's European Serious Organised Crime Threat Assessment report notes that AI allows these criminals to craft multilingual messages, create realistic impersonations, and automate processes, complicating detection efforts. The report also highlights the potential for fully autonomous AI-controlled criminal networks in the future, underscoring the need for robust regulatory frameworks and international cooperation to combat these evolving threats.	By Reuters		March 18,2025
5.2	OpenAI Proposes 'Freedom-Focused' AI Policy to White House	OpenAI has submitted a "freedom-focused" policy proposal to the White House's AI Action Plan, advocating measures to maintain U.S. AI dominance over China. Key recommendations include leveraging trade laws, reducing copyright restrictions for training data, investing in AI infrastructure, and preventing states	By Tina Nguyen		March 19,2025



V. AI POLICIES, REGULATIONS & STRATEGIES					
#	Highlights	Summary	Author	Source	Date
		from enacting restrictive AI laws, particularly targeting California's defeated SB 1047 bill. This move comes amid a surge in state-level AI legislation, with 893 bills introduced across 48 states in less than 80 days of 2025. OpenAI's stance reflects concerns over navigating a patchwork of state regulations that could hinder innovation.			
5.3	EU Charges Google and Apple Under Digital Markets Act	The European Commission has initiated actions against Google and Apple under the EU Digital Markets Act, alleging violations that favor their own services and restrict competition. Google is accused of promoting its services in search results and limiting developers' ability to direct users to alternative platforms. Apple has been ordered to improve interoperability with rival devices. These measures aim to foster fair competition and innovation in the tech industry. However, they risk straining EU-U.S. relations, with President Donald Trump threatening retaliatory tariffs.	By Rob Davies		March 19,2025



V. AI POLICIES, REGULATIONS & STRATEGIES					
#	Highlights	Summary	Author	Source	Date
5.4	Commerce Chief Seeks Help to Block China's Access to US Chip	U.S. Commerce Secretary Howard Lutnick has urged companies and foreign governments to help prevent China from obtaining U.S. semiconductor chips, asserting that Chinese AI company DeepSeek improperly used American-made chips. He emphasized concerns over national security and the potential consequences of losing access to Taiwanese chips. Lutnick called for export controls to be included in trade deals, encouraging countries to choose between aligning with the U.S. and Western values or prioritizing financial gain by aiding China. He also discussed efforts to boost U.S. production of critical goods like aluminum, steel, and semiconductors.	By Reuters		March 18, 2025
5.5	AI model providers see EU Commission build 'network of evaluators'	The European (EU) Commission is building a network of model evaluators to define how general-purpose AI models with systemic risk should be evaluated in accordance with the legal requirements of the AI Act and the GPAI code of practice. While the role of this network still needs to be formalized with an implementing act due in August, the participants might also provide a shortlist of "qualified" independent reviewers that AI companies and the AI Office might tap to carry out model evaluations on their behalf.	By Luca Bertuzzi		March 19, 2025



V. AI POLICIES, REGULATIONS & STRATEGIES					
#	Highlights	Summary	Author	Source	Date
5.6	X's Use of Personal Data for Grok AI Training Investigation	The Swiss Federal Data Protection and Information Commissioner (FDPIC) concluded its preliminary investigation into X's use of personal data to train its AI model, Grok. The probe revealed that X processed personal data without obtaining users' consent, breaching Swiss data protection laws. Consequently, users now have the right to object to their public posts being utilized for AI training. X has been advised to implement measures ensuring compliance with data protection regulations and to secure explicit user consent for data processing activities related to AI development.	By Swiss federal authorities		March 20, 2025
5.7	Commerce chief seeks industry help to prevent China from getting US chips	U.S. Commerce Secretary Gina Raimondo has urged the tech industry to assist in preventing China from accessing advanced American AI chips. Speaking at an event in Washington, she emphasized that national security depends on stopping sensitive technologies from reaching adversaries. Raimondo called on chipmakers and cloud providers to enhance internal controls and share information with the government. The Commerce Department has already restricted AI chip exports to China, but concerns remain about indirect access through cloud services. Raimondo stressed that maintaining America's lead in AI requires a collaborative effort between government and the private sector.	By David Shepardson		March 18, 2025
5.8	CAC issues Security Management	China's Cyberspace Administration (CAC) and Ministry of Public Security have issued new rules for facial recognition technology, effective June 1, 2025. The regulations require data collection to be purpose-specific and minimal, mandate separate user consent, and prohibit internet transmission without approval. High-risk uses need impact assessments, and storing over 100,000 facial templates	By Cyberspace Administration of China (CAC)		March 21, 2025


V. AI POLICIES, REGULATIONS & STRATEGIES					
#	Highlights	Summary	Author	Source	Date
	Measures for Facial Recognition Technology Applications	requires CAC filing. Security protocols ban using facial recognition as the sole verification method when alternatives exist, restrict scans in private spaces, and forbid coercive use. Organizations must also implement encryption, access controls, and auditing to safeguard biometric data and ensure compliance.			
5.9	UK minister in US to pitch Britain as global AI investment hub	UK Technology Minister Saqib Bhatti visited the US to position Britain as a global hub for AI investment. Highlighting the UK's "pro-innovation, pro-safety" regulatory approach, Bhatti emphasized a flexible framework that fosters AI development while ensuring ethical oversight. He pitched the UK's strengths, including access to high-quality data, leading universities, and strong public-private partnerships. The UK aims to distinguish itself from more rigid regulatory environments, appealing to global tech firms and investors. This move is part of Britain's broader strategy to lead in safe and responsible AI innovation on the world stage.	By Ryan Daws		March 20, 2025
5.10	UK's AI Copyright Proposal Sparks Debate	UK Technology Secretary Peter Kyle has urged critics of the nation's proposed AI copyright regime to embrace change rather than resist it. Addressing concerns from creative industries about potential impacts on intellectual property rights, Kyle emphasized the importance of adapting to technological advancements while ensuring fair compensation for creators. The proposed framework aims to balance innovation with the protection of artistic works, acknowledging the transformative potential of AI in various sectors. Kyle's stance highlights the	By Financial Times		March 24,2025

V. AI POLICIES, REGULATIONS & STRATEGIES					
#	Highlights	Summary	Author	Source	Date
		government's commitment to fostering an environment where technological progress and creative rights coexist harmoniously.			
5.11	AI's Potential Economic Impact	Bank of England Governor Andrew Bailey has renewed calls for coordinated international efforts to address escalating global trade tensions, highlighting particular concerns regarding technology and artificial intelligence. Bailey emphasized the importance of harmonized regulatory standards and collaborative policy frameworks to prevent fragmentation and economic instability stemming from tech-driven disputes. He warned that uncoordinated AI policies and protective economic measures could significantly disrupt global markets. His statements underscore the increasing urgency for unified global responses to technological advancements and associated economic challenges, advocating for cooperation to maintain international economic stability amid rapid technological innovation.	By David Milliken		March 24,2025
5.12	OpenAI Expands COO Brad Lightcap's Role to Drive Global Growth and Partnerships	OpenAI COO Brad Lightcap will take on an expanded role to drive global growth and partnerships, according to CEO Sam Altman. This strategic move reflects OpenAI's ambition to deepen its international footprint and forge stronger alliances across industries and governments. Lightcap's leadership will focus on operational scaling, partnership development, and policy engagement, signaling OpenAI's intent to balance innovation with regulatory collaboration. As global scrutiny of AI intensifies, OpenAI is positioning itself to navigate geopolitical	By Anna Tong		March 24,2025

V. AI POLICIES, REGULATIONS & STRATEGIES					
#	Highlights	Summary	Author	Source	Date
		dynamics while expanding responsible deployment of its technologies across borders.			
5.13	Sam Altman's Worldcoin in Talks with Visa for Stablecoin Wallet Integration	Sam Altman's Worldcoin project is reportedly in discussions with Visa to integrate its stablecoin wallet into Visa's global payment network, according to CoinDesk. The potential partnership would allow users to make everyday purchases with Worldcoin's digital currency, bridging traditional financial systems with blockchain-based identity and currency tools. This move reflects Worldcoin's strategy to expand access to digital finance and strengthen its real-world utility, while navigating evolving regulatory frameworks around stablecoins and digital identity.	By Reuters		March 24,2025
5.14	Former Intel CEO Pat Gelsinger Joins Faith-Based Tech Firm Gloo for AI Expansion	Former Intel CEO Pat Gelsinger has joined Gloo, a Colorado-based technology company focused on serving faith-based communities, to support its AI-driven outreach and engagement tools. Gloo provides churches and nonprofits with AI-enabled platforms for personalized communication, community-building, and data-driven ministry. Gelsinger's involvement aims to accelerate the firm's AI capabilities while aligning with his personal faith values. This move illustrates a growing trend of AI adoption in niche sectors like religious outreach, showing how AI tools are being tailored to support values-driven missions.	By Jeffrey Dastin		March 24,2025

V. AI POLICIES, REGULATIONS & STRATEGIES					
#	Highlights	Summary	Author	Source	Date
5.15	US-China Committee Head Urges Cooperation on AI to Avoid Conflict	The head of the US-China Relations Committee emphasized the urgent need for bilateral cooperation on artificial intelligence to mitigate risks and prevent potential conflicts. Speaking amid rising geopolitical and technological tensions, he highlighted that both nations have a shared responsibility to establish guardrails around AI development. The call for dialogue includes proposals for joint frameworks on safety, transparency, and ethical AI use. As global AI capabilities rapidly evolve, the appeal reinforces the importance of diplomacy and governance to ensure AI advances do not escalate international instability.	By Reuters		March 24, 2025
5.16	Sweden Proposes Bill to Let Police Use AI Facial Recognition	The Swedish government has proposed a bill allowing law enforcement to use AI-powered facial recognition technologies for criminal investigations. The move marks a significant shift in surveillance policy, aiming to enhance police efficiency while sparking debate over privacy and civil liberties. If passed, the legislation would permit real-time biometric analysis in public spaces, aligning Sweden with other European nations adopting AI in policing. Authorities emphasize safeguards and oversight mechanisms to prevent misuse, as the country navigates the balance between public safety and ethical AI deployment.	By Reuters		March 20, 2025

V. AI POLICIES, REGULATIONS & STRATEGIES					
#	Highlights	Summary	Author	Source	Date
5.17	In the Market: How Trump is driving Asia to diversify away from US	Fears over a potential second Trump presidency and his past protectionist trade policies are pushing Asian countries to reduce dependence on the U.S. Experts say nations like Japan, South Korea, and others are accelerating efforts to diversify supply chains and strengthen regional partnerships. This shift includes critical sectors like semiconductors and advanced tech—key components in AI development. As geopolitical uncertainty grows, Asia’s move toward economic self-reliance could reshape global AI infrastructure, investment flows, and hardware sourcing strategies, signaling a strategic pivot in how the region prepares for potential disruptions from U.S. policy shifts.	By Paritosh Bansal		March 20, 2025
5.18	Microsoft to Launch Three Data Centers in Malaysia by Q2 2025	Microsoft plans to open three data centers in Malaysia by the second quarter of 2025 as part of its broader strategy to expand cloud and AI infrastructure in Southeast Asia. The initiative supports Malaysia’s digital transformation goals and aligns with Microsoft’s commitment to responsible AI deployment and regional capacity building. These data centers will provide local businesses and governments with improved access to cloud services, AI tools, and data	By Rozanna Latiff		March 20, 2025

V. AI POLICIES, REGULATIONS & STRATEGIES					
#	Highlights	Summary	Author	Source	Date
		sovereignty. The move also reflects Microsoft’s growing investments in AI-aligned infrastructure worldwide.			
5.19	Swedish government proposes bill to allow police to use AI face-recognition	The Swedish government has introduced a bill that would allow police to use AI-powered facial recognition to investigate serious crimes and locate missing persons. The proposed legislation aims to balance public safety with privacy concerns and will be aligned with the European Union’s AI Act. If passed, it would mark a significant step in Sweden’s adoption of AI within law enforcement. The move highlights growing efforts across Europe to regulate the use of AI in sensitive areas like surveillance, raising important debates around ethics, oversight, and civil liberties.	By Reuters		March 20, 2025

The week’s chronicles reflect a pivotal moment in AI’s evolution, where technological breakthroughs intersect with strategic and societal considerations. Innovations in multimodal models, such as NVIDIA’s hybrid Mamba-Transformers and Tencent’s rapid 3D generation tools, demonstrate AI’s growing versatility in solving complex real-world problems. Hardware advancements, from Micron’s AI memory chips to AMD’s challenge to NVIDIA, underscore the critical role of scalable infrastructure in sustaining AI progress. Yet, these strides are tempered by urgent policy discussions: the EU’s push for semiconductor autonomy, debates over facial recognition regulations, and OpenAI’s call for balanced AI governance reveal the tension between rapid innovation and responsible oversight. As AI permeates sectors like gaming, healthcare, and finance, collaborations—such as NatWest’s integration of OpenAI tools—highlight its transformative potential. Moving forward, the democratization of AI through open-source frameworks and localized processing must align with global cooperation to address

security, privacy, and ethical challenges. This week's developments not only chart a path toward more capable AI systems but also emphasize the need for harmonized efforts to ensure these technologies serve humanity equitably and sustainably.