Article Name	Website Link	Date	General Topics/ Industries	Companies	AI summary	AI Key Words
AI Is Putting the Silicon Back in Silicon Valley	https://www.bloomberg.com/news/articles/2024-03-26/ai-chip-startups-like-matx-storm-silicon-valley	26-Mar-24	Competition, Hardware, LLM, Semiconductors, Software	Nvidia	Nvidia's dominance in AI chips has prompted companies like MatX to design their own semiconductors specifically for large language models (LLMs). MatX aims to create cheaper and faster AI-friendly chips to meet the growing demand for AI software. The company is betting that its chips will outperform Nvidia's GPUs by at least 10 times in training LLMs. MatX has raised $25 million and plans to release its chip next year. The rise of LLMs has sparked a surge of chip startups, but entering the chip industry is challenging due to the long design and manufacturing process. However, the shift to AI computing is expected to create new chip empires.	Nvidia, MatX, AI, silicon, silicon valley, chips, artificial intelligence, large language models, LLMs, OpenAI Inc., Google, tensor processing units, Nat Friedman, Daniel Gross, Mountain View, Shockley Semiconductor Laboratory, computing, http://amazon.com/, Microsoft, Groq Inc., Cerebras Systems Inc.
Inside the Creation of the World's Most Powerful Open Source AI Model	https://www.wired.com/story/dbrx-inside-the-creation-of-the-worlds-most-powerful-open-source-ai-model/	27-Mar-24	LLM	Databricks	Databricks has successfully built DBRX, a powerful open-source AI language model that outperforms other open-source models in various benchmarks. DBRX is similar in design to OpenAI's ChatGPT and even rivals the closed GPT-4 model. By open-sourcing DBRX, Databricks aims to challenge the secretive approach of prominent companies in the AI field and accelerate innovation. The company also plans to share insights into the creation process and hopes to assist industries like finance and medicine in leveraging AI technology. DBRX utilizes a neural network architecture and has around 136 billion parameters, demonstrating significant advancements in AI model efficiency.	Databricks, artificial intelligence, language model, OpenAI, ChatGPT, Meta, Llama 2, Mistral, Gemini, Grok AI, Elon Musk, GPT-4, Google, MosaicML, Nvidia H100s GPUs, EleutherAI
Google's new technique gives LLMs infinite context	https://venturebeat.com/ai/googles-new-technique-gives-llms-infinite-context/	12-Apr-24	LLM	Google	Google has developed a technique called Infini-attention that allows LLMs (Language Model Models) to work with text of infinite length by extending their "context window" without requiring additional memory. This technique enhances model performance, coherence, and summarization tasks while reducing memory requirements. LLMs with infinite context have applications in customizing models, improving performance on specific tasks, and reducing engineering efforts in model customization.	LLMs, Google
SoftBank to spend $960m to boost computing power for generative AI	https://asia.nikkei.com/Business/Technology/SoftBank-to-spend-960m-to-boost-computing-power-for-generative-AI2?utm_campaign=IC_one_time_free&utm_medium=email&utm_source=NA_newsletter&utm_content=article_link&del_type=3&pub_date=20240428093000&seq_num=18&si=0da7fc56-79df-4037-bd3f-e874dae0c0e2	22-Apr-24	Datacenters, LLM, Semiconductors, Supply Chain	Nvidia, Softbank	Japanese telecom company SoftBank is investing $960 million to boost its computing power for generative artificial intelligence (AI). The investment will make SoftBank's computing facilities among the highest in Japan and will be used to develop a Japanese-language-specific generative AI with world-class performance. SoftBank will purchase graphics processing units (GPUs) from Nvidia and will also loan them to other companies. The company is developing a large language model (LLM) and plans to complete a model with 390 billion parameters in fiscal 2024, with a high-performance model of 1 trillion parameters in development for next year.	Nvidia, SoftBank, Datacenters, LLM, Semiconductors, Supply Chain

Article Name	Website Link	Date	General Topics/ Industries	Companies	AI summary	AI Key Words
Apple Releases Small, Open-Source AI Models for On-Device Applications	https://www.theinformation.com/briefings/apple-releases-small-open-source-ai-models-for-on-device-applications?utm_campaign=%5BREBRAND%5D+%5BTI-AM%5D+Th&utm_content=1095&utm_medium=email&utm_source=cio&utm_term=124&rc=hm8aii	24-Apr-24	LLM, New Features	Apple	Apple has released a new family of small, open-source language models called OpenELM, signaling their focus on developing AI software for small devices. The models have significantly fewer parameters compared to recent releases from other companies, making them suitable for on-device applications. This release marks a departure from Apple's closed ecosystem approach and suggests their interest in licensing AI models from other companies.	Key Words/Topics: Apple, OpenELM, language models, artificial intelligence, on-device applications, parameters, open-source, Meta Platforms, Llama 3, Google, Microsoft, closed ecosystem, generative AI models, iPhone Companies: Apple, Meta Platforms, Google, Microsoft, OpenAI
Nvidia’s AI chatbot now supports Google’s Gemma model, voice queries, and more	https://www.theverge.com/2024/5/1/24146130/nvidia-chatrtx-google-gemma-ai-model-features-updates	1-May-24	Digital Assistants, LLM, New Features, Software	Nvidia	Nvidia is updating its ChatRTX chatbot with new AI models, including Google's Gemma, ChatGLM3, and OpenAI's CLIP, to enhance search capabilities for RTX GPU owners. The chatbot, which runs locally on a Windows PC, allows users to query personal documents and search photos. The update also introduces voice query support using Nvidia's integrated Whisper AI speech recognition system.	Nvidia, Google, Gemma, ChatRTX, Mistral, Llama 2, OpenAI, CLIP, RTX GPU, ChatGLM3, Whisper