AI News

In this issue of AI News magazine, we introduce a myriad of AI advancements, challenges, and future projections, aimed at keeping our readers informed and connected with the AI community globally.

We begin with the Claude 3 AI model family, the latest innovation in AI intelligence. This model puts us a step ahead in cognitive tasks, content creation, and analysis, with Haiku, Sonnet, and Opus poised to provide smarter, faster, and safer AI tools.

Ema, a San Francisco startup focusing on automating mundane tasks and revolutionizing the workplace, has emerged from stealth mode, promising to deliver a more unified and efficient AI solution for large enterprises.

On the technical front, Daniel and Michael Han's keen efforts to debug Google's AI model, Gemma, serve as a testament to the potential growth and enhancement of AI with community support.

While AI models like GPT-4 and Gemini continue to show impressive abilities, we explore the underlying principles and challenges that need to be addressed for a better understanding and control of these models.

The launch of Elon Musk's open-source AI chatbot Grok has sparked a broader conversation in the industry about open-source AI's role in pushing technological progress forward.

We cover significant projects like Punica's innovative way to efficiently serve multiple Large Language Models using Low Rank Adaptation, and Answer.AI's breakthrough in fine-tuning large AI models using just two consumer-grade GPUs.

Join us in meeting Devin, the first AI software engineer developed by Cognition and creating a new wave of potential AI and human collaboration in the tech world.

Catch up with the upcoming launch of Microsoft’s AI Copilot for Security, aiming to strengthen cybersecurity, especially in the face of continuous global cyber threats and sophisticated attacks.

Finally, we touch base on the urgent need for stronger safety measures to combat new hacking methods such as ArtPrompt, a vulnerability unveiled in 5 major AI chatbots.

As we delve into these topics, our aim is to inspire innovation and stimulate conversation on the AI frontier. While the future remains uncertain and full of potential challenges, the prospects for positive impact are limitless and exciting, encouraging us all to engage deeply with the world of artificial intelligence.

We invite you to immerse yourself in these stories that reflect both the achievements and the work ahead. Enjoy the read!

Introducing the next generation of Claude

2024-03-04

Meet the Claude 3 AI model family: Haiku, Sonnet, and Opus. These models set new standards in AI intelligence, excelling in cognitive tasks, analysis, content creation, and multilingual conversation. With impressive speed, advanced vision capabilities, and improved accuracy, they promise smarter, faster, and safer AI tools. Designed responsibly, they address biases and safety concerns. Opus and Sonnet are now available, with Haiku coming soon.
Read more...

Ema, a ‘Universal AI employee,’ emerges from stealth with $25M

2024-03-05

San Francisco startup Ema aims to revolutionize the workplace with its universal AI employee designed to automate mundane tasks. The company integrates over 30 large language models with domain-specific models for enhanced accuracy and data protection. With $25 million in funding and a team of experienced founders, Ema's products, Generative Workflow Engine and EmaFusion, are set to transcend traditional robotic process automation and AI task acceleration, offering a unified and efficient AI solution for large enterprises.

Read more...

Unsloth Fixing Gemma bugs

2024-03-07

Developers Daniel and Michael Han have successfully fixed a series of bugs in Google's AI model, Gemma. The bugs ranged from simple typos to complex issues like incorrect casting in Keras and precision errors in RoPE calculations. The fixes, which include improved handling of layer normalization and the GELU activation function, are now available in the latest Hugging Face transformers version and Colab notebooks. The duo continues to enhance Gemma and encourages community support through donations and engagement on their platforms.

Read more...

Large language models can do jaw-dropping things. But nobody knows exactly why.

2024-03-10

Machine learning, particularly large language models like OpenAI’s GPT-4 and Google DeepMind’s Gemini, is experiencing rapid progress through trial and error, despite a lack of understanding of the underlying principles. These models can generalize knowledge from specific examples to new situations, even solving math problems in untrained languages. However, challenges like overfitting persist, and the process remains more art than science. Interestingly, the largest models continue to improve with size, defying statistical expectations and conventional wisdom.
Read more...

Elon Musk to open-source AI chatbot Grok this week | TechCrunch

2024-03-11

Elon Musk’s AI startup xAI is set to open-source its chatbot Grok, positioning it as a competitor to OpenAI’s ChatGPT. Musk, who co-founded OpenAI, has sued the company, accusing it of abandoning its open-source commitments and becoming a closed-source entity under Microsoft’s influence. Grok, part of xAI’s $16 monthly subscription service, boasts real-time information access and an unfiltered approach to content. The move to open-source Grok aligns with Musk’s longstanding advocacy for open technology, reminiscent of Tesla’s open-sourced patents and Twitter’s open-sourced algorithms. The lawsuit has sparked a wider debate on the role of open-source AI, with industry figures like Vinod Khosla and Marc Andreessen weighing in on its implications for the future of artificial general intelligence (AGI) and technological progress. [Read more at TechCrunch…]

GitHub – punica-ai/punica: Serving multiple LoRA finetuned LLM as one

2024-03-12

Punica introduces a novel method to efficiently serve multiple Large Language Models (LLMs) using Low Rank Adaptation (LoRA). It leverages small matrices to modify pretrained model weights, running multiple LoRA models at the computational cost of one. The Segmented Gather Matrix-Vector multiplication (SGMV) CUDA kernel ensures efficiency and reduced latency. Punica outperforms other systems in benchmarks, offering up to 12 times the throughput in text generation tasks. It can be installed from a binary package or built from source, providing examples for various operations.
[Read more…]

Devin, the first AI software engineer

2024-03-13

Meet Devin, the first AI software engineer developed by Cognition, capable of autonomously executing complex engineering tasks and collaborating with human engineers. Devin has set a new record on the SWE-bench coding benchmark, resolving 13.86% of real-world GitHub issues unassisted. Cognition, backed by a $21 million Series A funding, is offering early access to Devin for engineering tasks.
Read more...

Microsoft's AI Copilot for Security launches next month with pay-as-you-go pricing

2024-03-14

Microsoft is set to introduce Copilot for Security, a generative AI chatbot aimed at aiding cybersecurity professionals. Leveraging OpenAI's GPT-4 and Microsoft's security data, the tool offers real-time insights into security incidents and threat summaries. Unique features include a collaborative pinboard and event summarization for reporting. Unlike the fixed pricing of Copilot for Microsoft 365, Copilot for Security will adopt a consumption-based model, charging $4 per hour starting April 1st, allowing businesses to scale their AI cybersecurity investment.

This initiative is part of Microsoft's broader strategy to bolster its software security, especially in light of recent cyberattacks by state-sponsored groups like Nobelium and vulnerabilities exploited in Microsoft's Azure cloud and Exchange Server. The company is actively overhauling its security measures to prevent future breaches and improve resilience against sophisticated cyber threats.
Read more at The Verge...

Answer.AI – Enabling 70B Finetuning on Consumer GPUs

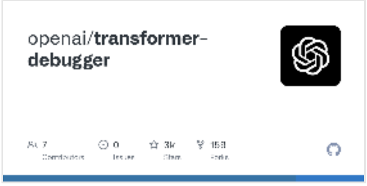
2024-03-14

Answer.AI has launched FSDP+QLoRA, an open-source project that enables the fine-tuning of large AI models with up to 70 billion parameters using just two consumer-grade GPUs. The project integrates Fully Sharded Data Parallel with quantization libraries, allowing efficient model sharding and mixed precision training. This breakthrough has been incorporated into Axolotl and the Hugging Face ecosystem, paving the way for more efficient training of large-scale models on limited hardware.

Read more...

GitHub – openai/transformer-debugger

2024-03-14



openai/**transformer-debugger**

OpenAI's Superalignment team has developed the Transformer Debugger (TDB), a tool that simplifies the exploration of small language model behaviors. TDB uses automated interpretability techniques and sparse autoencoders to investigate model decisions. The release includes a Neuron viewer React app, an Activation server, a simple inference library for GPT-2 models, and top-activating example datasets. The repository provides setup instructions and a guide for validating updates. The tool can be cited in research.

Read more...

ASCII art elicits harmful responses from 5 major AI chatbots

2024-03-16

Researchers have discovered a new hacking method, ArtPrompt, that uses ASCII art to trick AI assistants into providing responses they're programmed to reject. The study found that AI systems, including GPT-4, prioritize recognizing ASCII art over safety protocols, leading to potential illegal or unethical behavior. This highlights a significant vulnerability in AI's understanding of context and the need for stronger safety measures.

Read more...