AI News

Welcome to the latest issue of AI News Magazine, where new beginnings, strides forward in technology, and pressing cautionary tales shape our agenda.

Headlining is Apple's pursuit of an AI tool aimed at easing app development, reflecting a broader trend towards making AI technology more accessible, and foreshadowing an exciting phase of AI integration. An equally noteworthy development has been the introduction of Gemma, Google's latest AI model that exemplifies a commitment to safe and accountable AI innovation. However, the suspension of Google's Gemini AI model due to ongoing concerns over AI bias serves as a solemn reminder of the ethical hurdles we must overcome in the AI landscape.

This issue is also rich with ground-breaking strides in generative AI models. The debut of Stable Diffusion 3.0 is a testament to the potential of AI in the visual media realm. Contrastingly, the deep dive into the role of Large Language Models in prompt engineering highlights the potential implications for performance outcomes, signifying a move towards automatic prompt optimization.

A concerning revelation in this issue is the deployment of OpenAI’s GPT-4 for hacking, exemplifying the grave risks associated with the misuse of AI technology. This serves as a stark reminder to us all of the pertinent ethical discussions we must pursue in the realm of AI technology, especially around cybersecurity.

While on the topic of GPT-4, the newly released FireFunction-v1 offers a promising alternative with superior function calling capacity. Coupled with the announcement of novel AI assistant, Phind-70B, which challenges GPT-4's dominance in the coding assistant realm, it is clear that the competition in AI innovation is fierce and far from over.

With the successful implementation of Klarna's AI assistant and the release of the pioneering 1-Bit large language model, the BitNet b1.58, it's evident that the boundaries of AI are constantly being pushed.

Finally, Microsoft's commercial foray into AI with the launch of GitHub’s Copilot Enterprise, underscores the increasing prominence of AI in everyday business operations and offers an enticing glimpse into a more

intelligent, connected future.

Packed full of enlightening insights and the latest developments, we hope this issue of AI News Magazine inspires, informs and prompts key discussions about the AI landscape as it constantly evolves before our eyes. Enjoy the read.

Apple Developing AI Tool to Help Developers Write Code for Apps

2024-02-17

Apple is developing an AI tool for its Xcode environment to assist developers in writing app code, similar to Microsoft's GitHub Copilot. The tool will generate code from natural language requests and translate between programming languages. Apple is also integrating AI into app testing, Siri, and native apps, and enhancing Spotlight search. The focus on AI innovations is directed towards the upcoming iOS 18, iPadOS 18, and macOS 15 updates, with a phased approach to AI integration planned for the future.
Read more...

Gemma a quasi open model from Google

2024-02-21

Google introduces Gemma, a new AI model focused on safety and reliability, supported by a Responsible Generative AI Toolkit. Compatible with various frameworks and devices, Gemma allows fine-tuning on specific datasets. Developers can access Gemma for free on platforms like Kaggle and Colab notebooks, with resources and guides available for assistance.
Read more...

Google pauses AI-generated images of people after ethnicity criticism

2024-02-22

Google has suspended its AI model, Gemini, from generating images of people following criticism over inaccurate portrayal of historical figures of different ethnicities and genders. The move underscores the ongoing issue of AI bias, with investigations revealing unfavorable depictions of people of color and women. Google acknowledges the need for improvement, especially in historical contexts, while experts suggest that mitigating AI bias remains a challenging task.
Read more...

Stable Diffusion 3.0 debuts new diffusion transformation architecture to reinvent text-to-image gen AI

2024-02-22

Stability AI introduces Stable Diffusion 3.0, a state-of-the-art text-to-image generative AI model with enhanced image quality and performance. The model, built on a novel diffusion transformer architecture, excels in multi-subject prompts and improved typography. It also incorporates a new training method, flow matching, for better speed and performance. Stability AI plans to extend its capabilities to 3D and video generation, marking a significant advancement in generative AI models for visual media.
Read more...

Prompt engineering is a task best left to AI models

2024-02-23

Large language models (LLMs) have sparked a new focus on prompt engineering, a technique to craft prompts that yield better AI responses. Research by Rick Battle and Teja Gollapudi from VMware has highlighted the significant impact of subtle prompt variations on AI performance. They argue against the common trial-and-error approach, suggesting instead the use of automatic prompt optimization, where an LLM refines prompts to enhance performance on benchmark tests.

Their research tested whether smaller, open-source models could also serve as effective optimizers. Using models like Mistral-7B and Llama2-70B, they demonstrated that even with limited data samples, automatic optimizers can improve LLM performance. Interestingly, these optimizations can lead to unexpected strategies, such as a model's mathematical reasoning being improved by expressing a liking for Star Trek. This finding underscores the potential of LLMs to develop prompt strategies beyond human intuition.
Read more…

GPT-4 developer tool can hack websites without human help

2024-02-23

OpenAI's GPT-4 has demonstrated the alarming ability to hack websites and extract information from databases autonomously, a discovery made by researchers that raises concerns about the potential misuse of AI. This advancement indicates that even those without any hacking skills could potentially deploy AI to conduct cyber attacks, significantly lowering the barrier to entry for such malicious activities. Daniel Kang from the University of Illinois Urbana-Champaign highlights the ease with which the AI can operate independently, emphasizing the reduced need for technical expertise in carrying out cyber intrusions. The implications of this finding suggest a pressing need for discussions on AI ethics and cybersecurity measures.
Read more at New Scientist…

FireFunction V1 – Fireworks' GPT-4-level function calling model – 4x faster than GPT-4 and open weights

2024-02-23

Fireworks introduces FireFunction-v1, a superior function calling model that integrates external knowledge into large language model applications. It outperforms GPT-4 in real-world use cases, offering faster response times and open-source flexibility. The model excels in structured output generation, decision-making, and multilingual input accuracy. Available on Hugging Face and the Fireworks platform, it invites developers to participate in its ongoing development.
Read more…

Phind-70B closes the code quality gap with GPT-4 while running 4x faster

2024-02-25

Phind's new AI assistant, Phind-70B, matches OpenAI's GPT-4 Turbo in code quality while offering 4x faster inference. The model scores 82.3% on the HumanEval benchmark, slightly surpassing GPT-4 Turbo, and provides more responsive, detailed code generation. Phind-70B's speed and performance position it as a strong contender in the realm of AI coding assistants, potentially revolutionizing software development.
Read more...

echo-embeddings

2024-02-26

Echo embeddings enhance autoregressive language models by incorporating future token information, demonstrating strong performance on the MTEB benchmark. A pretrained model is available on HuggingFace, with code snippets for easy implementation. The method allows for sentence-level comparisons and the calculation of cosine similarity between different text inputs.
Read more...

Mistral AI releases new model to rival GPT-4 and its own chat assistant | TechCrunch

2024-02-27

Paris-based startup Mistral AI has launched Mistral Large, a large language model set to rival GPT-4 and Claude 2. The company has also introduced Le Chat, a beta chat assistant service. Shifting from an open-source approach to a paid API model, Mistral AI has partnered with Microsoft to offer its models to Azure customers, expanding its reach and aligning with Microsoft's AI hosting strategy.
Read more...

Klarna AI assistant handles two-thirds of customer service chats in its first month

2024-02-27

Klarna's AI assistant, developed with OpenAI, has handled 2.3 million customer service chats in its first month, equating to the work of 700 full-time agents. The multilingual assistant, available 24/7 in 23 markets, matches human agents in customer satisfaction and reduces repeat inquiries by 25%. Integrated into the Klarna app, it supports services from customer service to financial management, contributing to an estimated $40 million in profit for Klarna in 2024.
Read more...

The Dawn of 1-Bit Large Language Models

2024-02-28

The BitNet b1.58, a 1.58-bit large language model (LLM), matches the performance of 16-bit LLMs while reducing memory usage, latency, and energy consumption. Its ternary parameterization allows for more efficient computation and could lead to specialized hardware for 1-bit LLM inference. This breakthrough paves the way for deploying larger LLMs on edge devices and mitigates environmental and economic costs of massive models, marking a new era of efficient and accurate LLMs.
Read more...

GitHub's Copilot Enterprise is now generally available at $39 a month | TechCrunch

2024-02-29

GitHub has launched Copilot Enterprise, an AI-powered code completion tool for large businesses, available at $39 per month. The tool offers access to an organization's internal code and knowledge base, integration with Microsoft Bing, and a future fine-tuning feature for customized AI models. The company aims to integrate Copilot more seamlessly into developers' workflows, with no plans to differentiate pricing based on AI model size.
Read more...