AI News

Welcome to the latest issue of AI News Magazine, your one-stop shop for groundbreaking developments, intriguing insights, and ongoing debates within the rapidly evolving world of artificial intelligence.

Our featured story for this edition hails from the impressive minds at Google DeepMind and University of Southern California. They've made a significant breakthrough by developing SELF-DISCOVER, a unique method enabling large language models (LLMs) to dynamically compose reasoning structures for complex problem-solving. This innovation marks an exciting step toward human-AI collaborative problem-solving and prompts us to contemplate a future where AI's autonomy and reasoning capabilities are ever-evolving.

Our current issue further delves into the progressive roadmap of LLMs. An outstanding example is DEJAVU, a transformative method that enhances the efficiency of LLMs, rendering them six times faster than their standard counterparts, all without compromising learning capability.

In the realm of accessible technology, Hugging Face's new Hugging Chat Assistant and the launch of Google's AI-powered Gemini app deserve mention. These offerings put the power of AI in more hands, affording a myriad of users the ability to harness AI's potential for customizable chatbots and integrated tech interactions.

AI is also making strides in fine-tuning and learning processes. The advent of Self-Play fIne-tuNing (SPIN) shows how LLMs can be strengthened without additional human-annotated data. Meanwhile, OpenAI's ChatGPT proves that learning over time can vastly enhance user experiences.

But it's not just successes we spotlight in this issue. We also delve into the mixed reviews Google's Gemini Advanced plan has garnered, iterating that the perfect blend of AI efficacy lies in aligning the technology's capabilities with user expectations and needs.

Copyright, authenticity, and data control issues also feature prominently in this issue. We cover the latest developments in OpenAI's ongoing copyright infringement dispute, the unveiling of Sora, a text-to-video AI

model by OpenAI, and the robust privacy measures integrated into ChatGPT. As AI continues to evolve, these areas remain crucial for maintaining an ethical and fair AI landscape.

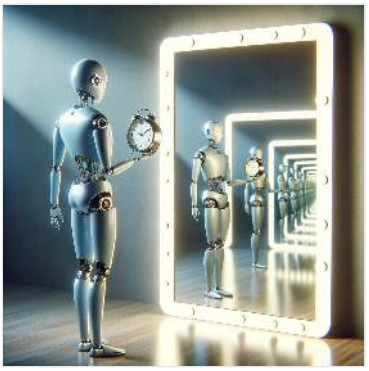
Lastly, our conversation wouldn't be complete without mention of the impressive progress noted in text-to-speech model BASE TTS and the revolutionary 'sampling-and-voting' technique introduced by Tencent Inc. researchers.

As always, we're thrilled to bring you the latest news from the AI industry; we believe these advancements hint at a future where AI technology permeates and enhances our daily lives to an extent hitherto imagined only in the realms of science fiction.

Enjoy the read!

DEJAVU: 6x faster transformers' inference

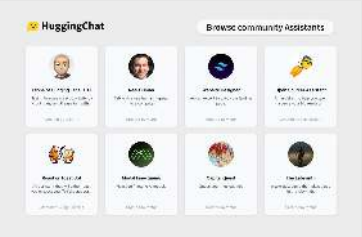
2024-02-03



The Deja Vu method introduces a way to make Large Language Models (LLMs) more efficient by implementing contextual sparsity, turning off components based on input. This approach preserves the model's learning ability and significantly reduces runtime costs, making it six times faster than standard transformer implementations. The method also reveals contextual sparsity in attention blocks, likened to mean-shift clustering, where denser regions gain more weight, resulting in stronger bonds and higher attention scores.
Read more...

Hugging Face makes it easier to create its custom chatbots.

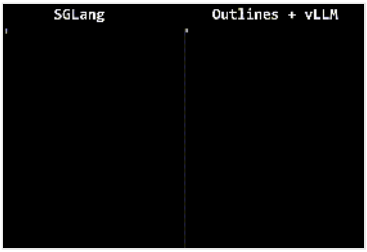
2024-02-04



Hugging Face has streamlined the process of creating custom chatbots with its new Hugging Chat Assistant, as announced by tech lead Philipp Schmid. This user-friendly feature allows individuals to generate their own chatbots with just a couple of clicks, and these bots are subsequently made public. Schmid highlights the simplicity and accessibility of the tool by comparing it to OpenAI's GPTs feature, emphasizing that it supports the use of various open large language models (LLMs) such as Llama2 or Mixtral. This development marks a significant step towards democratizing the creation of conversational AI, making it more accessible to a wider audience.
Read more at The Verge...

Fastest JSON Decoding for Local LLMs with Compressed Finite State Machine | LMSYS Org

2024-02-05



A new optimization technique significantly boosts JSON decoding in Large Language Models (LLMs), reducing latency by 50% and increasing throughput by 150%. The method combines finite state machine-based and interleaved-based methods for efficient decoding, even addressing tokenization boundary issues. Successfully tested in production use cases, this feature is now accessible in SGLang.
Read more...

Large Language Models Learn to Self-Compose Reasoning Structures

2024-02-09

Google DeepMind and University of Southern California researchers have developed SELF-DISCOVER, a method that enables large language models (LLMs) to dynamically compose reasoning structures for complex problem-solving. The system guides LLMs to select, adapt, and implement reasoning skills without training data or human involvement. Tests show that SELF-DISCOVER significantly improves LLM reasoning capabilities, outperforming other methods in efficiency and accuracy. This innovation could pave the way for collaborative problem-solving between humans and AI.
Read more...

Google introduces AI-powered Gemini app and casts aside Bard

2024-02-09

Google has launched the Gemini app, a new AI-powered application for Android smartphones, signaling a shift away from its previous chatbot, Bard. Gemini, which will also be integrated into Google's search app for iPhones, is expected to become the primary interface for users to interact with Google's AI technology. Alongside the free version, Google is introducing Gemini Advanced, a subscription service priced at $20 per month, featuring the "Ultra 1.0" AI capable of tutoring, programming assistance, and content creation. This advanced service includes 2 terabytes of storage and aims to leverage Google's nearly 100 million subscribers. The release of Gemini is part of a broader trend to integrate more AI into smartphones and intensifies the competition with Microsoft, especially in light of their ChatGPT-4 chatbot. As AI becomes more sophisticated, concerns about misuse and the need for regulation are growing, with Europe already implementing rules and the US considering similar measures.
Read more at euronews...

Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models

2024-02-10

Harnessing the power of Supervised Fine-Tuning (SFT) is crucial for the evolution of Large Language Models (LLMs). A new fine-tuning method, Self-Play fIne-tuNing (SPIN), has been proposed to strengthen LLMs without additional human-annotated data. SPIN utilizes a self-play mechanism, allowing the LLM to generate its own training data and refine its responses by comparing them to human-annotated examples. This process incrementally improves the LLM, maximizing the use of human-annotated data in SFT. Theoretical proofs confirm that SPIN's training objective is optimized when the LLM's policy matches the target data distribution. In practical tests, including on the HuggingFace Open LLM Leaderboard and other benchmarks, SPIN not only enhanced LLM performance but also surpassed models trained with direct preference optimization using extra GPT-4 data. These findings highlight the potential of self-play to reach human-level LLM performance without the need for expert human input.
Read more…

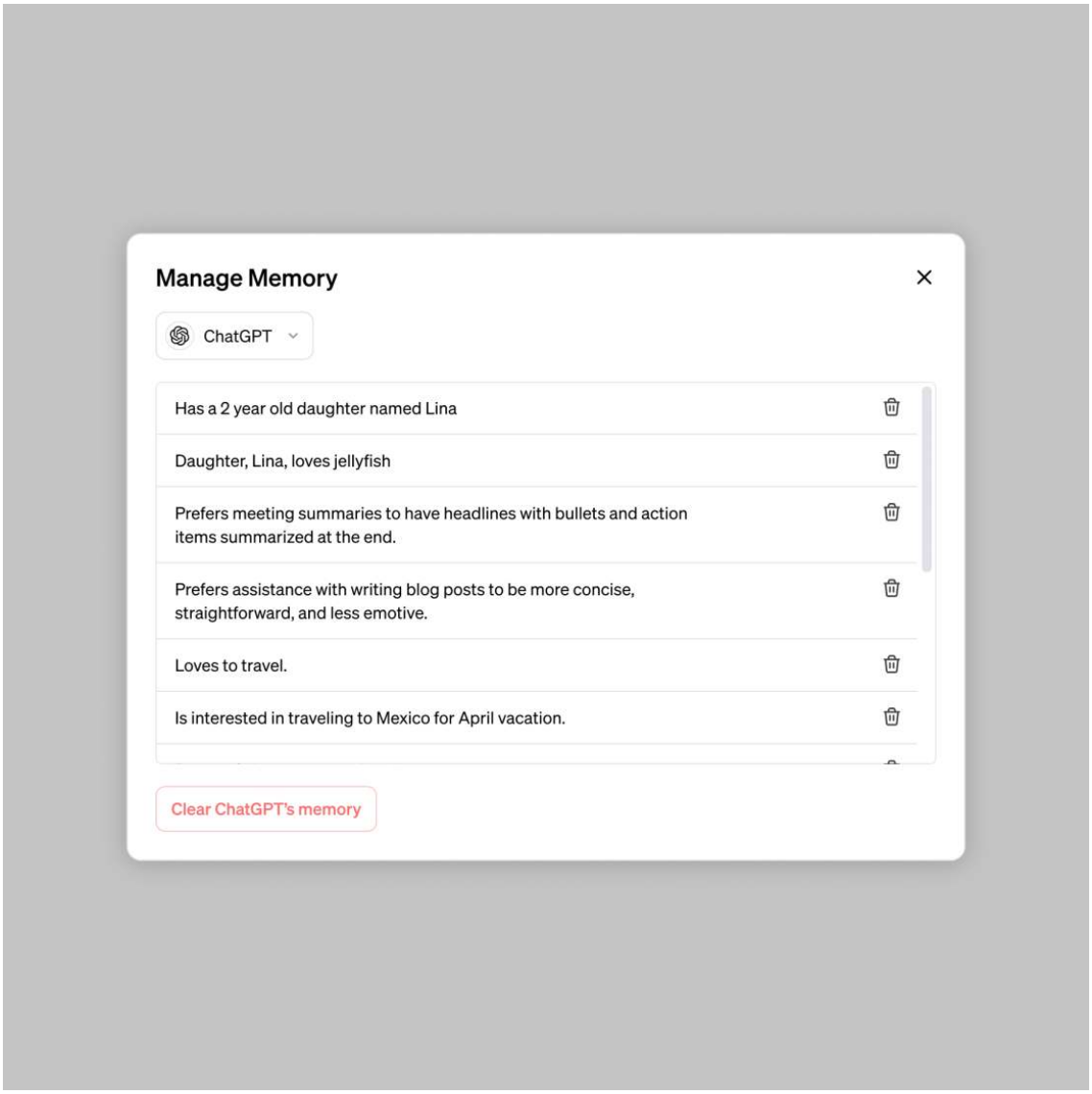
Google’s paid Gemini Advanced plan is getting mixed reviews

2024-02-12

Google’s Gemini Advanced, the paid version of its AI assistant, is eliciting a polarized response from users. While some are expressing dissatisfaction due to the assistant’s hallucinations and logical errors, others are finding value in its capabilities as a chatbot and research assistant. The user experience appears to hinge on the specific tasks Gemini Advanced is employed for, with its performance varying according to individual needs and preferences. The debate over its effectiveness is particularly lively on the Google Bard subreddit, where users are sharing their diverse experiences. As Google stakes its claim in the competitive AI landscape, the reception of Gemini Advanced underscores the subjective nature of technology adoption and the importance of aligning tools with user expectations.
Read more at XDA Developers…

Memory and new controls for ChatGPT

2024-02-13



OpenAI's ChatGPT now offers a memory feature for Enterprise and Team users, enhancing productivity by learning and adapting to user preferences over time. This advancement allows ChatGPT to retain information about a user's style, tone, and specific work-related

requirements, such as coding languages and data visualization preferences. As a result, users benefit from more personalized and efficient interactions, with ChatGPT applying learned preferences to tasks like drafting blog posts, coding, and generating business reports. Importantly, OpenAI ensures user control over their data, with the ability to manage memory usage and the option for Enterprise account owners to disable the feature entirely. The memory capability is part of a broader rollout aimed at improving the ChatGPT experience for professional users.
Read more…

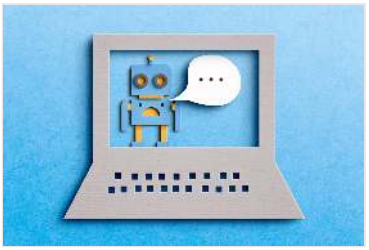
Judge rejects most ChatGPT copyright claims from book authors

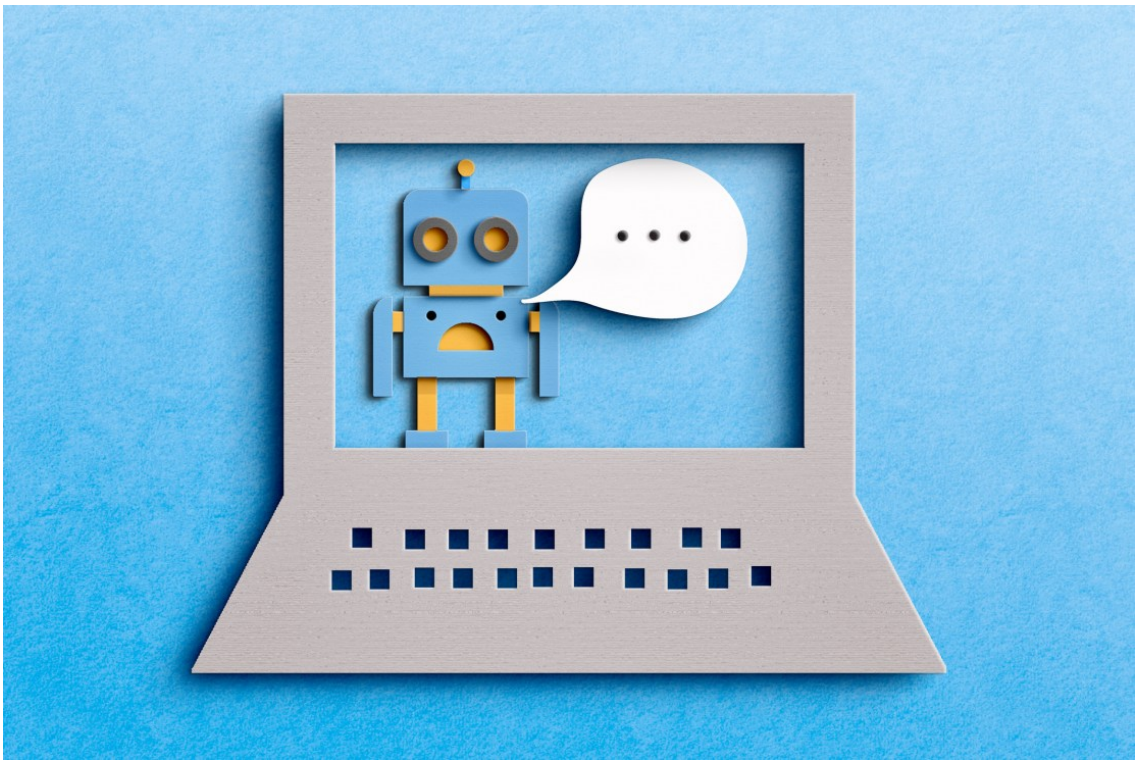
2024-02-14

A US judge has dismissed most copyright claims against OpenAI by authors accusing the firm of training its ChatGPT language model on pirated books. The judge ruled that the authors failed to provide sufficient evidence, except for direct copyright infringement. The case, which could impact copyright law in the AI era, will proceed under California's unfair competition law. The authors have until March 13 to amend their complaints.
Read more…

Largest text-to-speech AI model yet shows ’emergent abilities’

2024-02-15





Amazon researchers have developed BASE TTS, the largest text-to-speech model to date, with 980 million parameters and trained on 100,000 hours of speech. This model demonstrates emergent abilities, handling complex linguistic challenges such as compound nouns, emotional expressions, foreign words, and paralinguistics more effectively than previous models. The medium-sized version of BASE TTS, with 400 million parameters, notably exhibited a significant improvement in these areas. The model’s streamable nature allows for real-time, low-bitrate speech generation, with the potential to enhance accessibility features. Although impressive, the model remains experimental, and its source data has not been published due to potential misuse by bad actors.
Read more at TechCrunch…

Scaling Up Language Models with Agent Ensembles

2024-02-15

Boosting the number of agents in an ensemble can enhance the performance of large language models (LLMs) across various tasks, according to Tencent Inc. researchers. Their "sampling-and-voting" technique improved results for models like Llama2-Chat and GPT-3.5 Turbo, with smaller models matching or surpassing larger ones. The method could make deploying LLMs more affordable, potentially enabling wider access to powerful AI tools. The researchers aim to reduce the computational expenses of ensembling in future work.
Read more...

Gemini 1.5: A Giant Leap in Long-Context AI

2024-02-15

Google DeepMind's latest AI system, Gemini 1.5, can understand and reason over long context across multiple modalities. Its expanded context length allows it to process up to 10 million tokens of context, a significant increase from previous models. Gemini 1.5 excels in retrieving information from large data sets and learning new skills. Despite its long-context capabilities, it matches or exceeds the performance of previous models while requiring less resources. This advancement could enable new practical applications, but also raises questions about safety and control of such powerful AI systems.
Read more...

OpenAI introduces Sora, its text-to-video AI model

2024-02-15

OpenAI has revealed Sora, a text-to-video AI model that can generate videos from text prompts. Capable of creating complex scenes with multiple characters and detailed backgrounds, Sora can also enhance still images into videos. Despite promising demos, challenges remain, including accurately simulating physics and cause-and-effect relationships. The model is currently being tested for potential risks and authenticity issues to prevent its creations from being mistaken for real footage.
Read more...