Newsletter Data & AI

25/03/2024

THE SHORT OF IT ⏱️

- **World Models:** DeepMind introduced Genie, a world model trained on unlabeled video data to learn and create interactive environments, advancing pursuit of comprehensive world representations.
- **LLM and Cyber Security:** While large language models pose potential threats as hacking agents, they also face risks of being stolen via API calls that can retrieve pieces of closed-source models.

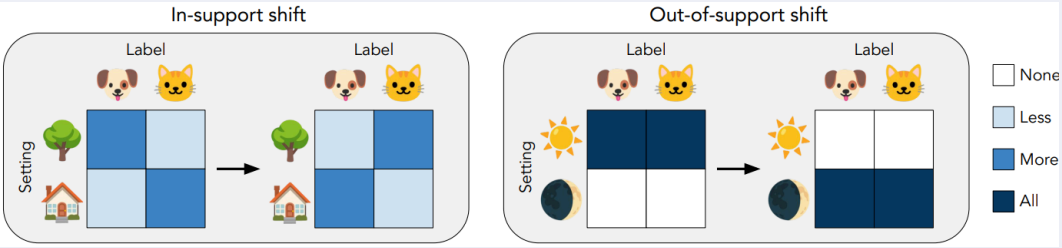
Trends

- [Paper] Why do Random Forests Work? Understanding Tree Ensembles as Self-Regularizing Adaptive Smoothers

  By framing tree ensemble models such as random forests and gradient boosting as adaptive, self-regulating smoothers, the study unlocks new insights into their efficacy. It reveals that their superior performance arises not just from variance reduction but through a combination of smoothing effects, improved function learning, and an expanded hypothesis space, challenging the traditional bias-variance framework.

- [Paper] Ask Your Distribution Shift if Pre-Training is Right for You

  Researchers from MIT investigate the variable success of pre-training in enhancing model robustness to distribution shifts, pinpointing its limitations in addressing poor extrapolation and training data biases. Their findings suggest pre-training aids in mitigating extrapolation challenges but falls short against dataset biases. They propose that integrating pre-training with bias prevention strategies and fine-tuning on small, de-biased datasets could markedly advance model robustness, presenting a strategic avenue for crafting more adaptable AI systems.



- [Paper] Stealing Part of a Production Language Model

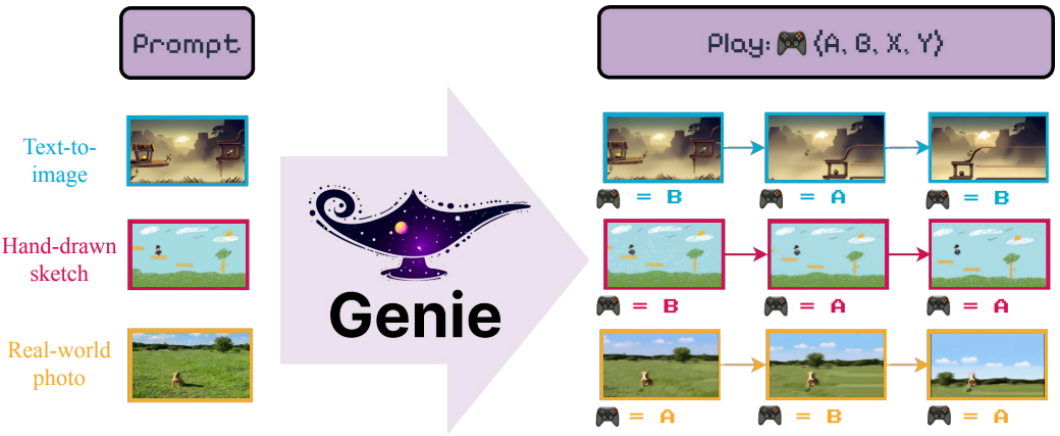
  In a pioneering effort, this research introduces an innovative attack strategy that pierces through the confidentiality of prominent language models, such as OpenAI's ChatGPT and

Google's PaLM-2, via standard API usage. By employing a cost-effective method to reveal the dimensions of the models' embedding projection layers, it illuminates key architectural insights. Additionally, the document proposes strategies for safeguarding against such breaches and ponders the future impact of refining these attack techniques.

State Of The Art

- [Paper] Genie: Generative Interactive Environments

  Google DeepMind's Genie is a novel generative interactive environment trained unsupervised from unlabeled Internet videos, capable of generating diverse, action-controllable virtual worlds through text, images, and sketches. This foundation world model, with 11 billion parameters, combines a spatiotemporal video tokenizer, an autoregressive dynamics model, and a latent action model to enable user interaction without ground-truth action labels. Genie's latent action space also facilitates the training of agents to imitate unseen video behaviors, heralding the development of adaptable generalist agents.



- [Paper] The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits

  Microsoft Research introduces BitNet b1.58, a pioneering 1-bit Large Language Model (LLM) variant where each model parameter is ternary {-1, 0, 1}. This model rivals traditional full-precision Transformer LLMs in perplexity and task performance while offering substantial improvements in latency, memory usage, throughput, and energy efficiency. BitNet b1.58 establishes a new benchmark for scaling future LLMs, promising high performance and cost efficiency. It also paves the way for a novel computation paradigm and the development of specialized hardware tailored for 1-bit LLMs.

Miscellaneous

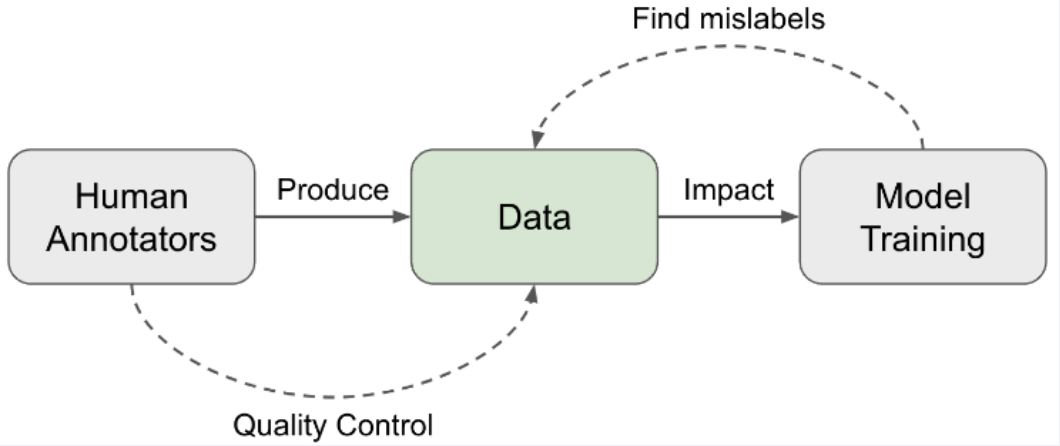
- [Blog] Show Me The Prompt.

  This blog post critiques automation libraries (e.g., guardrails, DSPy) that obscure LLM prompting processes. It advocates for using mitmproxy to reveal the actual prompts sent to

LLMs, like guidance and langchain, enabling users to critically assess and potentially refine or bypass these tools for enhanced LLM interactions.

- [Blog] Thinking about High-Quality Human Data

  Lilian Weng, an AI safety and alignment team leader at OpenAI, delves into the complexities of high-quality human data for deep learning in her blog. She underscores the importance of striking a balance between leveraging crowdsourced wisdom and mitigating biases, aiming to enhance data collection, annotation, and aggregation practices to refine model training with precise human insights.



- [Package] Mosaic

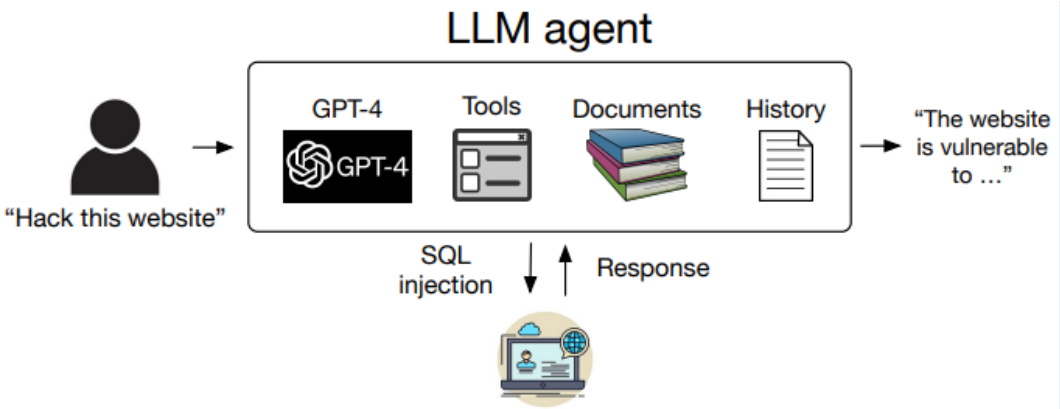
  `Mosaic` is an extensible framework that integrates databases with interactive views, enabling the exploration of vast datasets and the construction of data-driven web applications or Jupyter notebook interactions. It leverages DuckDB for efficient data processing and offers interoperability and extensibility for creating custom components, making it ideal for building scalable, interactive data visualizations.

- [Blog] The Paradox of Diffusion Distillation

  Sander Dieleman, a Research Scientist at DeepMind, examines strategies for distilling diffusion models, aiming to decrease sampling steps while retaining output quality. He delves into various distillation methods, such as progressive, guidance, and consistency models, highlighting their potential to simplify complex model processes. Dieleman's analysis navigates the challenges of optimizing sampling efficiency against model performance, offering insights into advancements in generative model optimization.

- [Paper] LLM Agents can Autonomously Hack Websites

  This study reveals the potential of large language models (LLMs) as autonomous agents capable of executing complex cyberattacks, including blind database schema extraction and SQL injections, without prior knowledge of vulnerabilities. Focusing on GPT-4, the research distinguishes its advanced tool use and context leveraging abilities from those of existing open-source models. The ability of GPT-4 to autonomously discover website vulnerabilities underscores the need for caution in the deployment of LLMs, highlighting significant cybersecurity implications.



- [Online Book] Spinning Up in Deep RL

  "Spinning Up in Deep RL" by OpenAI simplifies learning in deep reinforcement learning (RL) with tutorials, essential papers, and algorithm implementations. Targeting the gap in deep RL education, it prepares learners for AI safety contributions by demystifying RL concepts and algorithms. Emphasizing clarity and minimalism, the project aligns with OpenAI's goal of promoting AI safety and democratizing AI knowledge. Currently maintained to ensure ongoing accessibility, "Spinning Up" underscores OpenAI's dedication to the responsible development of AI technologies.

Latest Releases

- [Minor Release] Transformers 4.39.0

  HuggingFace Transformers 4.39.0 release adds several new models, including Cohere's Command-R for long context tasks, the improved multimodal LLaVA-NeXT v1.6, MusicGen Melody for music generation from text/audio, the position embedding-free vision model PvT-v2, and UDOP for document AI tasks. It expands the library's capabilities across modalities like vision, audio, and document processing.

- [Minor Release] Keras 3.1.0

  The 3.1.0 release of Keras introduces int8 inference support via model.quantize("int8"), a utility to set the backend, layers for mel spectrogram audio processing and image cropping, new random sampling ops, enabling int8 einsum operations, and improved axis handling across backends. Notable fixes include resolving issues with Functional model slicing, SpectralNormalization XLA compilation errors, and refactoring the axis logic.

Events

- [Conference] NVIDIA GTC 2024

  Nvidia's GTC 2024, held from March 18th to 21st in San Jose, spotlighted the latest AI advancements, with a major focus on generative AI across cybersecurity, conversational AI/NLP, computer vision, and more. Innovations in robotics driven by AI, edge computing for

autonomous systems, and other emerging technologies were also showcased. Recorded content remains available, capturing Nvidia's cutting-edge AI solutions unveiled over those insightful days.

Thank you for your engagement. We eagerly anticipate sharing further advancements in AI with you.