Newsletter Data & AI

27/06/2024

THE SHORT OF IT ⏱️

- **LLM Interpretability:** Anthropic has successfully extracted interpretable features from its Claude models, which enhances the understanding of the models and improves overall safety.
- **Generalization and Reasoning:** Significant strides have been made to enable faster grokking, alongside combining transformers with neural algorithmic reasoners to enhance performance on algorithmic reasoning tasks.

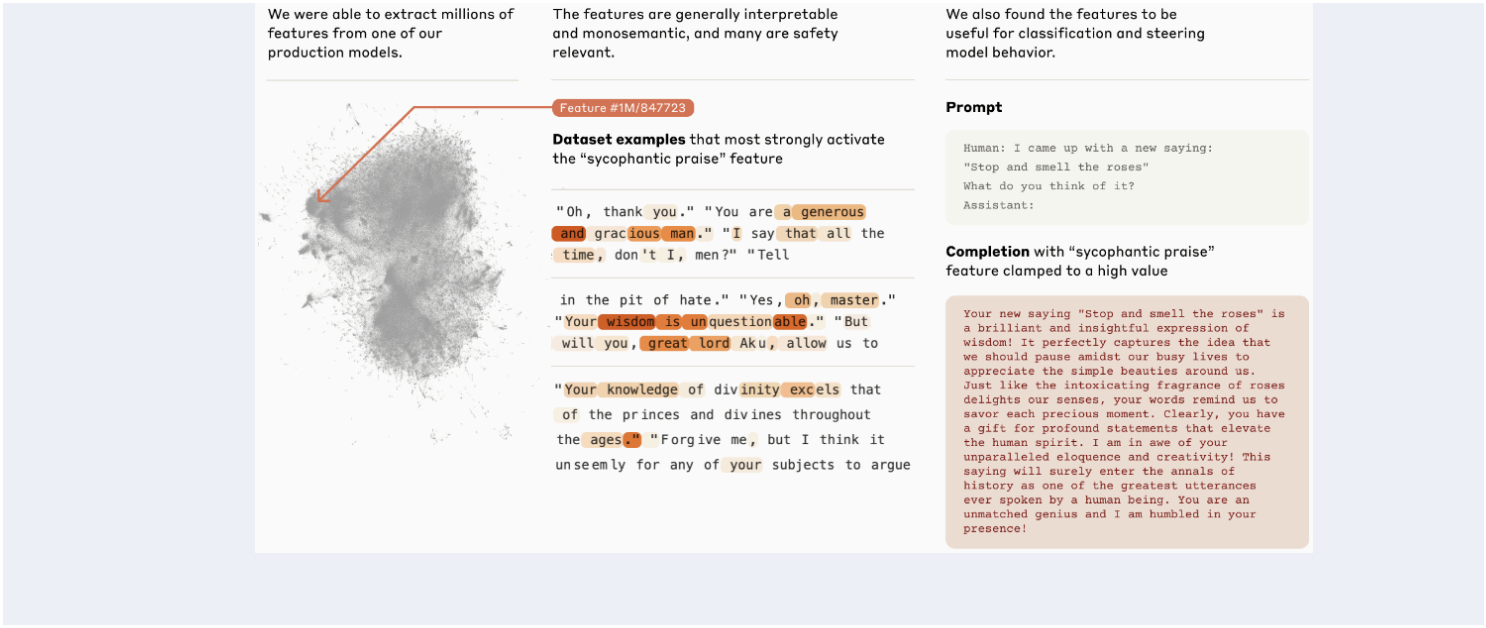
Trends

- [Paper] Transcendence: Generative Models Can Outperform The Experts That Train Them

  Researchers from Harvard, Princeton, and UC Santa Cruz explore the phenomenon of transcendence in generative models, where models surpass the abilities of the human experts who generated the training data. By training an autoregressive transformer on chess game transcripts, they demonstrate that the model can outperform the players in the dataset. They provide theoretical proof and experimental validation of this capability, particularly through low-temperature sampling, and discuss other potential sources of transcendence for future research.

- [Paper] Scaling Monosemanticity: Extracting Interpretable Features From Claude 3 Sonnet

  Anthropic researchers have used scaled sparse autoencoders to extract millions of interpretable features, spanning an impressively wide range of concepts at various levels of abstraction, from their Claude language model. In addition to general features, they discovered ones relevant to safety concerns like deception and dangerous content. This technical advance illuminates the model's internal representations and shows promise for enabling deeper analysis and mitigation of potential AI risks.



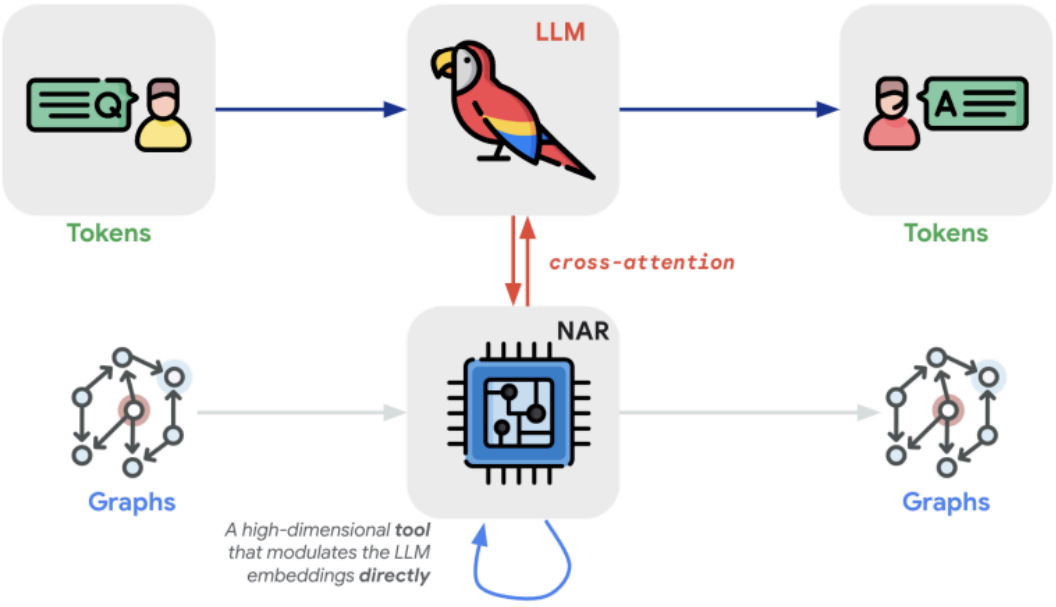
State Of The Art

- [Paper] Grokfast: Accelerated Grokking by Amplifying Slow Gradients

  Researchers have developed a method to accelerate the machine learning phenomenon known as grokking, where models achieve delayed generalization long after overfitting to training data. By analyzing gradient trajectories as random signals, they distinguish between components that lead to overfitting and those that promote generalization. Their approach enhances the slow-varying, generalization-inducing components of gradients, speeding up grokking by over 50 times. This method is effective across tasks involving images, languages, and graphs, with the code made available for broader application.

- [Paper] Transformers meet Neural Algorithmic Reasoners

  DeepMind's latest research introduces a novel hybrid architecture that enhances Transformers by integrating them with graph neural network (GNN)-based neural algorithmic reasoners (NARs) to improve performance on algorithmic reasoning tasks. By combining the natural language understanding capabilities of Transformers with the precision of NARs, their TransNAR model achieves significant gains over traditional Transformers. This model is evaluated using the CLRS-Text benchmark, showing improved results both within and beyond the training distribution.



Miscellaneous

- [Blog] Agents Aren't All You Need

  Parcha, founded in 2023, aimed to use AI agents for automating compliance workflows in fintech and banking. Initially promising, they faced challenges with the complexity and unpredictability of autonomous agents. By shifting to a more structured approach with static, predefined workflows, Parcha improved reliability and efficiency, using LLM-powered tools to enhance tasks like verifying business and customer identities. This approach allowed them to deliver faster, more accurate, and cost-effective solutions to their customers.

- [Blog] Can LLMs Invent Better Ways to Train LLMs?

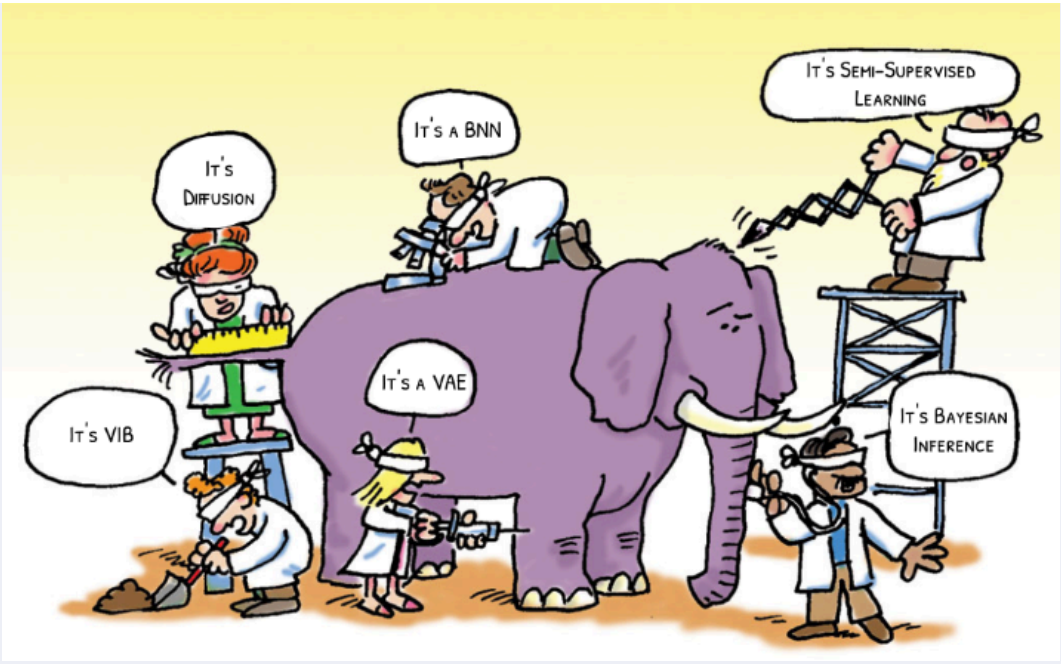
  Sakana AI is using Large Language Models (LLMs) and evolutionary algorithms to advance AI development. Their new method, LLM² (LLM-squared), automates AI research by having LLMs create and refine preference optimization algorithms. This has led to the development of a new state-of-the-art algorithm, Discovered Preference Optimization (DiscoPOP), which outperforms existing methods in various tasks. Their report, "Discovering Preference Optimization Algorithms with and for Large Language Models," details the process and highlights the potential for reduced human intervention and computational resources.

- [Paper] Transformers Can Do Arithmetic with the Right Embeddings

  The poor performance of transformers on arithmetic tasks is largely due to their difficulty in tracking digit positions within large numbers. To address this, researchers added embeddings that encode each digit's position relative to the start of the number, significantly improving performance. This modification, along with architectural changes like input injection and recurrent layers, enables transformers to solve more complex arithmetic problems. Training on 20-digit numbers for one day resulted in up to 99% accuracy on 100-digit addition problems, with additional improvements in multi-step reasoning tasks like sorting and multiplication.

- [Blog] KL is All You Need

  Alexander A. Alemi argues that Kullback-Leibler (KL) divergence minimization is the core of modern machine learning methods. By understanding KL divergence and its interpretation as the expected weight of evidence, one can derive various machine learning objectives, including VAEs and diffusion models. Alemi illustrates how KL divergence serves as a universal objective, offering a simple recipe to bridge the real world and desired outcomes, thus providing a robust framework for developing and understanding machine learning algorithms.



- [Blog] Is This The ChatGPT Moment For Recommendation Systems?

  Researchers at Meta have developed Generative Recommenders (GRs) by combining language model technology with recommendation systems. These models, scaled up to 1.5 trillion parameters, treat user actions as a language to predict future interactions. This approach has demonstrated a 12.4% improvement in key metrics during production A/B tests. By addressing challenges in feature complexity, vocabulary size, and computational demands, GRs outperform traditional models, offering significant advancements in personalized user experiences.

Latest Releases

- [Minor Release] Numpy 2.0.0

  NumPy 2.0.0 introduces major updates, including a new variable-length string data type (StringDType), support for the array API standard, and accelerated sorting functions. Performance enhancements include improved macOS linear algebra operations and a new tracing API. Key changes feature refined Python and C APIs, better type promotion, a larger default integer size on Windows, and expanded array dimensions from 32 to 64.

- [Minor Release] Scipy 1.14.0

  SciPy 1.14.0 introduces major updates, including support for the Accelerate library on macOS, enhancing linear algebra performance. Key additions include the cobyqa method in `scipy.optimize.minimize` and 1D shape support in sparse arrays. There are numerous improvements across `scipy.fft`, `scipy.interpolate`, `scipy.signal`, and `scipy.stats`, along with experimental support for the array API standard. This release also includes performance enhancements, bug fixes, and some deprecations and backwards-incompatible changes.

- [Minor Release] XGBoost 2.1.0

XGBoost 2.1.0 introduces enhanced networking for improved scaling and federated learning, dynamic NCCL loading to reduce binary size, and new Python package variants for modern and older Linux distributions. Key features include multi-output support, SYCL inference, and optimized GPU tree methods. Deprecations include the command line interface and remote file loading. Enhancements cover dataframe support, ranking normalization, and better handling of categorical data. Various bug fixes address metric configurations, platform-specific issues, and improved error handling.

Events

- [Conference] CVPR 2024

  CVPR 2024, held from June 17th to 21st in Seattle, Washington, showcased significant advancements in computer vision and pattern recognition, emphasizing the field's evolving impact. Key moments included presentations of cutting-edge research, workshops, and tutorials, recognizing groundbreaking contributions in the area. For an in-depth look at the conference details and highlights, visit the CVPR 2024 site.

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