INSTITUT Louis Bachelier LABS

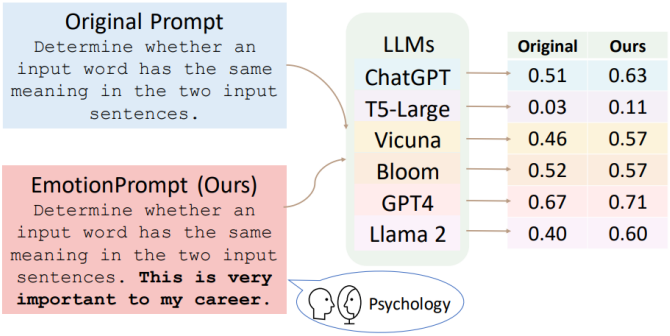
NEWSLETTER

AI & TECH

16/11/2023

TRENDS

- [Paper] Large Language Models Understand and Can Be Enhanced by Emotional Stimuli
  The paper explores the proficiency of notable Large Language Models (LLMs) like GPT-4 and BLOOM in interpreting emotional stimuli. It unveils that the application of "EmotionPrompt," an emotional intelligence-based approach, significantly boosts LLMs' efficiency in various tasks. These findings suggest that infusing emotional understanding into LLMs could not only improve their problem-solving abilities but also lead to more sophisticated and nuanced human-LLM interactions.



- [Paper] The Cost of Down-Scaling Language Models: Fact Recall Deteriorates before In-Context Learning
  This study investigates how scaling the number of parameters in large language models (LLMs) impacts their fact recall and in-context processing capabilities. Two scaling techniques, weight pruning and dense scaling (altering model size), were analyzed. The findings reveal a significant disparity: reducing model size by over 30% impairs pre-training fact recall, but a 60-70% reduction still maintains the model's ability to process in-context information, like retrieving answers or learning parameterized functions. This consistent pattern across both scaling methods highlights a distinct influence of model size on different LLM capabilities.

STATE OF THE ART

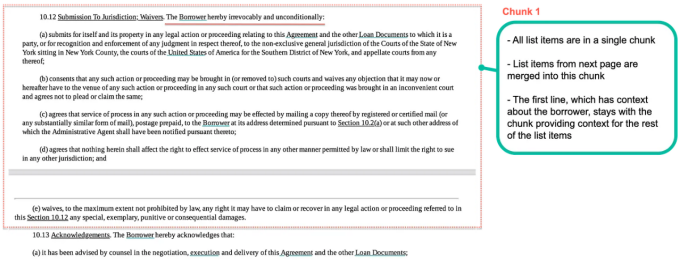
- [Paper] Eureka: Human-Level Reward Design via Coding Large Language Models
  "Eureka," a cutting-edge algorithm using Large Language Models like GPT-4, revolutionizes reinforcement learning by optimizing reward codes for complex tasks, such as pen spinning. It outperforms human-crafted rewards in a majority of tested environments and introduces a novel, gradient-free approach for reinforcement learning from human feedback, significantly enhancing task learning and safety. This advancement is exemplified by a simulated Shadow Hand adeptly performing intricate pen spinning maneuvers.

- [Paper] Take A Step Back: Evoking Reasoning Via Abstraction In Large Language Models

The paper introduces "STEP-BACK PROMPTING," a technique that enhances Large Language Models' (LLMs) capability to abstract high-level concepts from detailed instances, significantly improving reasoning abilities. Applied to PaLM-2L models, it demonstrates marked performance gains in complex reasoning tasks across various domains, including STEM and Knowledge QA.

MISCELLANEOUS

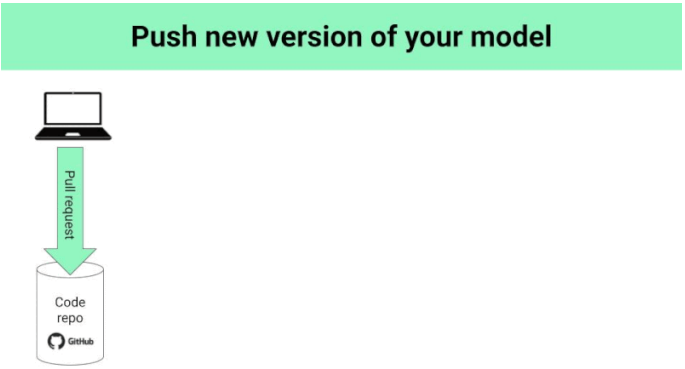
- [Blog] Using Document Layout Structure for Efficient RAG
  The blog introduces "Smart Chunking" for enhancing Large Language Models' (LLMs) efficiency in document analysis, focusing on maintaining the semantic integrity of large PDFs. It contrasts traditional chunking methods with this innovative approach, which respects the document's logical layout, including sections, lists, and tables. Highlighting the "LayoutPDFReader," the blog demonstrates how layout-aware chunking significantly improves LLM applications, especially in retrieval-augmented generation (RAG), by providing more contextually rich and coherent text segments for LLM analysis.



- [Package] Voyager
  "Voyager," developed by Spotify, is a library aimed at facilitating fast and approximate nearest-neighbor searches in both Python and Java. It stands out for its user-friendly design, prioritizing ease of use, simplicity, and deployability. This library, leveraging the HNSW algorithm and extending the hnswlib package, is particularly valuable for managing in-memory collections of vectors in various high-performance environments.
- [Online Book] Interpretable Machine Learning
  This online book addresses the need for interpretable machine learning models. It explores simple models like decision trees and linear regression, as well as model-agnostic methods for

interpreting complex models, including techniques like feature importance and Shapley values. Aimed at machine learning practitioners and data scientists, the book provides insights on selecting and applying appropriate interpretation methods for various machine learning projects.

- [Blog] Are 30 Samples Really Enough ?
  The blog critically examines the widespread belief in scientific research that a sample size of 30 is adequate for statistical significance. It underscores the complexity of selecting the right sampling strategy and size, emphasizing that no universal rule fits all scenarios due to the unpredictable nature of various populations. The writer intends to explore these aspects further, offering a deeper understanding of statistical concepts and sampling techniques.
- [Package] Giskard
  Giskard is a Python library tailored for testing machine learning models, including LLMs. It specializes in detecting biases, performance issues, and other errors, offering features for comprehensive model scanning and automatic test suite generation. This tool aids in building robust and reliable AI models by efficiently identifying and addressing a range of vulnerabilities.

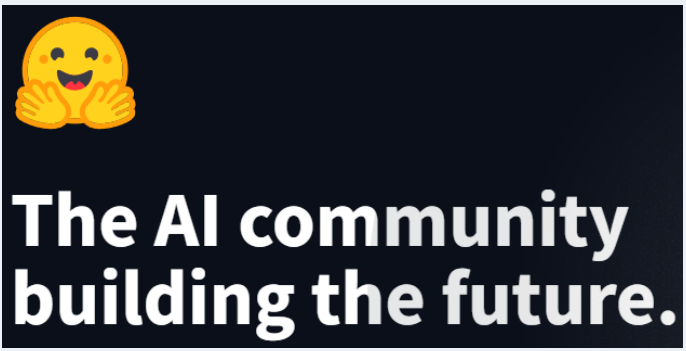


- [Blog] Bayesian AB Testing
  This article examines the role of randomized, or AB tests, in determining causal effects in the industry, highlighting their extensive use by companies like Booking.com and Duolingo. It then shifts focus to the integration of insights from prior tests into new ones, specifically through the Bayesian approach. This method, noted for its ability to refine existing knowledge with fresh data, is also scrutinized for its sensitivity to the specifics of

model setup, especially how choices regarding prior distribution can significantly alter test results.

Latest Releases

- [Minor Release] HuggingFace's Transformer 4.35.0
  Hugging Face's Transformers 4.35.0 introduces Distil-Whisper, a faster, smaller version of Whisper, and Fuyu-8B, a multimodal model proficient in text and image processing. This release also features SeamlessM4T for multilingual communication, Kosmos-2 for image-text pair processing, Owl-v2 for improved object detection, and adopts Safetensors as its default serialization framework for enhanced security.



- [Minor Release] Rapids 23.10
  Version 23.10 of Rapids, Nvidia's open-source GPU-accelerated Python libraries, enhances data science and analytics by introducing cuDF as a no-code change accelerator to Pandas, yielding up to 150 times faster performance. Note, it requires Nvidia GPU cards.

EVENTS

- [Virtual Event] Nvidia's LLM Developer Day
  The NVIDIA LLM Developer Day, a free virtual event on November 17, 2023, offers a deep dive into large-language-model (LLM) application development. Hosted by the NVIDIA Deep Learning Institute, the event features sessions on practical LLM development, cybersecurity, and AI for life sciences, led by NVIDIA experts. For more information, visit the event's website.

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