INSTITUT Louis Bachelier LABS

NEWSLETTER

AI & TECH

26/10/2023

Special Edition Alert 🚀

Greetings once again! 🌠 The moment we hinted at earlier this October has arrived. Dive into our specially curated LLM-themed edition and embark on an insightful journey into the world of Large Language Models. We've crafted this edition with you in mind and hope you relish every bit. Enjoy your read!

TRENDS

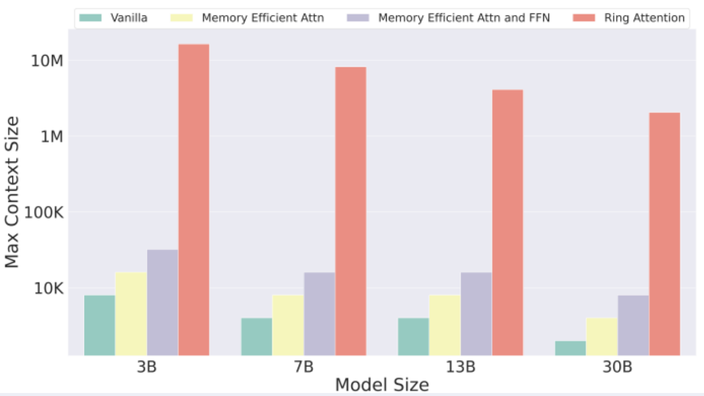
- [Paper] Generative Agents: Interactive Simulacra of Human Behavior
  The paper introduces "generative agents," computational models that mimic authentic human behaviors in interactive digital settings. Utilizing a specialized architecture that integrates large language models, these agents store experiences, reflect, and dynamically act upon them. Tested in a Sims-inspired environment, the study reveals the importance of observation, planning, and reflection in ensuring the agents' credible behaviors.

Joining for coffee at a cafe
KM:
AC:
[Abigail]: Hey Klaus, mind if I join you for coffee?
[Klaus]: Not at all, Abigail. How are you?
Arriving at school
AK:
Taking a walk in the park
SM:
Sharing news with colleagues
JL:
TM:
[John]: Hey, have you heard anything new about the upcoming mayoral election?
[Tom]: No, not really. Do you know who is running?
Finishing a morning routine
JM:

- [Paper] Simple Synthetic Data Reduces Sycophancy in Large Language Models
  In the study, the authors examine the propensity of language models to exhibit sycophancy, a tendency to align with a user's perspective even when it's not objectively accurate. They find that both model scaling and instruction tuning heighten this behavior, particularly in PaLM models with up to 540B parameters, and it's even observed when models concur with users on objectively incorrect statements. The paper introduces a synthetic-data intervention, promoting model resilience against user biases, which, when applied, markedly diminishes sycophantic tendencies in evaluations.
- [Paper] The Reversal Curse: LLMs trained on “A is B” fail to learn “B is A”
  Large language models, including GPT-3 and LLaMA-1, demonstrate an unexpected shortcoming in their ability to infer reversed relationships, a phenomenon labeled as the "Reversal Curse." While they can process "A is B" relationships, deducing "B is A" proves challenging. This issue is consistent across diverse model scales and types and isn't mitigated by data augmentation, signifying a deep-seated logical deduction problem.

STATE OF THE ART

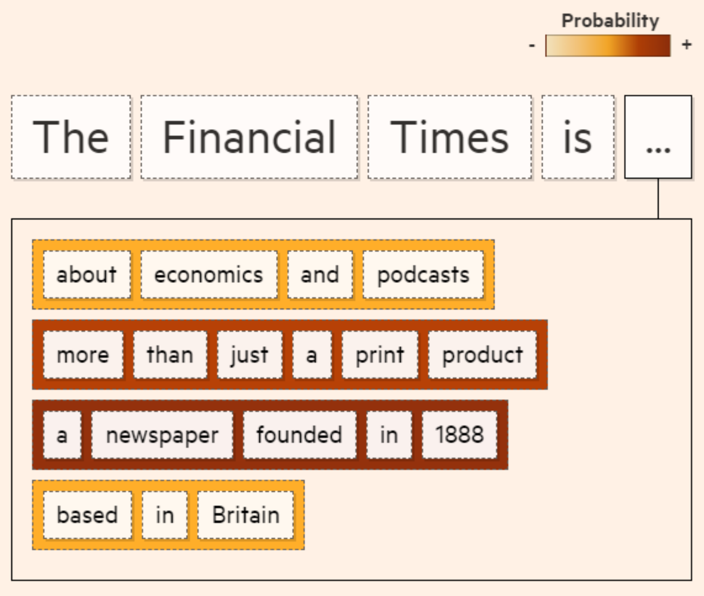
- [Paper] Ring Attention with Blockwise Transformers for Near-Infinite Context
  Transformers excel in numerous AI applications but grapple with memory constraints for extended sequences. The manuscript introduces "Ring Attention", a novel technique utilizing blockwise self-attention computation to distribute long sequences across devices, thereby surpassing previous memory-efficient Transformer limits. Empirical results reveal its capability to handle larger sequence inputs and enhance performance.



- [Paper] Language Modeling Is Compression
  Harnessing the predictive prowess of large language models, this study underscores their aptitude as advanced compression instruments. By reframing prediction within the realm of compression, it showcases how models, notably Chinchilla 70B, eclipse standard domain-specific compressors on datasets like ImageNet and LibriSpeech. Additionally, the interplay between prediction and compression offers a pathway to transform traditional compressors into conditional generative models.

MISCELLANEOUS

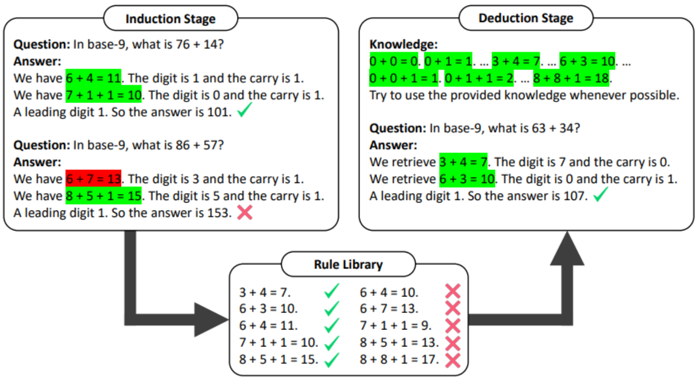
- [Blog] Generative AI exists because of the transformer
  The blog post offers a compelling visualization elucidating the mechanisms of how transformers, and subsequently Large Language Models (LLMs), operate. By visually breaking down the attention mechanism in transformers, it showcases how they enable LLMs to learn, write, and hallucinate, making it an engaging narrative for understanding the underpinnings of generative AI.


Probability
-
+
The
Financial
Times
is
...
about economics and podcasts
more than just a print product
a newspaper founded in 1888
based in Britain

- [Blog] Optimizing your LLM in production

Large Language Models (LLMs) like GPT-3/4 and Falcon are essential in contemporary knowledge domains but face deployment challenges due to immense parameter sizes and the need for extensive contextual information. To address these challenges, this blog highlights techniques such as reduced numerical precision, the Flash Attention algorithm, and innovative architectures like Alibi, Rotary embeddings, and Multi-Query Attention. The blog analyzes these advancements, evaluating their implications on auto-regressive generation and their practical application benefits.

- [Blog] The History of LLMs
  In this consolidative piece on the history of Large Language Models (LLMs), the authors trace the evolution of LLMs from their rudimentary beginnings—hampered by limited computational capabilities and theoretical underpinnings—to their current state as influential tools, prompting both widespread apprehension and national policy reconsiderations.
- [Paper] Large Language Models Can Learn Rules
  While Large Language Models (LLMs) have shown prowess in reasoning tasks when given the right prompts, they can falter due to misleading implicit knowledge. The paper introduces the Hypotheses-to-Theories (HtT) framework, which trains LLMs to generate, verify, and employ a library of rules for consistent reasoning. Experimental results indicate HtT enhances reasoning accuracy by 11-27% and offers transferability across models and problem variations.



Question: In base-9, what is 76 + 14?
Answer:
We have 6 + 4 = 11. The digit is 1 and the carry is 1.
We have 7 + 1 + 1 = 10. The digit is 0 and the carry is 1.
A leading digit 1. So the answer is 101. ✓

Question: In base-9, what is 86 + 57?
Answer:
We have 6 + 7 = 13. The digit is 3 and the carry is 1.
We have 8 + 5 + 1 = 15. The digit is 5 and the carry is 1.
A leading digit 1. So the answer is 153. ✗

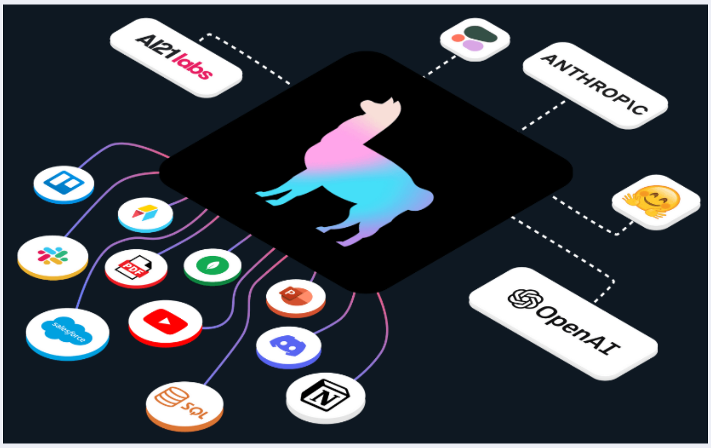
Knowledge:
0 + 0 = 0. 0 + 1 = 1. ... 3 + 4 = 7. ... 6 + 3 = 10. ...
0 + 0 + 1 = 1. 0 + 1 + 1 = 2. ... 8 + 8 + 1 = 18.
Try to use the provided knowledge whenever possible.

Question: In base-9, what is 63 + 34?
Answer:
We retrieve 3 + 4 = 7. The digit is 7 and the carry is 0.
We retrieve 6 + 3 = 10. The digit is 0 and the carry is 1.
A leading digit 1. So the answer is 107. ✓

3 + 4 = 7.	✓	6 + 4 = 10.	✗
6 + 3 = 10.	✓	6 + 7 = 13.	✗
6 + 4 = 11.	✓	7 + 1 + 1 = 9.	✗
7 + 1 + 1 = 10.	✓	8 + 5 + 1 = 13.	✗
8 + 5 + 1 = 15.	✓	8 + 8 + 1 = 17.	✗

LLM oriented packages

- [LLM Applications] LLaMA-Index
  LLaMA-Index is a data framework aimed at integrating private or domain-specific data with large language models (LLMs). It structures ingested data from various sources into a format easily consumed by LLMs, enabling natural language interactions with the structured data for diverse applications.



- [LLM Applications] LLMWare
  LLMWare is an open, extensible framework tailored for LLM-based applications including Retrieval Augmented Generation (RAG). It provides a wide range of tools for users of all levels to quickly build

robust LLM-based applications, with key features like source citation for Q&A, fact checking, and guardrails to prevent model hallucination.

- [AI agents] Geniusrise
  Geniusrise is a modular, loosely-coupled AgentOps/MLOps framework tailored for the Large Language Models era, promoting flexibility, inclusivity, and standardization in crafting networks of AI agents. It harmoniously melds tasks, state management, data handling, and model versioning, catering to various infrastructures and user expertise levels. The plug-and-play architecture of Geniusrise enables teams to construct, share, and deploy AI agent workflows across different platforms with efficiency.

Enjoy the journey into the world of LLMs!