Are LLMs Prescient? A Continuous Evaluation using Daily News as the Oracle

Hui Dai, Ryan Teehan, & Mengye Ren
New York University

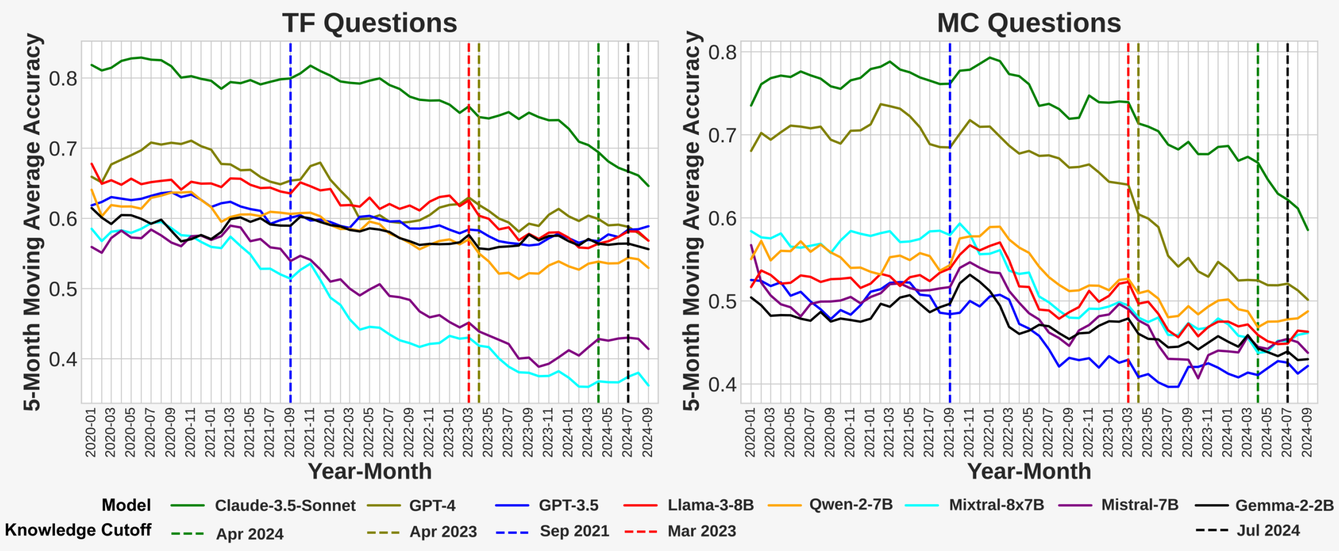
Project Website & Dataset

Motivation: Daily-updated Benchmark for LLMs' Continuous Evaluation

Daily Oracle: Automatically generated QA dataset from daily news to assess LLMs' temporal generalization and forecasting abilities.

Existing Benchmark	Our Benchmark
Static	**Daily-updated**
• **Quickly become outdated** due to new models and new training data	• Daily news provides a natural setting for continuous evaluation of LLMs
• Without temporal dimension	• **Can access how the future prediction capabilities of LLMs evolve over time**
• **Cannot do continuous evaluation**	

How Do Models Perform on Daily Oracle Overtime?



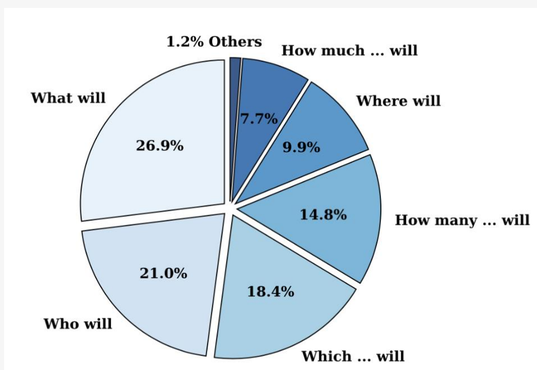
- Decline 20.14% on TF questions, and 23.26% on MC questions!
- Gradual decline in the recent past & rapid decline in the near future
- Consistent performance decline after September 2021

Daily Oracle Dataset

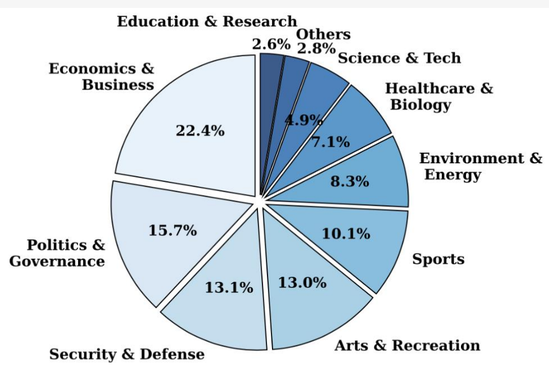
Question Type	Total
True/False	16,082
Multiple Choice	13,906
Total	29,988

- Jan. 2020 ~ Sept. 2024*
- On average ~17.3 QA pairs per day

* Daily Oracle is updated daily. For the current analysis, a subset of the data is used



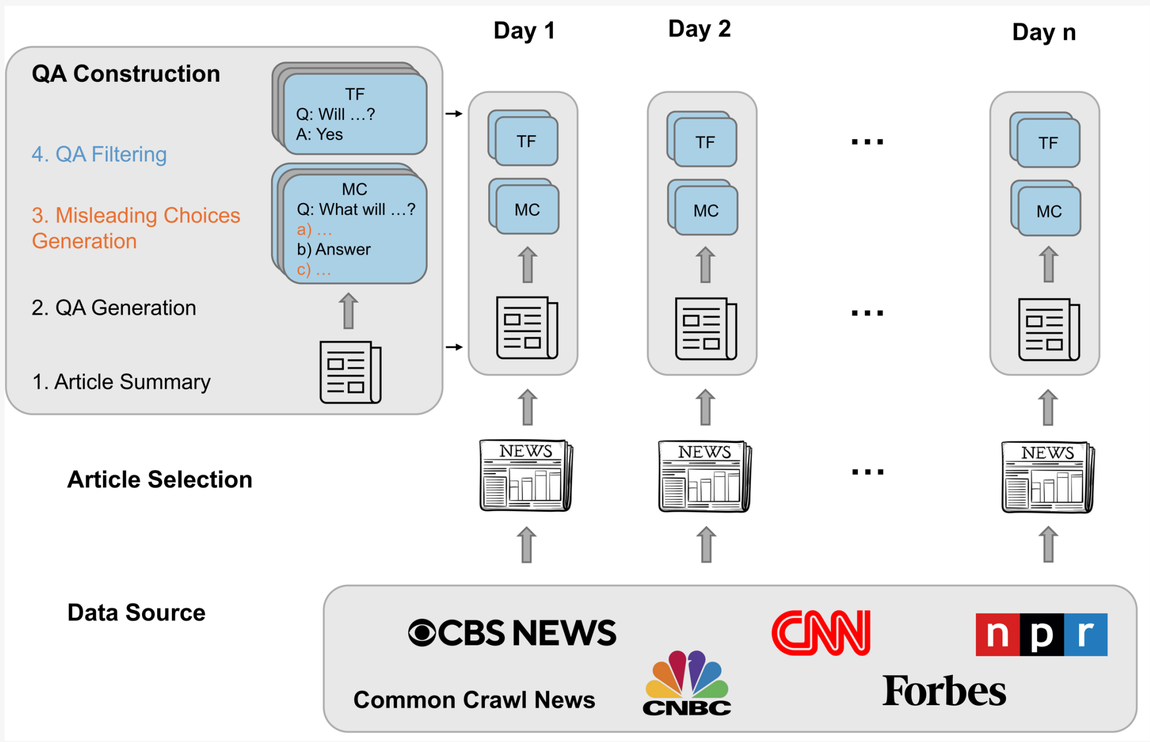
(a) MC question types for Daily Oracle.



(b) Category Breakdown for Daily Oracle.

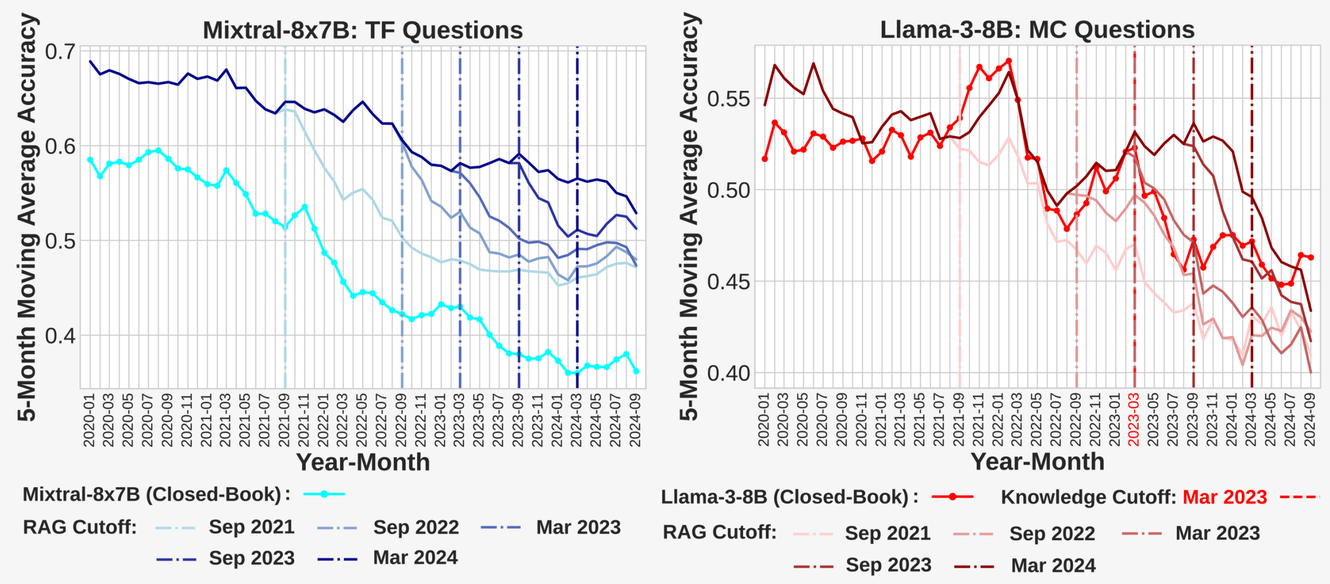
Type		Category	Question and Answer
TF		Politics & Governance	Will the prosecution's key witness in the New York hush money trial in April 2024 be someone other than Michael Cohen? –**No.**
TF		Politics & Governance	Will the House Energy and Commerce Committee vote unanimously to advance a bill that could potentially ban TikTok if ByteDance does not sell the app by March 2024? –**Yes.**
MC	What	Science & Tech	What will be the starting price range for the Google Pixel 8a as of May 2024? A.$599–$649 B. $199–$249 C. $750–$800, D. $499–$559. –**D.**
MC	Who	Sports	Who will go on the injured list before the New York Mets' game on May 29, 2024? A. Pete Alonso B. Edwin Diaz C. Jeff McNeil D. Francisco Lindor –**B.**
MC	Which	Arts & Recreation	By May 2024, on which streaming service will "The First Omen" become available for subscribers? A. Disney+, B. Hulu, C. Amazon Prime Video, D. Netflix –**B.**
MC	How many	Science & Tech	How many U.S. states will the path of totality cross during the total solar eclipse on April 8, as reported by February 2024? A. 15 B. 10 C. 20 D. 6 –**A.**
MC	Where	Healthcare & Biology	Where will the second known U.S. case of bird flu in a human be reported by March 2024? A. California, B. Texas, C. New York, D. Florida –**B.**
MC	How much	Economics & Business	How much will Apple, Inc. (AAPL) be up year-to-date by the end of June 2024? A. Up 149.5% B. Just over 19% C. 9.7%. D. 27%. –**C.**

Dataset Construction



Can RAG Save the Declining?

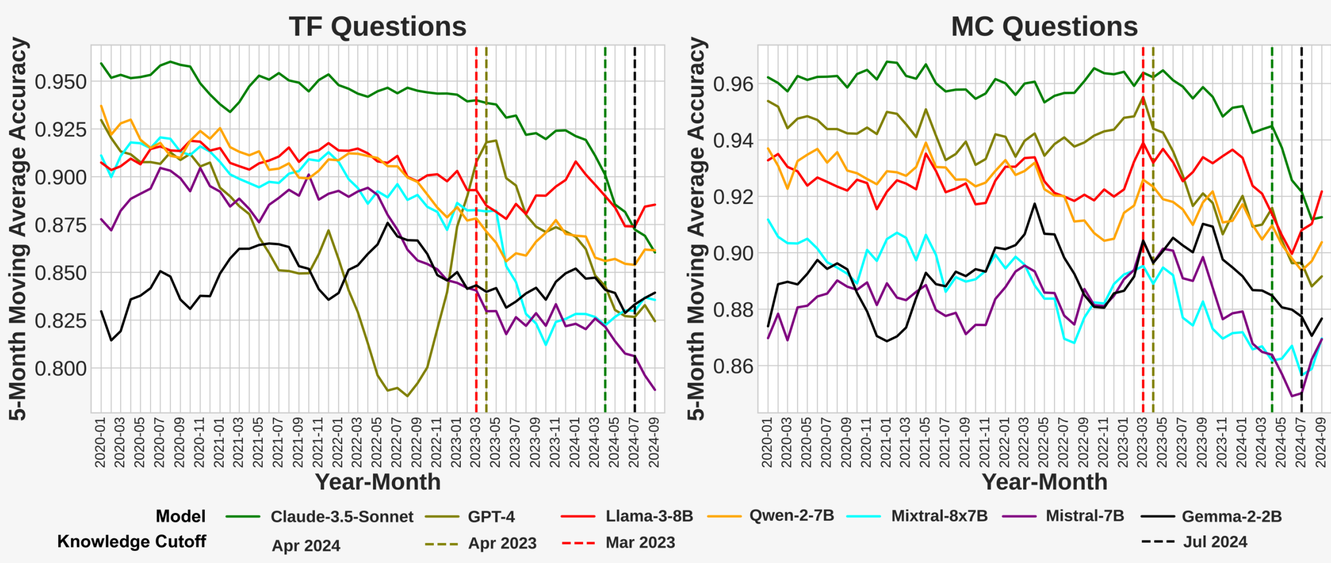
Q: How does access to news articles up to different RAG cutoffs influence LLM performance?



- More updated information may help with the performance
- However, **the overall performance decline pattern persists**

What if the Gold Articles are Given?

Setting: We give models direct access to the gold article from which the question is generated, framing the task as reading comprehension



- **~90%** accuracy demonstrates **answerability**
- The **downward trend still exists**
- Likely due to **out-of-date representations** in LLMs
- **In-context knowledge updates are insufficient, and continuous model updates are necessary**

Conclusion

- We introduce Daily Oracle, a continuously updated QA benchmark leveraging daily news to evaluate the temporal generalization and future prediction capabilities of LLMs
- Model performance degrades over time, even with RAG/gold articles
- Our findings underscore **the necessity for ongoing model updates with more current information**
- **We call for continual learning methodologies to bridge this performance gap**