agentic
learning
ai lab

ICML
International Conference
On Machine Learning

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

Hui Dai, Ryan Teehan, Mengye Ren
https://agenticlearning.ai/daily-oracle

Project Website

TL;DR: Daily Oracle is an automatically generated QA dataset from daily news to quantify how outdated the LLMs are.

Key Takeaways

- With Daily Oracle, we can **see how LLMs' performance degrades** over time.
- **This decline persists even with RAG/access to gold articles.**
- The decline comes from the **missing future knowledge** and **outdated language representation.**
- Underscore the **necessity for ongoing model updates** with more current information.

Problems

- LLMs have knowledge cutoff dates
- Most of the benchmarks are static
- But the world is changing!

➡

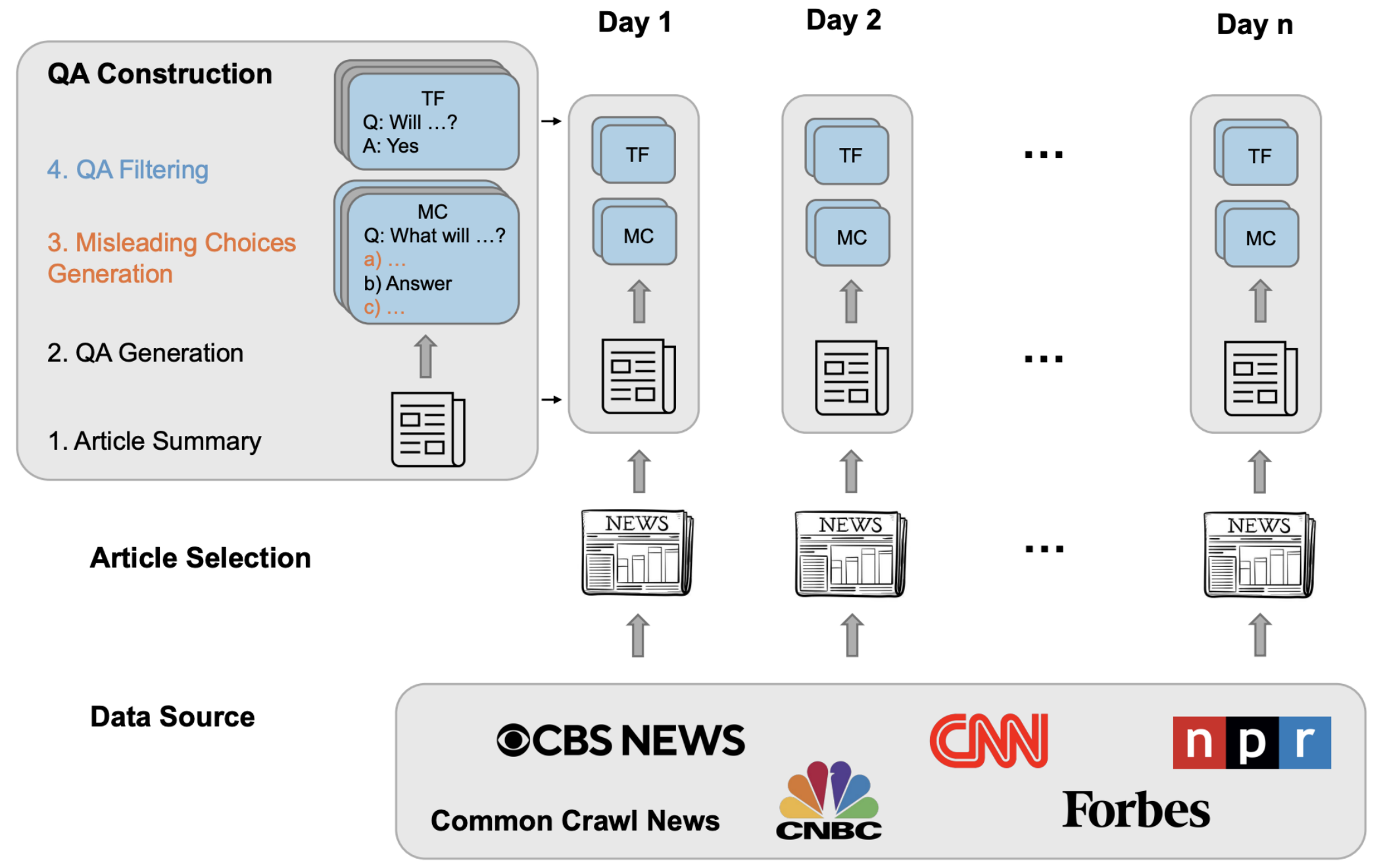
Goal

- A daily updated dataset for continuous evaluation
- Assess LLMs' temporal generalization and forecasting abilities

Daily Oracle Dataset

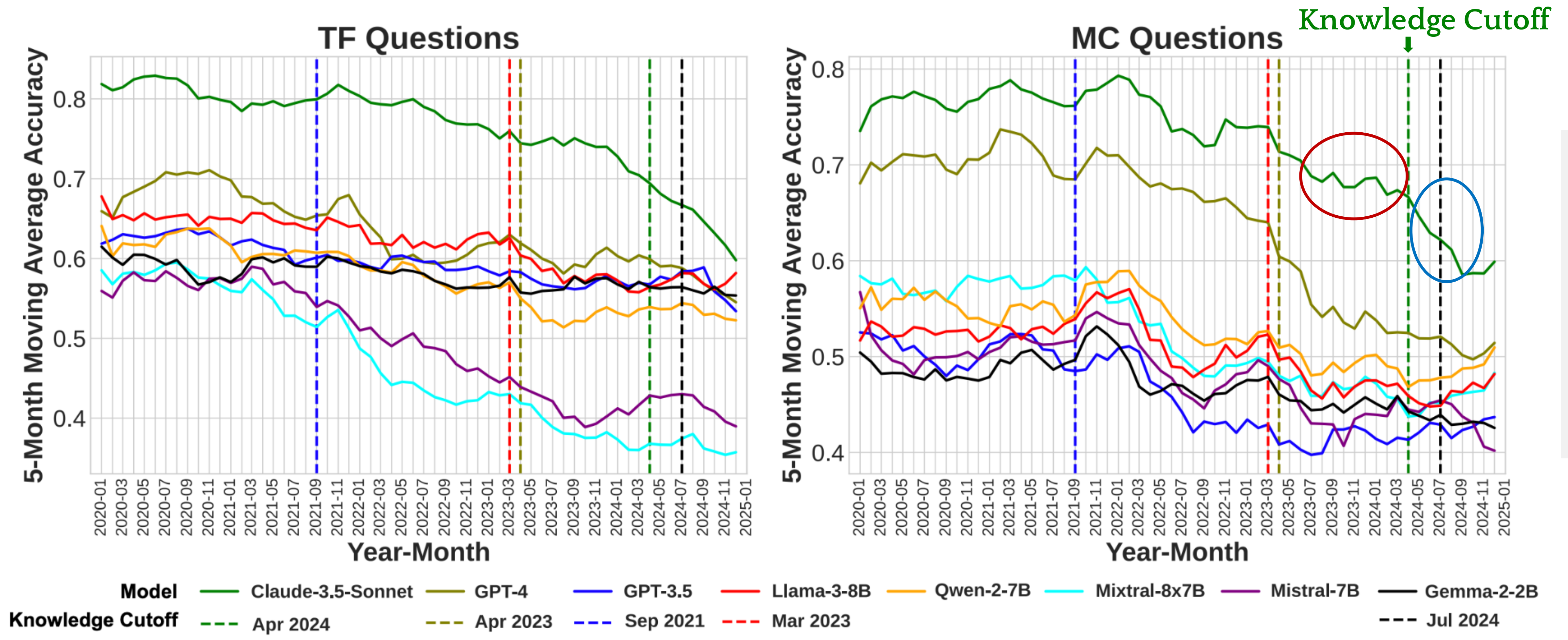
- **Question Types:** True/False & Multiple Choice
- **Time Span:** 2020.01 – present
- **Size:** ~17.2 QA pairs per day (31,510 in total)*

* with the current version until 2024.12



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.**

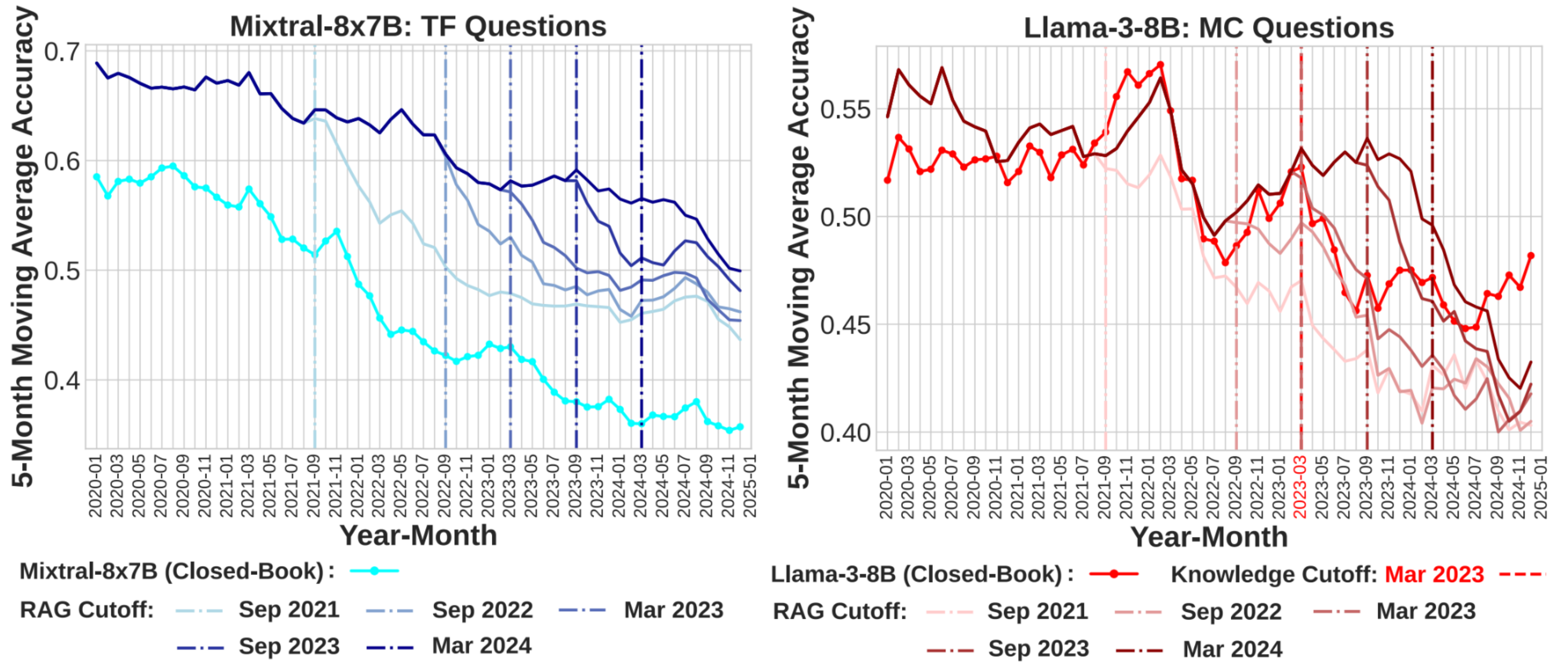
We reveal a clear performance degradation pattern in LLMs' forecasting accuracy over time



1. On average, models decline 21.55% on TF questions and 11.33% on MC questions.
2. Gradual decline in the recent past & rapid decline in the near future
3. Consistent performance decline after September 2021

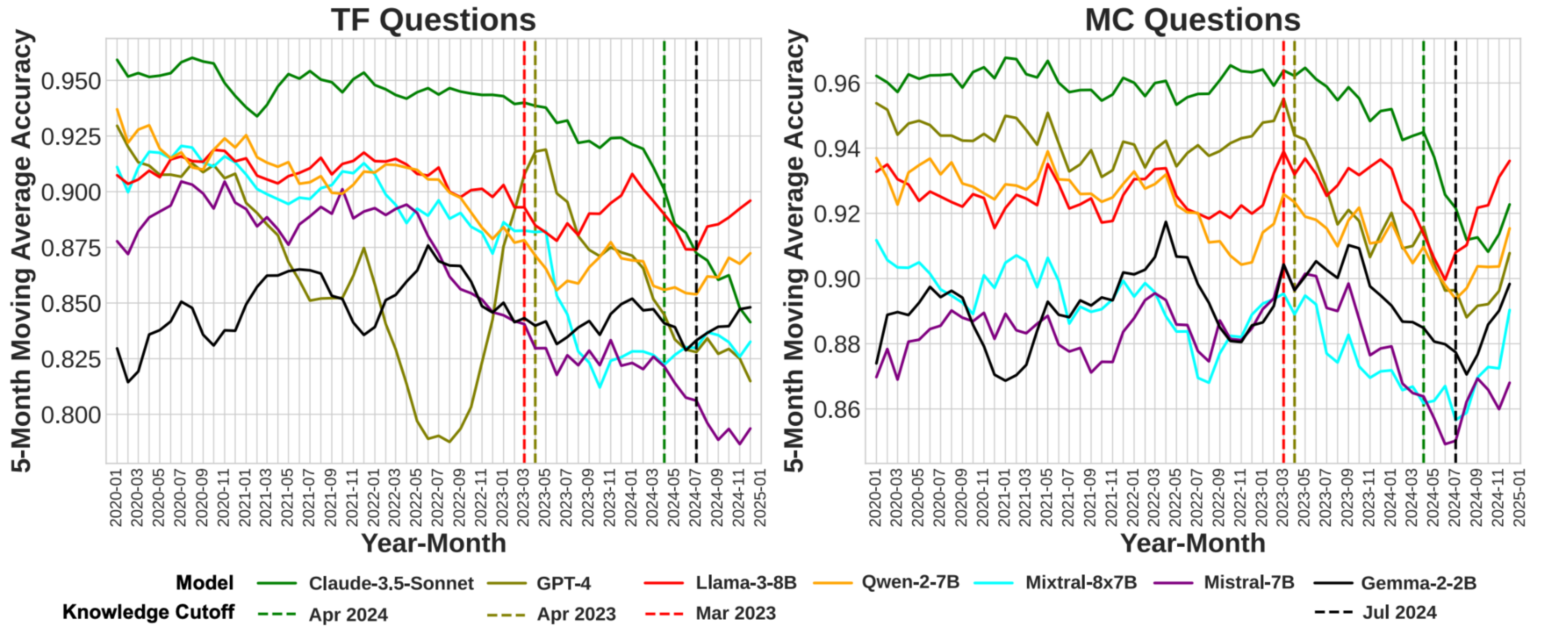
RAG cannot save the decline

- Allow models to access news articles up to different RAG cutoffs.
- Setup: BM25 retriever + top-5 articles
- With RAG, the overall performance decline pattern still exists!



Surprisingly, decline persists even with gold articles!

- ~90% accuracy demonstrates answerability.
- In-context knowledge updates are insufficient -> Representations are outdated
- Continuous model update is necessary.


TF Questions
MC Questions
5-Month Moving Average Accuracy
Year-Month
Model
Claude-3.5-Sonnet
GPT-4
Llama-3-8B
Qwen-2-7B
Mixtral-8x7B
Mistral-7B
Gemma-2-2B
Knowledge Cutoff
Apr 2024
Apr 2023
Mar 2023
Jul 2024