maintain these abilities.

Notably, the average YoY accuracy declines provide further insight. Before the knowledge cutoff, the average YoY decline across all models was relatively moderate. However, post-knowledge cutoff, we observe steeper declines in many models, with GPT-4 showing the most drastic drop in MC performance, declining by 18.54%, compared to just 4.23% before the cutoff. This contrast highlights that while LLMs manage to retain a baseline of past knowledge with small degradation, their ability to forecast future events deteriorates much more rapidly as they move beyond their training data, struggling with temporal generalization.

Among different models, Claude-3.5-Sonnet (Anthropic, 2024) significantly outperforms all others, while GPT-4 excels in MC questions but its performance in TF is not as remarkable as in MC. GPT-3.5, Qwen-2-7B (Yang et al., 2024) and Llama-3-8B (Dubey et al., 2024) show smaller temporal declines than GPT-4 in both TF and MC questions. Interestingly, Mistral-7B (Jiang et al., 2023) and Mixtral-8x7B (Jiang et al., 2024) show the most pronounced drops in TF accuracy, with scores falling below the random baseline 50% due to increased answer refusals, as shown in Figure 9. Gemma-2-2B (Team et al., 2024) exhibits the most consistent performance with the smallest average YoY decline, likely due to its more recent knowledge cutoff date.

Results for the Constrained Open-Book Setting. In Figure 4, we present the results of the constrained open-book setting, with Mixtral-8x7B on TF questions and Llama-3-8B on MC questions across different RAG cutoff dates.[8] For Mixtral-8x7B, as the RAG cutoff dates extend to closer to the resolution dates, we observe a clear improvement in performance, indicating the model benefits from increasingly updated information retrieval. However, there are noticeable performance drops immediately after each RAG cutoff date when compared to providing information up to the day before the resolution date. This highlights the importance of keeping up-to-date information for optimal RAG performance. Interestingly, RAG does not uniformly enhance performance. Llama-3-8B may perform worse than the closed-book setting when the RAG cutoff is prior to the knowledge cutoff dates, suggesting outdated information may negatively impact performance. Conversely, for more recent RAG cutoff dates that extend beyond the knowledge cutoff, significant performance improvements are observed (as illustrated by the curves with cutoffs in September 2023 and March 2024). Notably, across all different RAG cutoffs, the overall performance decline pattern persists, likely due to outdated internal representations and the model's inherent knowledge limitations.

Results for the Gold Article Setting. Figure 5 shows that when given access to the gold articles from which the questions are generated, LLM performance can approach around 90%, demonstrating the answerability of Daily Oracle.[9] However, most of the models still show declining trends. This is noteworthy because, ideally, LLMs are expected to achieve consistent accuracy regardless of the article's publication date when answers are directly accessible. However, the outdated representations hinder their ability to consistently generate correct answers, even in a reading comprehension setting.

4.3. Discussion

LLMs' Performance Degradation Pattern Over Time. We observe LLMs' performance evolution patterns in Figure 3: (1) *Gradual Decline in the Recent Past:* In the months before the knowledge cutoff date, which we call the *recent past*, we observe a gradual decline in model performance, as seen in Llama-3-8B, GPT-4, and Claude-3.5-Sonnet, likely due to a lack of representation of recent news in the training data. (2) *Rapid Decline in the Near Future:* In the *near future*, which we define as the months following a model's knowledge cutoff date, sharp performance drops are observed in several models in MC questions. For instance, the decline in Claude-3.5-Sonnet and GPT-4 accelerates soon after their knowledge cutoffs. Most of the models, however, do not lose all the predictive power at once, as evidenced by the further decline into the farther future.

We explore this further by analyzing the slope of accuracy as a function of time. In Figure 6, we show how the slope changes as we fit a regression to an increasingly larger window of data, until we reach the full set of accuracies. Specifically, using the 5-month moving average of each model's accuracy on MC questions (visualized in Figure 3), we start by fitting a linear regression line on the first 10 months of data. We then add an additional month and compute a new regression on the larger window, repeating until we reach the final month, and applying an exponential decay weighting to past data to reduce the influence of distant observations. With this, we can analyze how the slope of our regression line changes as each month is added to the data. The slope in each case is negative after the cutoff data and for Claude-3.5-Sonnet, GPT-4, and Llama-3-8B, the slope eventually or immediately becomes more negative than it was at any point preceding the cutoff. Both Claude-3.5-Sonnet and Llama-3-8B have a crossover from positive to negative slope in late summer 2022, July and August, respectively, while GPT-4's seems to occur slightly earlier, in March of 2022. For GPT-3.5, GPT-4, and Llama-3-8B, the slope becomes

[8] Refer to Appendix B.4 for results of other models in the constrained open-book setting.

[9] Results for GPT-3.5 are provided and discussed in Appendix B.3, as this older model performs relatively poorly and including it on the same scale would obscure the trends of other models.

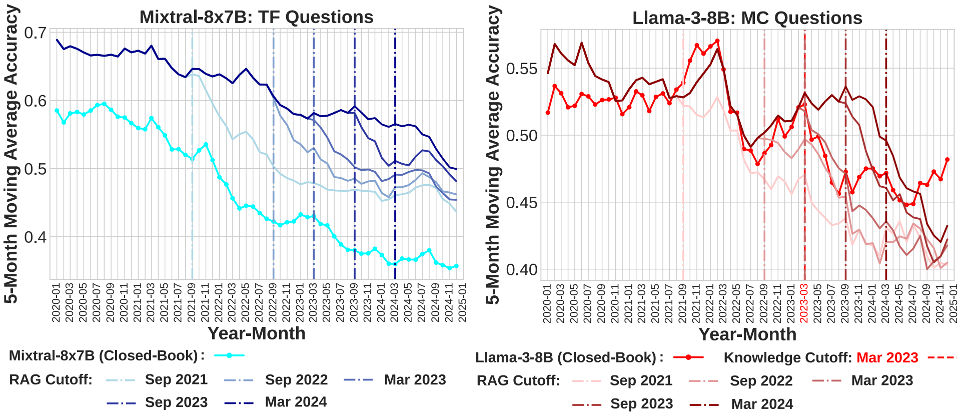

Figure 4. Results for the constrained open-book setting, evaluating Mixtral-8x7B on TF questions and Llama-3-8B on MC questions with different RAG cutoff dates.

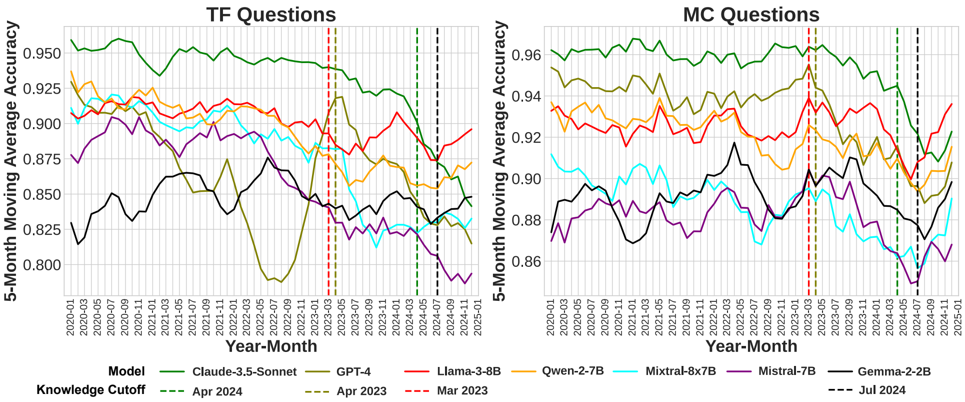

Figure 5. Results for the gold article setting. Most of the models struggle with temporal generalization, even when provided with gold articles containing the answers.

increasingly negative not long after the knowledge cutoff, giving evidence for a rapid decline in the *near future*. For example, Llama-3-8B's slope is around -0.06% per month near its cutoff in March 2023, but declines by more than 2 times to approximately -0.13% per month by the end of 2023. Likewise, the period preceding the cutoff shows a milder decline, with models like GPT-3.5 and Llama-3-8B exhibiting slightly negative but consistent slopes (approximately -0.01% and -0.06% per month respectively, over the four months leading up to the cutoff). This suggests a gradual decline in the *recent past*.

Need for Continuous pre-training. The overall decline trend may come from two sources, the missing knowledge of future and a lack of up-to-date language representation. The absence of relevant future information can lead to two outcomes: either the model makes uninformed or incorrect predictions, or, in some cases, more likely to refuse to answer altogether. We observe this latter behavior notably in Mistral-7B and Mixtral-8x7B, where refusal rates are significantly higher compared to other models, as shown in Figure 9(b).[10] The lack of knowledge can be partially recovered with information retrieval, as seen in the constrained open-book and gold article settings. For instance, Figures 9, 10, and 11 show that Mixtral-8x7B's refusal rate drops from 14–28% in closed-book to 3–15% with open-book retrieval, and further to 0.5–4.2% with gold articles. However, accuracy still declines over time. Notably, the gold article setting provides an "upper bound" of open-book retrieval. The remaining performance drop despite full access to relevant information suggests that the models' internal representations are outdated. This indicates continuous pre-training of LLMs (Jang et al., 2022; Jin et al., 2022; Ke et al., 2022a;b; Yıldız et al., 2024) is still needed in the context of news event forecasting.

TF & MC Comparison. All models except for Claude-3.5-Sonnet struggle with TF questions, where the degrada-

[10] See Appendix B.2 for more discussion of the refusal behavior.

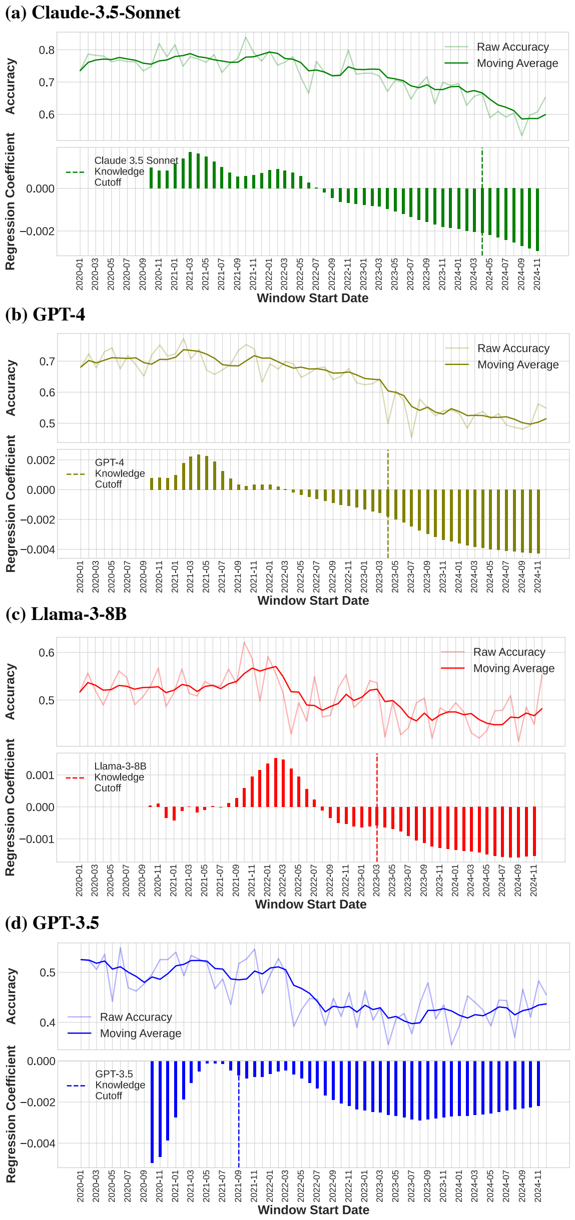

Figure 6. Coefficients for regressing accuracy on the MC questions against time, as the number of months grows. Using an initial window of 10 months, we progressively add data for additional months to our regression and plot the coefficient (slope) for the regression of accuracy against time. For our regression, we use the moving average of accuracy and apply exponentially decaying weights to older months (i.e., given a window of k months, we weight x_t with λ^{k-t}; in this case $\lambda = 0.995$).

tion trends towards the random baseline accuracy of 50%, indicating that predicting if a future event will happen or not can be sometimes challenging for LLMs. In contrast, on MC questions, models tend to perform much better than the random baseline at 25%. There are two potential reasons that can explain the disparity. First, TF questions can be considered more open-ended than MC because the "No" answer contains other possible open-ended outcomes. Second, since the distractor choices are created by an LLM, they may not be as likely to happen as the true answer.

Consistent Performance Decline After September 2021. Interestingly, Figure 3 reveals a higher rate of performance decline around September 2021, which is the knowledge cutoff date of GPT-3.5, across all models, particularly for MC questions. In contrast, performance remains relatively stable prior to this date. We hypothesize that this trend arises because the period up to September 2021 may be overrepresented in many pre-training corpora (Raffel et al., 2020; Gao et al., 2020; Kobayashi, 2018; Gokaslan & Cohen, 2019; Zhu et al., 2015; Rae et al., 2020; Tiedemann, 2016; Saxton et al., 2019), compared to more recent periods. Another potential cause of this imbalance is an increasing number of websites restricting access to web crawlers after the rise of ChatGPT (Longpre et al., 2024).

Limitations. On the data generation side, the generated questions as well as the distractor answers could contain biases from an outdated LLM, making the benchmark less reliable in the long run unless we upgrade the models. Additionally, generating questions from news articles can introduce bias by focusing only on events that have definitively occurred, overlooking potential events that never occur and thus never appear in the news. On the evaluation side, our paper proposes the continuous evaluation benchmark but at the time of the writing there isn't a long enough time horizon on each model, especially after the cutoff dates, for a thorough analysis. Ideally, we would like to analyze the relation between the effect of knowledge and RAG cutoff dates but the trend seems to be weak within the time horizon available.

5. Conclusion and Future Work

We introduce Daily Oracle, a continuously updated QA benchmark leveraging daily news to evaluate the temporal generalization and future prediction capabilities of LLMs. Our experiments reveal that while LLMs maintain a degree of predictive power over future events, their prediction accuracy exhibits a significant smooth decline over time. Although RAG mitigates the effect of outdated knowledge, a strong and noticeable decline remains. Our findings in the gold article setting further emphasize the importance of disentangling missing knowledge from the lack of up-to-date representations. In the future, alongside maintaining Daily Oracle, we plan to incorporate a broader range of models and explore how continuous pre-training and efficient adaptation can address the performance degradation challenges presented in our work.