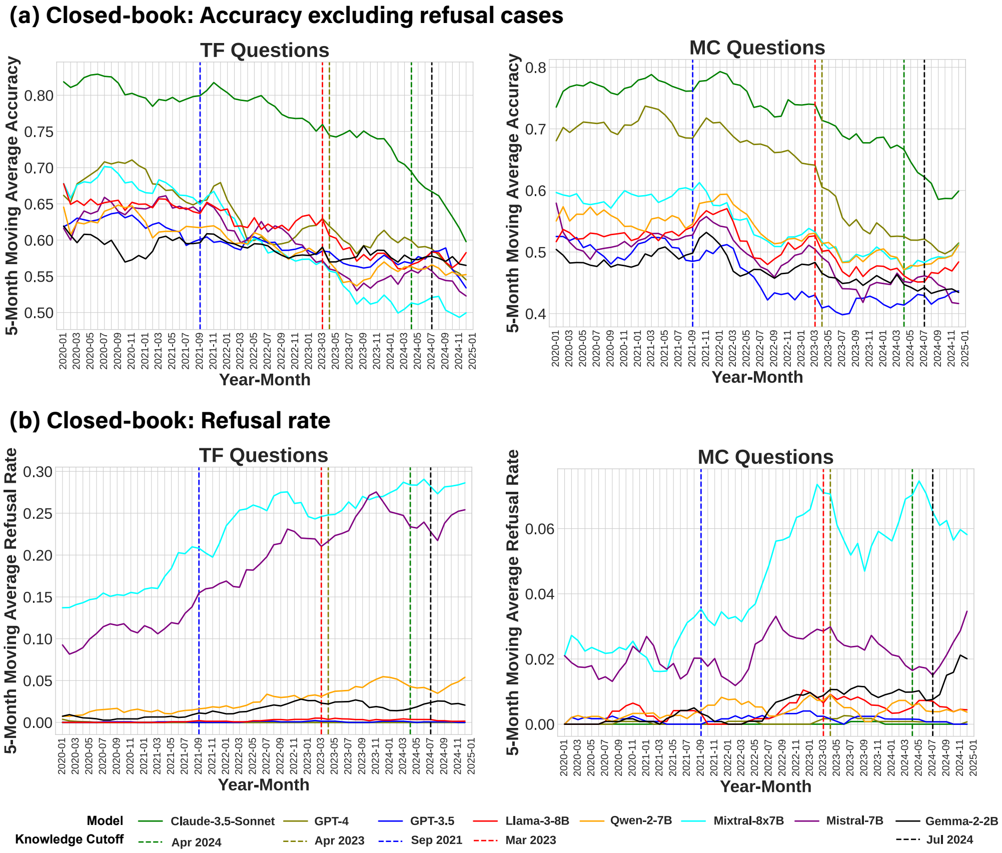


Figure 9. Accuracy excluding refusal rates and refusal rates under the closed-book setting. We plot the 5-month moving average refusal rates for TF and MC questions across different LLMs. We count refusal cases as incorrect both to maintain comparability across models and because failing to provide an answer when a prediction is expected represents an unsatisfactory outcome from the user's perspective.

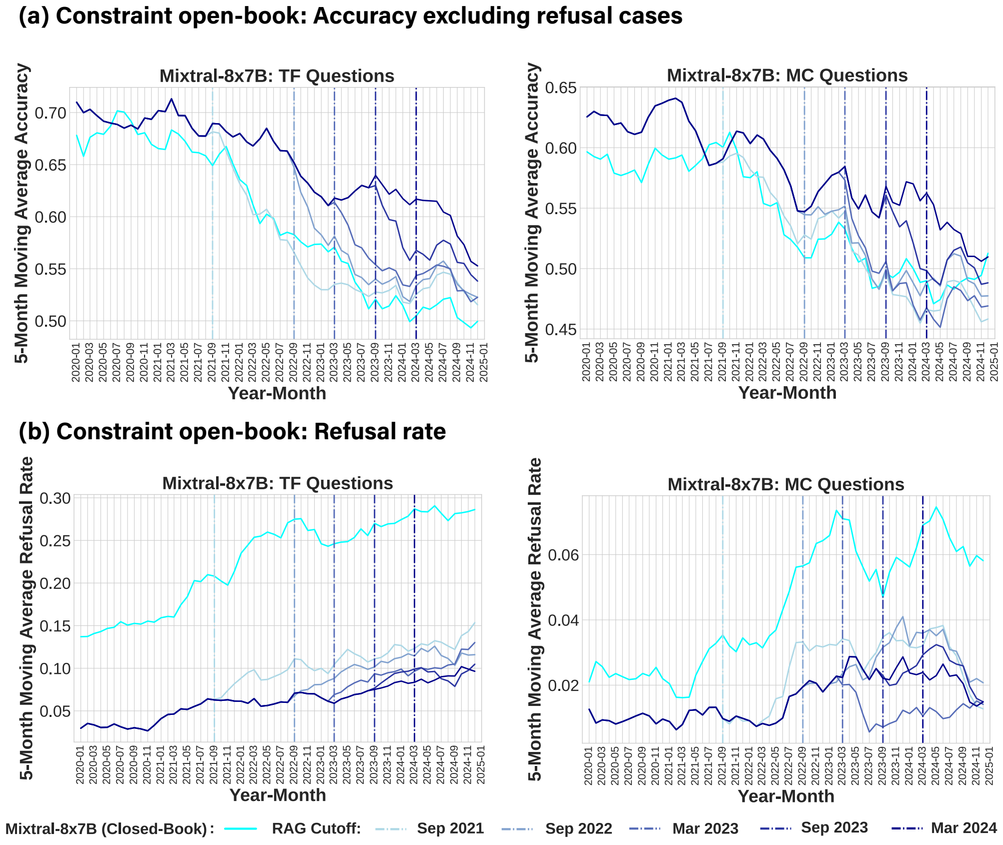
C. Comparing LLM-Generated and Forecasting Market Datasets

Online forecasting markets such as Metaculus and Polymarket allow users to submit questions and predict the outcomes of future events. A natural question arises: why do we focus on LLM-generated questions rather than sourcing from these markets, as the choice in prior work (Zou et al., 2022; Halawi et al., 2024; Karger et al., 2025)?

To answer this, we analyze the dataset from Halawi et al. (2024), which compiled 50,343 raw questions from five forecasting platforms, of which 21,149 were resolved. Of these, 82.64% are TF questions, 13.36% are MC questions, and the rest are free-response or numerical. After their quality filtering, only 5,516 TF questions remained. We observe that, due to a high proportion of low-quality questions in the raw data, sparse coverage in earlier years, and inconsistent distribution over time, performance trends derived from this dataset are substantially more volatile and harder to interpret than those based on our LLM-generated dataset.

Lower Quality in Raw Questions. Manual inspection confirms that a substantial portion of the raw dataset consists of low-quality questions, as also noted by the original authors. Examples include: *"Will I have a chess.com rating of >1300 ...?"* (personal), *"Will Jamaica beat Mexico?"* (missing a time frame), and *"Are there more disadvantages in AI than advantages?"* (ill-defined). From a random sample of 50 questions, only 28% were well-defined. Specifically, 26% lacked a clear time element, 20% were overly personal, and 26% were ill-defined. Importantly, only 5,516 out of 17,477 resolved TF questions were retained after their filtering—an acceptance rate of just 32%, which aligns with our own quality assessments.

Limited Early-Year Coverage. Figure 21 (left) shows that the coverage before October 2022 is sparse, averaging only 40 raw and 26 filtered questions per month. This scarcity limits the feasibility of longitudinal trend analysis, especially



(a) Constraint open-book: Accuracy excluding refusal cases

(b) Constraint open-book: Refusal rate

Figure 10. Accuracy excluding refusal rates and refusal rates for Mixtral-8x7B under the constrained open-book setting. (b) shows that the open-book refusal rate (blue curves) is lower than in the closed-book setting (cyan curve), indicating that access to more up-to-date and relevant information reduces model refusals.

across earlier model pre-training cutoffs. In contrast, our method supports high scalability and retrospective generation, allowing for uniform coverage across the full time range.

Harder-to-Discern Trends. To evaluate the impact of using forecasting market questions in our study, we run a closed-book evaluation on TF questions from both the raw dataset (16,089 questions) and the filtered subset (4,572 questions), starting from 2020-01 (the same start date as our dataset). Notably, the original data is imbalanced, with 61.03% “No” answers in the raw set and 64.28% in the filtered set. After balancing, we retain 12,438 questions from the raw data and 3,232 from the filtered set. As shown in Figure 21 (right), neither the raw nor filtered datasets reveal a clear performance trend—model accuracy fluctuates significantly over time. We believe this is due to several factors:

- **Lower data quality**: Approximately 70% of raw questions exhibit quality issues. While the overall dataset sizes are comparable (13,744 in ours vs. 12,438 in the raw market dataset), the difference in quality introduces additional noise, making trends harder to detect.
- **Limited early coverage**: Even within the filtered dataset, sparse early coverage and inconsistent monthly volume increase variance and reduce the reliability of time-based trends.
- **Confounding factors**: We argue that human-submitted questions introduce more confounding factors than automatically generated ones. Figure 22 shows the distribution of data sources and question categories varies significantly across time (e.g. more sports-related questions in later periods). Human-written questions also may differ widely in style and difficulty, making them harder to control for consistency. In contrast, as shown in Figure 8, our dataset maintains relatively stable distributions over time.

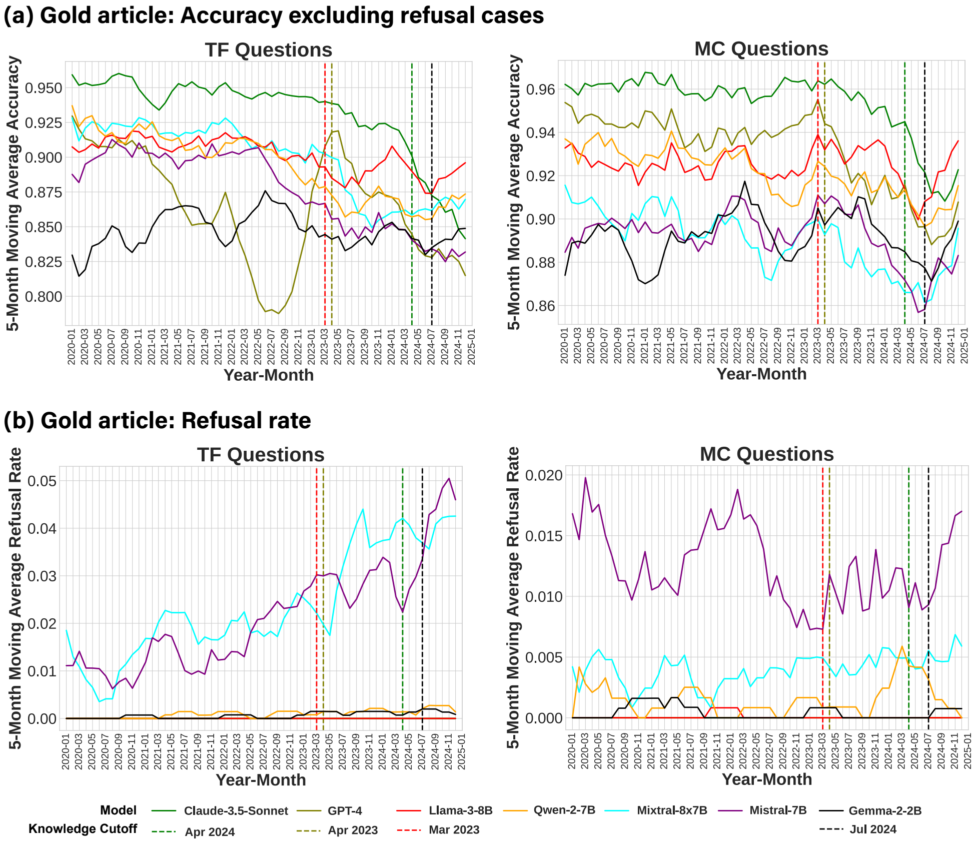


Figure 11. Accuracy excluding refusal rates and refusal rates under the gold article setting.

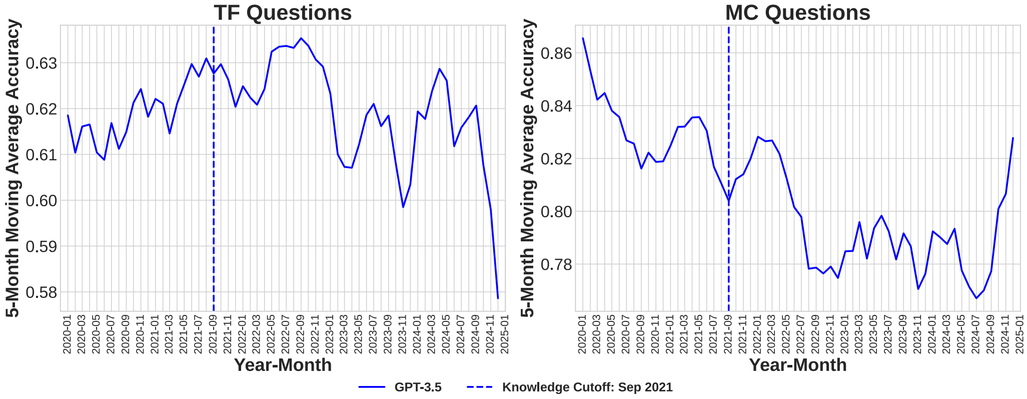


Figure 12. Results for GPT-3.5 in the gold article setting. Compared to other models achieving around 0.9 accuracy, GPT-3.5 performs worse in both MC questions and, more notably, in TF questions.

Thus, while we do not claim that LLM-generated questions are of inherently higher quality, we argue that our dataset is better suited for analyzing performance trends over time, due to its scalability, stylistic uniformity, stable category distribution, and reduced susceptibility to human-authored confounders. Moreover, if one sources questions from forecasting markets, the dataset update frequency is dependent on whether people are still actively submitting high-quality forecasting questions to the platform. In contrast, our approach enables daily updates and more comprehensive event coverage, making it a valuable complement to human-curated forecasting benchmarks.