E Details on Optimization of Our System

E.1 Hyperparameter Sweep

Throughout the experiment, we set the retrieval date to be the midpoint between a question's open and resolve date. At this time point, the crowd aggregates achieve 0.160 Brier score, averaged over all questions in our validation set.

All the hyperparameter sweeps below evaluate all questions in the validation set.

Search query prompt. We sweep over 6 candidate prompts for generating search queries. The top 2 prompts lead to retrieved articles that have average relevance rating of 3.08 and 3.09, while other prompts below 3.04. Among all articles with rating at least 4, the average rating is 4.37 and 4.38 via the top 2 search query prompts, which is also the highest among all candidates.

Summarization prompt. We sweep over 5 candidate prompts for summarization and evaluate the resulting Brier scores. The best summarization prompt gives a Brier score of 0.193 and the second gives 0.201. In this step of hyperparameter search, the ordering of the summaries, article count and reasoning prompt are randomly chosen for each question to avoid confounding.

Article ordering and count. In this step, we sweep over both orderings of articles (by recency or relevance), and over 5 candidate choices of $k \in [5, 10, 15, 20, 30]$. We run our full system on all questions in the validation set. Presenting 15 article summaries and ordering them by relevance gives the lowest Brier score of 0.177 on GPT-4-1106-Preview. Similar performance can be achieved by presenting 20 articles summaries.

Reasoning prompt. We hand-craft 15 prompts for eliciting forecasts and reasonings. The best prompt (Figure 15) achieves a Brier score of 0.167 on the validation set (while fixing the optimal hyperparameter choices found by the optimization stages above). Two other top prompts get 0.170 and 0.174. The best prompt is given in Figure 15.

Ensembling. We implement 5 ensembling methods, including mean, median, geometric mean, trimmed mean, and a variant of universal self-consistency (USC) (Chen et al., 2023). The last two approaches are defined as follows:

- For the trimmed mean, we assign uniform weights over the input forecasts, reduce the weight of the forecast furthest from the median by half, redistribute the half weight uniformly to the other forecasts, and finally output the weighted average. We remark that this is not a standard implementation of trimmed mean, and it is set this way since we only aggregate a small number (i.e., 6) of forecasts in our system.
- For USC, we present the (external) reasoning-prediction pairs to a final LM, which is then prompted to form an aggregated forecast. In this hyperparameter sweep, we use GPT-4-1106-Preview as the aggregator model.

Ensemble Method	Brier Score
Mean	0.1656
Median	0.1651
Geometric Mean	0.1655
Trimmed Mean	**0.1649**
USC (Chen et al., 2023)	0.1691
Baseline (No Ensemble)	0.1676
Human Crowd	0.1600

Table 14: **Brier scores across different ensembling methods on the validation set.** "Baseline" refers to the average Brier score of the base predictions (i.e., the inputs to ensembling).

We evaluate all these methods on the validation set by generating 6 base reasonings for ensembling, using our optimal system setup. Trimmed mean achieves the lowest Brier score; see Table 14 for the results. The USC method, in contrast, does not demonstrate improvement over the baseline.

E.2 News API Evaluations

We justify our choice of the news API's. To begin with, we experiment with 5 eligible APIs to news corpus that accept retrieval date ranges, which, for our purpose, must be specified to prevent leakage. In particular,

we assess Google News (accessed via Python open source package `gnews`), NewsCatcher, Newsdata.io, Aylien, and NewsAPI.org.

To assess the quality of their retrievals, we first take 24 unresolved forecasting questions. Next, we prompt GPT-4-1106-Preview to generate two search queries for each of these 24 questions, similar to the first stage of our retrieval system (Section 4). We use these queries to search for articles via all 5 APIs, restricting the retrieval range to the last 24 hours.

Finally, we prompt GPT-4 to rate the relevance of the articles with respect to the original questions. Higher scores indicate greater relevance. We compute the sum of scores of all retrieved articles for each API option. As a result, NewsCatcher and Google News achieve the highest scores of 35 and 39, respectively. The other three APIs, Newsdata.io, Aylien, and NewsAPI.org, score 16.5, 30.5, and 23.5.

E.3 Relevance Rating Approximations

We prompt GPT-3.5 Turbo to score the relevance of all retrieved articles (from Google News and NewsCatcher) with respect to the question. Our prompt is given by Figure 14, where the question's metadata along with an article text is provided to the model in context. The prompt asks the LM to rate the relevance of an article to the given question at the scale of 1–6, where 6 is considered "most relevant" and 1 "irrelevant". In our system, we filter out articles with ratings below 4.

Methods. Due to cost constraints, we cannot afford to evaluate relevance scores using the full article texts. We experimentally explore 3 cost-saving alternatives to approximate full text-based ratings:

(i) Title-only. We only give the article title to the model for relevance rating. Unfortunately, via manual inspection, we find the web scraper[5] sometimes fails to to load the full text of an article page, usually due to rendering error or paywall. In such cases, the article content may be incomplete or simply a error message, whereas the title is retrieved and appears relevant. Consequently, the model may be misled by the title. Therefore, we believe that this approach is unviable.

(ii) Title + first 250 words. We give the article title and the first 250 words to the model for rating.

(iii) Embedding similarity. We embed the article text and question text metadata and compute their cosine similarities. We threshold by the similarities to filter articles.

Experiment. We experiment with approach (ii) and (iii) above. For (ii), we prompt GPT-3.5-Turbo and Mixtral-8x7B-DPO for relevance rating with the same prompt template (Figure 14). For (iii), we use OpenAI's `text-embedding-3-large` as the embedding model.

To generate raw articles, we randomly sample 15 questions from the validation set, run our retrieval system up to relevant rating, and collect 169 articles in total.

We evaluate the relevance scores of the full texts via GPT-4 (considered as gold labels) and compare with the approximations described above. An article is said be relevant if its rating is at least 4 from the full text query via GPT-4. We compute the recall and precision of the approximate methods as follows.

- At a threshold of 4, title + first 250 words via GPT-3.5-Turbo gives recall of 0.73 and precision of 0.65.
- At a threshold of 3, title + first 250 words via Mixtral-8x7B-DPO gives recall of 0.70 and precision of 0.63.
- At a threshold of 0.48, the cosine similarity method gives recall of 0.73 and precision of 0.54.

In sum, fixing recall at above 70%, querying GPT-3.5-Turbo with title + first 250 words attains the highest precision. The average token length of the articles is 1087.6. Hence, the method of using first 250 words (or roughly 330 tokens) saves about 70% of the cost.

[5] https://pypi.org/project/newspaper4k/

F Example Outputs

We provide 5 qualitative example outputs from our fine-tuned GPT-4 model on the test set.

1. Rephrased and Expanded Question: Is it likely that Kevin McCarthy, the former House Speaker, will make a public announcement or have his resignation reported by credible sources, indicating his departure from the United States House of Representatives, before the end of November 30, 2023? This question takes into account the current political climate, McCarthy's recent loss of the Speaker position, his consideration of not seeking re-election, and the potential impact of his decision on the Republican Party and the 20th District of California. The resolution of this market hinges on whether McCarthy will officially declare his intention to vacate his seat within the specified timeframe.
2. Reasons why the answer might be no:
- McCarthy has not made a final decision and has expressed uncertainty about his future in Congress, which could mean he is still weighing his options and may not choose to resign within the given timeframe. (Strength: Strong)
- The articles suggest that while there is speculation and consideration of resignation, there is no definitive statement or action indicating an imminent resignation. (Strength: Strong)
- McCarthy may be waiting to see how the political landscape shifts or to make a decision after discussing with his family, which could extend beyond the November 30 deadline. (Strength: Moderate)
- Given his history of strong financial support and political influence, McCarthy may feel compelled to stay and influence the political process, especially with the upcoming 2024 elections. (Strength: Moderate)
3. Reasons why the answer might be yes:
- There are reports and hints from McCarthy himself that he is considering resignation and may want to "get the hell out," which could indicate a potential for an early departure. (Strength: Moderate)
- The pressure from political opponents and the recent loss of the Speaker position may contribute to a decision to resign, especially if he feels his influence is waning. (Strength: Moderate)
- The possibility of a special election being called could provide an opportunity for a controlled transition of power, which might be appealing to McCarthy. (Strength: Weak)
4. Aggregate Considerations:
While there is speculation and some suggestive comments from McCarthy about the possibility of resignation, there is no concrete evidence or announcement that he will resign by the specified date. His history of fundraising and political influence, as well as the lack of a definitive statement about his future, suggest that a resignation by November 30 is not certain.
5. Initial Probability:
Given the considerations above, the initial probability of McCarthy resigning by November 30 seems low.
6. Adjusted Probability Considerations:
- The calculated probability should not be excessively confident in either direction due to the uncertainty expressed by McCarthy himself.
- The base rate for such high-profile resignations in the face of uncertainty is not well-established, which adds to the difficulty in predicting this event.
- Intuitively, politicians often weigh their options carefully, and a decision to resign is not typically rushed, especially without a clear strategic reason.
7. Final Answer: *0.20*

Figure 17: A response of the fine-tuned model on the question "Will Kevin McCarthy resign from the House by Nov 30?" (resolved NO) at the last retrieval index. Inspecting the original prompt (not shown here), we notice that our retrieval system presents articles, published on November 30, that clearly state that no official announcement had yet been made about Kevin McCarthy's potential resignation. On the day of November, 30, the community prediction had dropped to 1%, but our system hedges by claiming 20%.