C Dataset: Curation and Further Analysis

C.1 Data Collection and Curation

Scraping. To compile our dataset from the forecasting platforms, we query their APIs or scrape the questions' webpages for initial data gathering. For Metaculus, we first extract basic information via the API and scrape the resolution criteria from webpage. INFER (CSET) and Good Judgment Open data are gathered via web scraping, since no API provides the full data we need. Polymarket's data, except for community predictions, is obtained from their API. Manifold's data is fully scraped via API.

Question: {question}
Background: {background}
Options:

- Science & Tech
- Healthcare & Biology
- Economics & Business
- Environment & Energy
- Politics & Governance
- Education & Research
- Arts & Recreation
- Security & Defense
- Social Sciences
- Sports
- Other

Instruction: Assign a category for the given question.
Rules:

1. Make sure you only return one of the options from the option list.
2. Only output the category, and do not output any other words in your response.
3. You have to pick a string from the above categories.

Answer: {{ Insert answer here }}

Figure 8: **Prompt for categorizing questions based on the provided options**. The prompt presents the forecasting question, along with 11 candidate category choices, and prompts the model to classify the question into one of the categories.

Assigning categories. There is no standard, uniform categorization of the forecast questions across the platforms. We prompt GPT-3.5-Turbo to assign one of the 11 categories to each question. See Figure 8 for the category set and the prompt we use.

Screening and curation. From manual examination, we notice that the initial dataset contains questions that are ambiguously formulated or overly personal. In a preliminary screening phase, we prompt GPT-3.5 to identify and exclude these questions. See Figure 9 for a prompt to detect ill-defined questions, where we provide several few-shot examples.

I'm trying to assess the quality of an old forecasting dataset.
Here is a forecasting question from the dataset: {question}.
Please flag questions that don't sound like binary forecasting questions by outputting "flag". If it sounds like a reasonable question, output "ok".

Examples of strings that should be flagged:

- Will I finish my homework tonight?
- Metaculus party 2023
- Will Hell freeze over?
- Heads or tails
- Will this video reach 100k views by the EOD?

Examples of strings that should not be flagged:

- Will Megan Markle and Prince Harry have a baby by the end of the year?
- Will the Brain Preservation Foundation's Large Mammal preservation prize be won by Feb 9th, 2017?
- Will there be more novel new drugs approved by the FDA in 2016 than in 2015?

If a question is already resolved, that doesn't mean it should be flagged. When in doubt, mark it as "ok".
Your response should take the following structure:
Insert thinking: {{ insert your concise thoughts here }}
Classification: {{ insert "flag" or "ok" }}

Figure 9: **The prompt for flagging ill-defined forecasting questions** in our dataset. The prompt contains several few-shot examples where the questions are ill-defined. A LM is prompted to filter out any questions of similar nature.

We then manually examine to eliminate all questions of low quality. This includes those with few community forecasts or trading engagement on platforms such as Manifold and Polymarket, as well as any ill-defined questions that GPT-3.5 is unable to identify during the initial screening.

C.2 Further Statistics and Samples

We give a list of detailed statistics and plots on our data:

- Figure 10 visualizes the location mentions in all the questions from our full dataset.
- Table 11 gives the distribution of questions and forecasts across platforms in our full dataset.
- Table 12 showcases a complete data sample in our curated set.
- Table 13 shows a list of questions with how community predictions shift over time.
- Figure 11a shows the opening dates of the questions in the full dataset.
- Figure 11b shows the percentage of questions that receives the retrieval date at index $k = 1, 2, 3, 4, 5$.

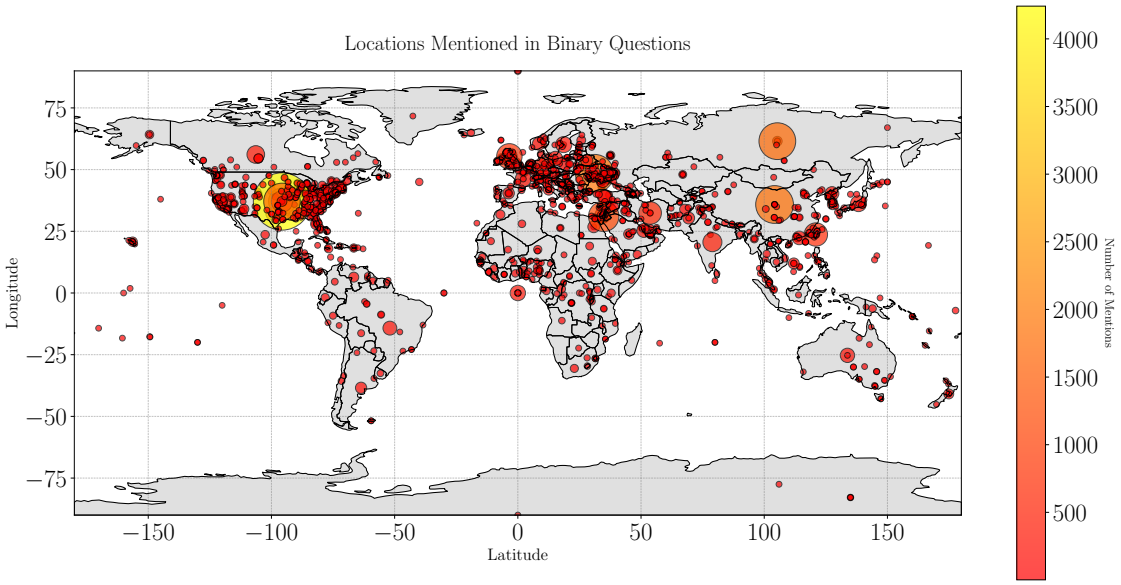
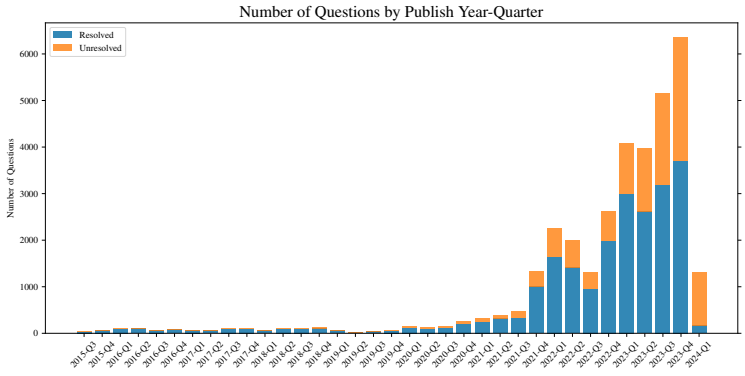


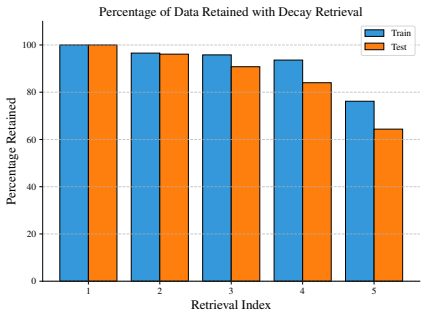
Figure 10: **Location mentions in all binary questions in our full dataset**. We visualize all location mentions in our full dataset on a world map. This shows that the dataset provides a diverse coverage of topics across the globe.

Platform	Questions (All)	Predictions (All)	Questions (Binary)	Predictions (Binary)	Brier Score (Binary)
Metaculus	8,881	638,590	4,862	387,488	.130
INFER	308	73,778	192	47,918	.079
GJOpen	2,592	743,671	1,168	342,216	.134
Manifold	24,284	1,997,928	20,319	1,387,668	.155
Polymarket	12,689	3,720,640	7,123	1,879,035	.158

Table 11: **Raw dataset statistics across platforms**. The Brier scores are calculated by averaging over all time points where the platforms provide crow aggregates.



(a) **Distribution of the opening dates of the questions in our full datasets**, ordered by year-quarter. Activity on these platforms has sharply increased over the past two years.



(b) **Percentage of questions that receive retrieval** at each retrieval index (1–5). The late retrieval indices can miss certain questions, since questions may resolve much earlier than the close time.

Figure 11: **Question publish time distribution and retrieval dates**.