3.1.1 QUESTIONS AND RESOLUTION VALUES

We bring in questions from two broad types of sources: **markets** and **datasets**.[1] We select multiple, reliable sources from each type, ensuring the diversity of our benchmark. See Table 1 for details.

Markets We compile market questions from a handful of prediction markets and forecast aggregation sites that elicit human predictions on questions across a wide range of topics.[2] When selecting questions from market sources to add to the question bank, we choose those with high levels of active human trading on these platforms (liquidity) as these questions tend to be of higher quality than those with low levels of human trading.[3] We store the latest market and resolution values for each question.

Datasets We create dataset questions from well-established and well-maintained datasets that track real-world events (e.g., ACLED (Raleigh et al., 2023), a geopolitical database that tracks worldwide conflict, including facts like the number of protests in Niger each month). With these sources, we can generate questions using a fixed question template (e.g., "Will the number of protests in Niger increase by at least 10% over this month's value by the resolution date?").[4]

Question Bank Table 1 lists the sources of market and dataset questions, along with the number of questions available in our question bank, whence we sample questions for our benchmark runs. In addition to the main questions (with sample size N), we also construct additional *combination* questions by choosing pairs of questions within each source. We describe combination questions in more detail in Section 3.2.

Table 1: Question bank composition, grouped by source type (market or dataset).

Source	URL	N	$\binom{N}{2}$
RFI	randforecastinginitiative.org	18	153
Manifold Markets	manifold.markets	405	81,810
Metaculus	metaculus.com	722	260,281
Polymarket	polymarket.com	915	418,155
Market Total		2,060	760,399
ACLED	acleddata.com	3,220	5,182,590
DBnomics	db.nomics.world	52	1,326
FRED	fred.stlouisfed.org	166	13,695
Wikipedia	wikipedia.org	428	91,378
Yahoo! Finance	finance.yahoo.com	509	129,286
Dataset Total		4,375	5,418,275
Question Bank Total		6,435	6,178,674

3.1.2 QUESTION METADATA

After we automatically update the question bank with new forecasting questions and resolution values, the data are processed in several ways. This generates more information about the questions and creates another sampling option when creating the question set. Following Halawi et al. (2024), we use `gpt-4o-mini` to categorize questions by subject and to filter out low-quality questions. We report the number of questions by category and source in Table 15, where we display the breakdown of the standard questions from our question bank (N from Table 1) across question categories by source, after removing low-quality questions.

[1] Licensing information can be found in Section C.1.

[2] Hereafter, for brevity, we refer to questions that come from both prediction markets and forecast aggregation sites as "market" questions. We also refer to theses sources collectively as "market" sources.

[3] See Table 8 for an example question pulled from a market source.

[4] See Table 9 for an example question pulled from a dataset source.

3.2 QUESTION SETS

LLM question set We release a set of 1,000 forecast questions for LLMs every other Sunday at midnight UTC. We sample an equal number of questions from each source to ensure representativeness. Within each source, we then uniformly sample questions across all question categories, aiming for an equal distribution from each category within each source. This ensures that models cannot be overfit to a specific type of question or topic. Limiting the number of questions generated to 1,000 also ensures that costs for testing LLMs on the benchmark are capped for development teams with fewer resources.

Human question set The human question set is comprised of 200 forecast questions sampled directly from the LLM question set. When sampling, we do our best to maintain proportionality across question sources and across categories within each question source; this ensures the question set addresses as many domains as possible.

Forecast horizons For questions derived from dataset sources, the distribution of resolution dates is the $forecast\_due\_date + n$ days, where $n \in \{7, 30, 90, 180, 365, 1095, 1825, 3650\}$. In other words, for each dataset question we ask for 8 forecasts that differ only in their resolution date. For questions derived from market sources, we ask for only 1 forecast: the probability that each question will resolve positively (will the event underlying the question occur, or not). This setup will allow us to evaluate both human and LLM forecasting performance over the short, medium, and long term.

Combination questions There are two types of questions, each comprising half of the question set. The first type is a standard forecasting question with a binary outcome, e.g., "Will inflation (core CPI) be above 3% next month?" We construct the second type, *combination* questions, by pairing two standard questions. For combination questions, we ask for forecasts on all Boolean combinations of the two questions (i.e., $P(Q1 \cap Q2)$, $P(Q1 \cap \neg Q2)$, ...). Considering the extensive number of standard questions and potential combinations in our question bank (as we could potentially combine 3, 4, or more standard questions together), we effectively have access to millions of possible forecasting questions from which we can sample. We show the number of two-question combination questions in the question bank as it stands at time of writing in the right-hand column of Table 1. This setup implies that for market combination questions, each forecaster will provide 4 forecasts, whereas for dataset combination questions, each forecaster will provide 32 forecasts (4 for each Boolean combination of $Q1$ and $Q2$ at each of the 8 forecast horizons).

Combination questions require forecasters to consider the covariance structure of different events, some of which are more independent than others. For instance, the best forecasts of whether the S&P500 (a key U.S. stock market index) will reach an all-time high and whether Spain will win the next Men's World Cup are likely independent. However, the best forecast of whether the S&P500 will reach an all-time high and whether the U.S. will enter an economic recession must account for the likely strong correlation between these events.

Of the 1,000 questions in the LLM question set, 500 are combination questions. Each combination question is composed of two standard questions from the same question set. This means that LLMs will also provide forecasts for the individual components of these combination questions, since they're drawn from the existing standard questions. Importantly, none of the 200 questions in the human question set are combination questions.[5]

Timeline of forecasting round To compare human performance to LLMs, we periodically run surveys asking the general public and superforecasters to forecast on our question sets (see Section 4). We produce the question sets 10 days before the forecast due date to allow for time to create and run a human survey. LLM teams receive their question set 24 hours before the due date, even though it was generated at the same time as the human question set. This constrained time frame gives teams less time to be able to game the system. We thus elicit forecasts, obtaining comparable forecast sets on the due date from both LLMs and humans who faced the same information environment.

[5]We exclude combination questions because expert human forecasters are expensive; we maximize their relevance by having them focus on standard questions.

3.3 RESOLUTION

We resolve forecast sets nightly by gathering the latest information about which events have or have not occurred. All questions are ultimately resolved to ground truth.

Evaluating performance on market questions For market questions, ground truth is not available until the question has been resolved on the platform. Until then, we evaluate performance by calculating the squared distance between the forecasted value and the platform's crowd forecast (an aggregate of human forecasts reported on the platform) from the preceding day.[6] This provides a good estimate of forecast performance on unresolved questions since crowd forecasts tend toward ground truth as the resolution date approaches.

Once a market question has officially been resolved, we score the forecast against the resolution value, creating a definitive score for the question. We are thus able to estimate forecast performance on the entire set of market questions (resolved and unresolved), incorporating all information available to markets on a nightly basis.

Evaluating performance on dataset questions Datasets can be updated as new information becomes available. Hence, questions derived from datasets are continuously resolved to the value from the latest available data. As the resolution dates (ranging from a 7-day to a 10-year horizon) for dataset questions come due, a new round of forecasts is evaluated. We thus are able to evaluate forecasting performance over different time horizons.

Missing forecasts We select 1,000 questions for the LLM question set to make the forecasting task impractical to complete manually within the 24-hour window after the question set is released. And we obligate all LLMs to forecast on all questions to ensure comparability of scores across models and human-based aggregates. When a model does not submit forecasts on certain questions or time horizons, we consider that a model error and impute a naive forecast for the model to ensure comparability across models over time.

For market questions, since we only ask for forecasts on the outcome of the question, missing forecasts are assigned the value of the crowd forecast on the forecast due date. Some may argue this is overly-generous, but we did not want forecasters to have a competitive advantage by simply scraping the market websites themselves.

For dataset questions, we impute the value 0.5 (which represents an uninformed 50% forecast) to forecasts across all time horizons. Empirically, top models report valid forecasts on all questions and are not affected by this imputation procedure.

3.4 LEADERBOARDS

We generate leaderboards that rank models and humans by average overall score. The main leaderboard highlights the top forecasting submission across all questions and can by sorted by performance on the question type (market or dataset) and by resolution status (resolved or unresolved). The leaderboard is updated nightly after scoring forecasts against the latest data, market resolution values, and crowd forecasts.

3.5 DATASETS

As a key output of ForecastBench, we generate four datasets that grow over time.[7]

General public forecast dataset Every time we run the public survey described in Section 4, we provide multiple independent forecasts and rationales for every one of the 200 forecast questions and report the accuracy of the median public forecast. See Section B.2.1 for details.

[6] The crowd forecasts are updated nightly in the Question Bank as described in Section 3.1.1.

[7] See Appendix A for licensing details and Appendix B for data dictionaries.