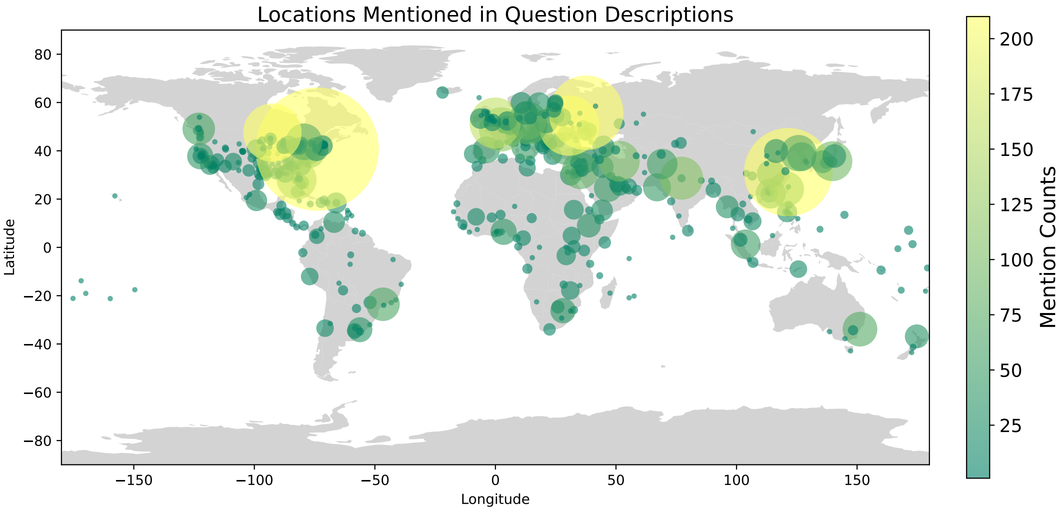

Figure 3: Autocast contains questions about locations across the world. The questions in the dataset mention over 500 cities, spanning six continents.

design an architecture for this purpose (albeit with limits on article length and time horizon), drawing inspiration from Wang and McAllester (2020).

Calibration. Calibration is important in forecasting (Tetlock and Gardner, 2016). Even expert forecasters will be highly uncertain about some outcomes of interest. Such forecasts will be more useful in the form of calibrated probabilities than as point estimates. Thus forecasters are evaluated with proper scoring rules, which incentivize calibration. There is an extensive literature on improving the calibration of deep learning models (Guo et al., 2017; Nguyen and O'Connor, 2015; Lin et al., 2022; Minderer et al., 2021; Kull et al., 2019b), mostly for classification with a fixed set of classes. One part of Autocast requires models to forecast continuous quantities varying over multiple orders of magnitude, which has not been explored in prior work.

Truthful question-answering. Current language models often generate falsehoods when answering questions (Shuster et al., 2021; Lin et al., 2021), and they also achieve poor calibration when giving probabilistic answers (Hendrycks et al., 2021a) to human knowledge questions. However, for questions with a known ground truth answer, we expect models to improve as a result of scale, fine-tuning, and information-retrieval from reliable sources (Bai et al., 2022; Nakano et al., 2021; Hadfield-Menell et al., 2016; Turner et al., 2020; Wainwright and Eckersley, 2019). Yet humans also want models to give calibrated and truthful answers to questions that are too difficult or costly for us to answer ourselves (Irving et al., 2018; Evans et al., 2021; Leike et al., 2017; Hendrycks et al., 2021d; Reddy et al., 2020; Nahian et al., 2021). Forecasting is useful for this purpose. Forecasting questions are challenging but eventually become easy to evaluate. By contrast, it may be difficult for humans to evaluate superior answers to open problems in fundamental philosophy or science.

3 The Autocast Dataset

Forecasting Questions. We collected all available forecasting questions from three public forecasting tournaments (Metaculus, Good Judgment Open, and CSET Foretell), which resulted in 6,707 questions total. These questions tend to have broad public interest (e.g., national rather than local elections) and clear resolution criteria. Most questions are not already covered well by specialized forecasts (such as weather forecasts). The questions are either true/false, multiple-choice, or involve forecasting a numerical quantity or date (see Table 1 for examples). In these forecasting tournaments, participants begin forecasting a question on a given day (the "start date") and update their forecasts multiple times up until the "close date." At some later time, the forecast is *resolved* and participants are scored based on all their forecasts. (Note the resolution date is often just after the closing date but not always. The resolution can also happen *before* the planned closing date: e.g. when forecasting when an event will occur.) Thus the "crowd" forecast (which aggregates over participants) is a time-series of forecasts from the start to close date.

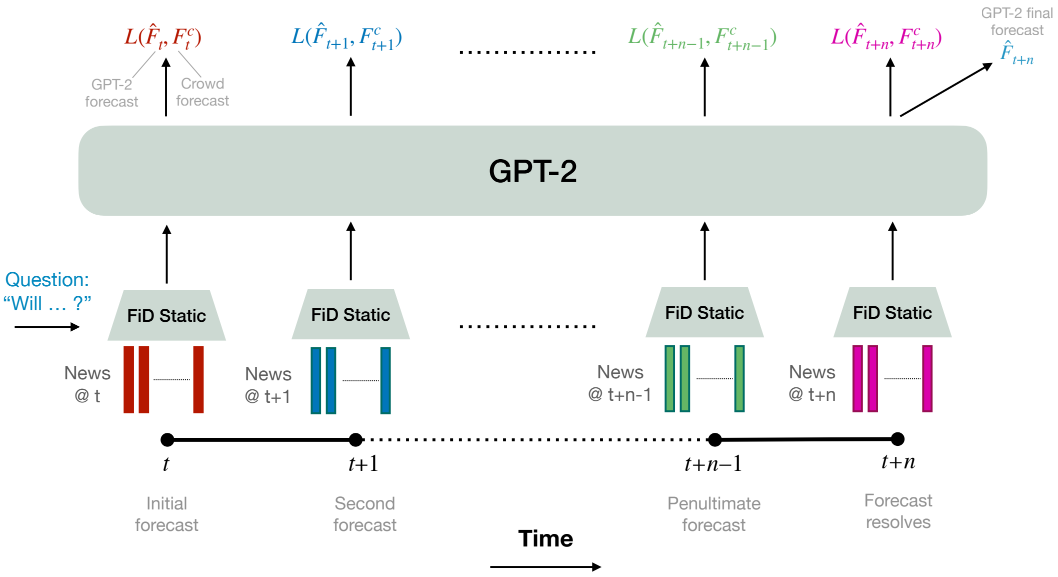

Figure 4: **Illustration of our FiD Temporal model.** Forecasts are made each day (from start date to resolution) by GPT-2. The input to GPT-2 is the top-1 daily news article retrieved by BM25, which is encoded by FiD Static (a T5 model). In training, GPT-2's target is the average of daily crowd predictions (denoted 'F^c_t' for day t) and the resolved outcome. Like human forecasters, GPT-2 accumulates news information over time and updates its predictions.

Autocast includes the question, the start and close dates, the answer (if the question has resolved), and the time-series of crowd forecasts (Figure 1). Half of the questions have not yet resolved and correspond to ongoing tournaments. Some of these questions concern events decades in the future, requiring reasoning over long time horizons. These questions can still be used as training data by using the crowd forecast as the target (as a high-quality proxy for the ground truth). However, the test set only includes resolved questions. Our dataset also includes metadata that is helpful for forecasting. There is detailed background information about the question (including precise terms of resolution) and also links to relevant information posted by tournament participants. We include more details in the appendix.

Train and test split. It is standard in ML for the test set to be drawn from the same distribution as the train set. However, randomly splitting our questions into train and test without considering the date would not simulate the conditions of forecasting. For example, a test question ("Will Trump win the 2020 election?") could come from an earlier date than a related training question ("Will President Biden pass the stimulus?"). Thus, we split our questions using a date cut-off of mid-2021, which means that questions in the test set resolve from mid-2021 to mid-2022. Note that if a model is pre-trained on data from after mid-2021, this will also not simulate forecasting faithfully. In both train and test sets, we implement dataset balancing for the true/false questions. To flip a label, we negate the question using OpenAI's GPT-3-175B Edit model (Brown et al., 2020) and manually check for correct negation.

Contemporaneous news as context for forecasts. When a human is making a forecast at time t, they use past and present ($\leq t$) information sources but are not exposed to any information from the future ($> t$). If they forecast again at $t + 1$, they will have updated on new information that was generated from t to $t + 1$. These conditions can be simulated for ML models by (a) pre-training on text generated before time t, and (b) providing the model with new information generated between t and $t + 1$. To this end, we provide a corpus of news articles scraped from CommonCrawl news (Nagel, 2016; Hamborg et al., 2017) that is organized by publish date. The articles were derived from diverse sources between 2016 to mid-2022 and total more than 200GB of data.

Model	Parameters	T/F	MCQ	Numerical	Score	Average
Random	–	50.0	22.1	34.5	18.8	18.8
UnifiedQA	0.2B	45.4	23.5	34.5	17.2	19.5
	0.8B	48.2	23.5	34.5	18.6	
	2.8B	54.9	25.1	34.5	22.8	
T5	0.2B	61.3	24.0	20.5	32.4	32.9
	0.8B	60.0	29.1	21.7	33.7	
	2.8B	60.0	26.8	21.9	32.5	
FiD Static	0.2B	62.0	29.6	24.5	33.5	37.2
	0.8B	64.1	32.4	21.8	37.4	
	2.8B	**65.4**	35.8	19.9	**40.6**	
FiD Temporal	0.6B	62.0	33.5	23.9	35.8	**37.8**
	1.5B	63.8	32.4	21.0	37.6	
	4.3B	62.9	**36.9**	**19.5**	40.1	
Human Crowd	–	92.4	81.0	8.5	82.5	82.5

Table 2: Model accuracy on the Autocast dataset for each question type: true/false (T/F), multiple-choice question (MCQ), and numerical (Numerical). For Numerical, lower is better. For other metrics, higher is better. The model FiD Static (based on T5) retrieves the top 10 news articles over the period, while FiD Temporal (based on GPT-2 with T5 encoder) retrieves the top 1 article each day. Averaging over all model sizes, we find that the FiD Temporal achieves the best average.

3.1 Dataset Analysis

Distribution of Questions. The questions in Autocast cover a very wide variety of topics. We divide the questions into five main categories: Economy, Politics, Science, Social, and Other. Each category contains numerous subcategories for a total of 44 subcategories ranging from foreign policy to AI. We list all subcategories in the Supplementary Material. The questions also cover a wide geographical distribution, as shown in Figure 3. Overall, Autocast tests both breadth of subject matter and depth (since questions ask for quantitative predictions about a specific, operationalized variable).

Adding new questions over time. The number of questions submitted to forecasting platforms is rapidly increasing (Figure 2). If trends continue, in two years there will be twice as many questions available. Autocast is a living dataset and will be updated periodically with new questions. This will provide both more data for training and a new set of test questions (to assess overfitting).

Human forecasts. The human crowd forecast for a given question becomes more accurate from the start to closing date, as shown in Figure 6. This is what we would expect if humans are updating their forecasts as more information comes out. In contrast to most ML benchmarks, the human crowd judgments are probabilistic. This allows us to evaluate their calibration. In the Supplementary Material, we show that crowd forecasts are well-calibrated.

Distribution shift. We expect a distribution shift over time in both the questions being asked and in the answers. For example, there will be fewer questions about Ukraine before 2022. This distribution shift is inherent to forecasting and so it is crucial that models can manage it. We do find a shift in the distribution of question categories. For example, the number of questions in the Social category increased from 12.6% in the training set to 28.2% in the test set, possibly due the Covid-19 pandemic (which is included in Social).

4 Experiments

4.1 Baselines

The *Crowd* baseline uses the final aggregate human forecast before the closing date. The *Random* baseline uses the analytically computed random accuracy for true/false and multiple-choice questions.