Algorithm 1 RMS Calibration Error

Input: A set of N examples each with label $\{y_i\}_{i=1}^{N}$ and C predicted confidence intervals $\{(l_i^c, u_i^c)\}_{c=1,i=1}^{C,N}$ corresponding to C confidence levels $\{\mathcal{I}^c\}_{c=1}^{C}$ (e.g., $\mathcal{I}^C = 0.95$). Set bin size to M.
function AdaptiveRMS
  Sort the examples by labels y_n in ascending order.
  Assign a bin label $b_k = \left\lfloor \frac{k-1}{M} \right\rfloor + 1$ to each by splitting sorted examples into chunks of M.
  Let $\{B_i\}_{i=1}^{b}$ be the set of bins and B_i the subset of examples in bin i.
  for $c = 1, \ldots, C$ **do**
    Calculate empirical containment for bin i

$$\widehat{p}_i^c = \frac{1}{|B_i|} \sum_{k \in B_i} \mathbb{1}(y_k \in [l_k^c, u_k^c])$$

    Calculate root mean squared calibration error

$$\text{RMS}^c = \sqrt{\frac{1}{b} \sum_{i=1}^{b} (\widehat{p}_i^c - \mathcal{I}^c)^2}$$

  end for
  Output $\frac{1}{C} \sum_{c=1}^{C} \text{RMS}^c$
end function

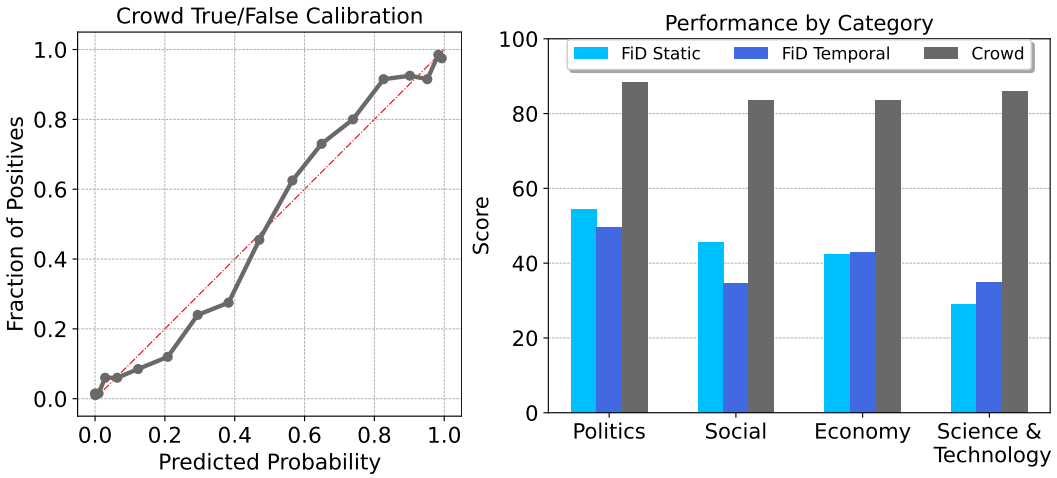

Figure 8: Left: Crowd forecasts for true/false questions have good calibration. Right: The per-category performance of baselines. Score indicates the combined score metric.

Calibration Dataset Statistics. The dataset of numerical questions gathered for our calibration evaluations has training, validation, and test sets containing 32,200, 3,443, and 6,170 examples respectively.

B Additional Dataset Information

B.1 Dataset Details

The Autocast dataset contains 6,707 unique questions in total, spanning three question types, including resolved and unresolved. After we balance the true/false questions by adding negated questions, the true/false question count doubles, making the grand

	T/F	MCQ	Numerical
Train	3187	753	471
Test	775	176	341
Total	3962	929	812

Table 5: The number of resolved questions in Autocast, grouped by question type.

Category	Percentage	Subcategories
Politics	31%	Geopolitics, Security and Conflict, Elections, Foreign Policy, Leader Entry/Exit, Law, Economic Policy, US Policy, Ukraine
Social	22%	COVID-19, Social Issues, Environment, Effective Altruism, Sports, Entertainment, Health, Society, Pandemic, Animal Welfare, Metaculus, Climate, Education
Science & Tech	21%	Technology, Computing, Biological Sciences, Physical Sciences, Computer Science, Biology, Human Sciences, AI, Mathematics, Tech
Economy	20%	Business, Finance, Industry, Economic Indicators, Infrastructure, Microelectronics, Semiconductors
Other	6%	Other, Open

Table 6: The percentage of Autocast questions in each category, and the subcategories belonging to each category. Autocast questions have fairly even coverage of a wide variety of topics.

total 9,757. The numbers of training and test examples are shown in Table 4 for ease of reference. The numbers below are based on the expanded dataset using true/false balancing. The Autocast training set we experiment with does not include unresolved questions. This training set contains 4,411 examples, and the test set contains 1,292 examples. To prevent leakage of future information, the train set consists of all questions that closed or resolved before 5-11-2021 and the test set consists of all questions that closed or resolved after 5-11-2021. In addition, we also release 1,974 unresolved train questions having a publish date before 5-11-2021 and 2,305 unresolved test questions published after 5-11-2021. Note that our baselines do not use any unresolved questions, so there is a guarantee of no leakage. However, training with auxiliary training signals from unresolved questions (e.g., crowd forecasts) requires additional care to ensure no leakage. Namely, crowd forecasts from after 5-11-2021 must not be used.

Per-Category Performance. In Figure 8, we show performance by category using the combined score metric. Science & Technology questions are the most challenging for the FiD Static and FiD Temporal baselines, while the crowd predictions perform similarly on all question categories. There is a substantial gap between models and crowd predictions, but crowd predictions are still far from a perfect score of 100%.

Computation of Crowd Forecasts. The human crowd forecasts are directly obtained from forecasting platforms, and the precise meaning depends on the platform. For example, for Metaculus questions the crowd forecast represents the median forecast with the recent player predictions weighted more. For Good Judgment Open questions, it represents the median of the recent 40% of forecasts. In all cases, the crowd forecast aggregates previous individual forecasts at a given time.

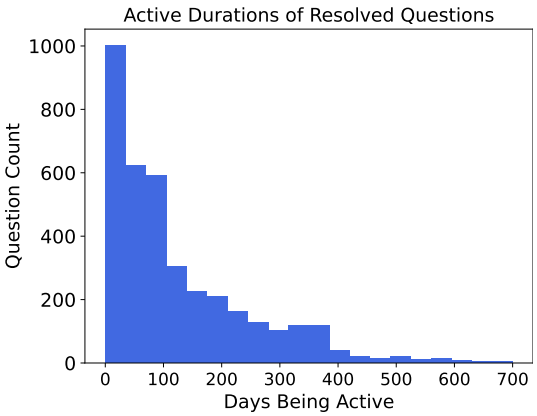

Figure 9: We visualize the distribution of the duration of the active periods for Autocast questions. Questions vary greatly in terms of how long they are active in the forecasting market, with questions taking up to years to resolve.

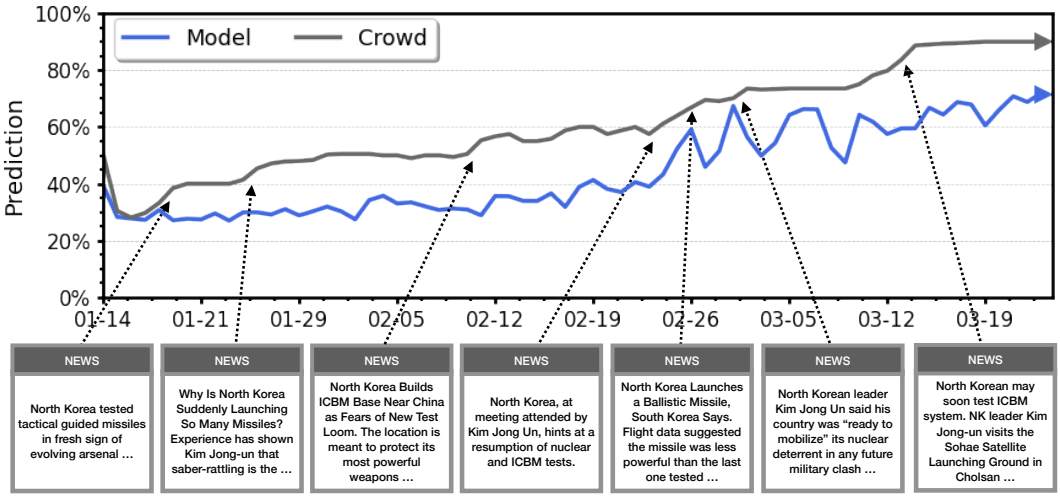

Figure 10: The same example from the Autocast dataset shown in Figure 1, illustrating how the crowd forecast is influenced by news articles published throughout the prediction period.

C X-Risk Sheet

We provide an analysis of our paper's contribution to reducing existential risk from future AI systems following the framework suggested by (Hendrycks and Mazeika, 2022). Individual question responses do not decisively imply relevance or irrelevance to existential risk reduction.

C.1 Long-Term Impact on Advanced AI Systems

In this section, please analyze how this work shapes the process that will lead to advanced AI systems and how it steers the process in a safer direction.

1. **Overview.** How is this work intended to reduce existential risks from advanced AI systems?

   Answer: This work builds towards improving institutional decision making and systemic safety. In short, this could help resolve matters of fact that influence policies and decisions made by political leaders in an increasingly complex modern world, putting humanity in a better place to deal with the global turbulence and uncertainty created by AI systems when they rapidly reshape society. A fuller motivation for "ML for Improving Epistemics" is described in Hendrycks and Mazeika (2022).

2. **Direct Effects.** If this work directly reduces existential risks, what are the main hazards, vulnerabilities, or failure modes that it directly affects?

   Answer: This directly works against failure modes such as eroded epistemics and hazards such as highly persuasive or manipulative AI systems.

3. **Diffuse Effects.** If this work reduces existential risks indirectly or diffusely, what are the main contributing factors that it affects?

   Answer: This work could lead to improved decision making, epistemics, and collective intelligence. Automated forecasting tools could eventually assist various levels of the sociotechnical hierarchy, including congress and legislatures; government regulatory agencies, industry associations, user associations, etc.; and company management. This lowers the risk of conflict that would accelerate the weaponization of AI, so it diffusely works against weaponized AI failure modes.

4. **What's at Stake?** What is a future scenario in which this research direction could prevent the sudden, large-scale loss of life? If not applicable, what is a future scenario in which this research direction be highly beneficial?