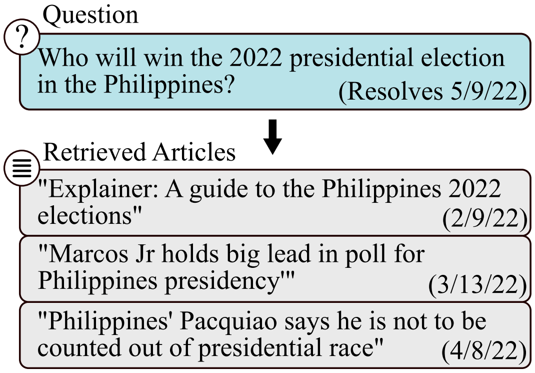
For numerical questions, random predictions are sampled uniformly from the bounded range of possible answers specified in each question.

Models without retrieval. We evaluate *UnifiedQA-v2* (Khashabi et al., 2022) and *T5* (Raffel et al., 2020) models of various sizes. These models are trained on a variety of tasks, enabling strong generalization on many unseen language tasks. Using zero-shot prompting for UnifiedQA, we report results on classification questions. The UnifiedQA models were not trained on numerical questions, hence, we report random performance to enable comparison with other baselines. T5 is fine-tuned using its original output head for true/false and multiple-choice questions. To output numerical answers with T5, we add an additional linear output head.

Retrieval-Based Methods. We investigate whether retrieval models can improve performance by selecting relevant articles from the news corpus included with Autocast. Importantly, news articles after the close time or resolution time of a question are not available for retrieval, so retrieved articles only include information about the past. For all retrieval methods, we use Fusion-in-Decoder or FiD (Izacard and Grave, 2021) to encode articles retrieved by BM25 (Robertson et al., 1994; Thakur et al., 2021) with cross-encoder reranking. FiD uses T5 to encode retrieved passages along with the question and can be viewed as a minimal extension of T5 for incorporating retrieval. We truncate retrieved articles to a maximum length of 512 tokens.



Who will win the 2022 presidential election in the Philippines? (Resolves 5/9/22)

"Explainer: A guide to the Philippines 2022 elections" (2/9/22)

"Marcos Jr holds big lead in poll for Philippines presidency'" (3/13/22)

"Philippines' Pacquiao says he is not to be counted out of presidential race" (4/8/22)

Figure 5: Articles retrieved by BM25 for a Politics question in the Autocast dataset with publication dates in parentheses. The articles are retrieved from 200GB of news and are highly relevant to making an informed forecast.

The *FiD Static* baseline uses the top 10 retrieved articles after reranking, which is the standard method for retrieval-augmented prediction. The *FiD Temporal* baseline leverages the intermediate crowd predictions (before the question is resolved) as auxiliary supervision. The intuition is that crowd predictions will change based on rational incorporation of new evidence, and these updates will not be captured by just training on the final outcome. For each day between the question's open and close date, we generate an embedding of the top news article using the frozen fine-tuned FiD Static model. These embeddings are then treated as input embeddings to an autoregressive model (GPT-2 (Radford et al., 2019)), which is fine-tuned to predict the average of the daily crowd prediction and the ground truth outcome. We illustrate this method in Figure 4. Figure 1 shows predictions from an FiD Temporal model over time for an example question.

4.2 Metrics

For true/false and multiple-choice questions, we evaluate models using percent accuracy. For numerical questions, we use ℓ_1 distance, bounded between 0% and 100%. We denote these question types as T/F, MCQ, and Numerical, respectively. To evaluate aggregate performance, we use a combined Score metric $(\text{T/F} + \text{MCQ} - \text{Numerical})/2$, which has an upper bound of 100%. A score of 100% indicates perfect prediction on all three question types. Note that since numerical question responses are normalized between 0% and 100%, the combined Score metric is lower-bounded at -50%. We also report the Average score, which averages the combined metric of all model sizes.

4.3 Forecasting Evaluation

Setup. We fine-tune the T5 baseline for 10 epochs with a batch size of 8, an initial learning rate of 5×10^{-5} with linear decay schedule, and a weight decay of 1×10^{-2}. The maximum sequence length of the T5 model is set to 512. We train FiD Temporal models for 5 epochs with a constant learning rate of 5×10^{-5}. Hyperparameters are selected based on early experiments. Additional details are in the Supplementary Material.

Results. We show results in Table 2. Although UnifiedQA-v2 obtains strong performance on various natural language benchmarks, it obtains close to random zero-shot performance on Autocast, showing the difficulty of forecasting. Fine-tuned T5 performs better, but multiple-choice accuracy is still at nearly random chance levels. Retrieval-based methods substantially outperform both UnifiedQA-v2 and T5, showing a relative increase in the Average score of 93% and 15%, respectively. Moreover, retrieval-based methods become more effective as parameter count increases, which suggests that the models learn to extract relevant information from retrieved articles.

Comparing the FiD Static and FiD Temporal baselines, we see that the Average score is slightly higher for FiD Temporal. However, the largest FiD Static model has the highest individual score. Thus, our temporal training strategy for incorporating the auxiliary crowd predictions neither harms nor helps compared to the static retrieval baseline. Future work could develop more effective ways of using these auxiliary training signals.

4.4 Model Analysis

Relevance of Retrieved Articles. We find that the retrieved articles are often highly relevant to the question. In Figure 5, we show examples of articles retrieved by BM25 from the news corpus in Autocast. Baseline models have access to the article text, but for brevity we only show the article title. The articles give information that is clearly relevant for making an informed forecast. Note that the T5 backbone for the baselines was pre-trained on data from before 2020, far before the timeline of the question, so retrieval provides vital information that models would not otherwise have. This suggests that large improvements on Autocast could come from integrating information from retrieved articles more effectively. We expect that more sophisticated retrieval methods would also improve performance, although efficiency becomes a concern when using large retrieval methods.

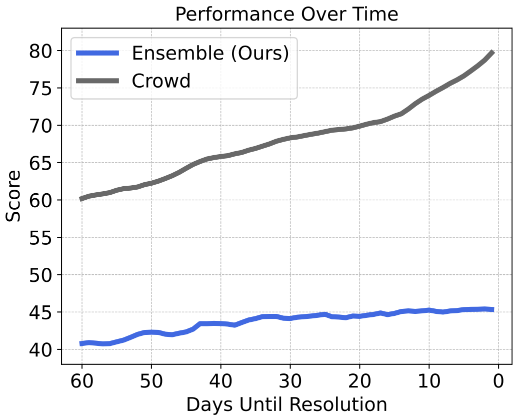


Figure 6: For the crowd and an ensemble of the two largest FiD Temporal models, prediction score increases as the resolution date grows nearer. This trend may be due to more relevant information becoming available over time, which the model can access through retrieval from the news corpus.

Detailed Performance. In the Supplementary Material, we show the performance of baseline methods on a more granular level. The per-category results indicate that Science & Technology is the most challenging category for models, whereas human forecasters have relatively consistent performance across categories. Inspecting the subset of questions that have been active for at least two months, we also find that the accuracy of the human crowd forecast and a model ensemble steadily increases over time up to the resolution date (Figure 6). This is to be expected, as more information about the eventual outcome is available closer to the time (e.g. election polls become more accurate). For this plot, we show an ensemble of the two largest FiD Temporal models, which has slightly higher final performance than the individual models and a clearer trend over time.

5 Calibrated Prediction of Numerical Quantities

In our results, we evaluate baselines on the accuracy of their point estimates, rather than their calibration. However, the eventual goal for Autocast is for models to achieve good calibration as well as accuracy. Here we describe an auxiliary dataset that helps with this goal for the challenging case of calibration on numerical questions.

The IntervalQA Dataset. In the Autocast training set, numerical quantities range over many orders of magnitude. Furthermore, Autocast has fewer than $1{,}000$ numerical training questions. This problem of making *calibrated* predictions for quantities over many orders of magnitude using text inputs has not been addressed in work on calibration for language models. To this end, we curate

Parameters	Point Estimate Distance	Conf. Interval Length	RMS Calibration Error
22M	20.8	2072.4	19.1
44M	20.3	1115.7	16.6
86M	19.6	763.1	16.9
304M	**18.1**	**305.4**	**13.5**

Table 3: Results for DeBERTa-v3 models trained to output confidence intervals on our dataset of numerical predictions. The high dynamic range of the targets leads to large confidence intervals, but median interval size decreases with larger models as does RMS Calibration Error.

IntervalQA, an auxiliary dataset of numerical estimation problems and provide metrics to measure calibration. The problems in the dataset are not forecasting problems but instead involve giving calibrated predictions for fixed numerical quantities. The questions were sourced from NLP datasets covering diverse topics and with answers varying across orders of magnitude: SQuAD, 80K Hours Calibration (80k, 2013), Eighth Grade Arithmetic (Cobbe et al., 2021), TriviaQA (Joshi et al., 2017), Jeopardy, MATH (Hendrycks et al., 2021b), and MMLU (Hendrycks et al., 2021a). We filtered these datasets for questions with numerical answers, which yielded about 30,000 questions.

5.1 Metrics

We evaluate whether confidence intervals are calibrated. Concretely, if a method outputs 80% confidence intervals on each test example, we would like the true prediction target to fall inside of these intervals 80% of the time. Additionally, we would like for models to be calibrated across their entire dynamic range of outputs. To measure this, we compute *RMS Calibration Error* similarly to Nguyen and O'Connor (2015) and Hendrycks et al. (2019), but with fixed confidence levels $c \in \{50\%, 55\%, \ldots, 95\%\}$ and such that calibration is sensitive to dynamic range. We describe this metric in detail in the Supplementary Material. Low RMS Calibration Error indicates that models are calibrated across their entire dynamic range. We also compute the median prediction error between the predicted point estimate and the ground-truth target (*Point Estimate Distance*) and the median interval length averaged across all confidence levels (*Conf. Interval Length*).

5.2 Experiments

We fine-tune DeBERTa-v3 models (He et al., 2020) to predict a point estimate and a set of confidence intervals corresponding to the confidence levels in the RMS calibration error metric. On a high level, we use a loss with three components: (1) MSE loss between the predicted point estimate and the ground-truth target, (2) MSE loss between the boundaries of the predicted confidence intervals and the ground-truth target for boundaries that are on the wrong side of the target, (3) a penalty on the length of the predicted intervals to encourage finer predictions. The models are trained for 5 epochs with a batch size of 100. A detailed description is in the Supplementary Material. We show results in Table 3. All three metrics decrease with model size.

6 Conclusion

We introduced Autocast, a dataset for measuring the ability of neural networks to forecast future world events. The dataset contains thousands of forecasting questions from public forecasting tournaments, including ground truth outcomes and aggregated human predictions. We also curated a large corpus of news items from the Common Crawl news corpus, enabling rigorous evaluations without information leakage. We evaluated numerous baseline algorithms and demonstrated that model size and information retrieval can improve forecasting performance. To better evaluate calibration for numerical prediction, we introduced IntervalQA, a large collection of numerical prediction questions with a wide dynamic range of prediction targets, and evaluated state-of-the-art language models. Our results show significant room for future improvement.