Methods / Metrics	Accuracy (%, ↑)			Brier score (↓)		
	yes/no	*multi*	*all*	*yes/no*	*multi*	*all*
Random	48.6	25.3	37.8	0.684	0.827	0.750
ESIM-ELMo (closed-book)	63.3	45.8	54.5	0.515	0.897	0.706
$\text{BERT}_{\text{BASE}}$ (closed-book)	66.2	41.5	54.7	0.511	0.715	0.606
$\text{BERT}_{\text{LARGE}}$ (closed-book)	67.3	45.4	57.6	0.447	0.653	**0.543**
BiDAF++ (Clark and Gardner, 2018)	51.7	30.1	40.9	0.478	0.898	0.688
$\text{BERT}_{\text{BASE}}$, MDS	63.1	39.1	52.0	0.504	0.716	0.603
$\text{BERT}_{\text{BASE}}$, AGG (Maxpool)	67.2	39.1	54.2	0.453	0.701	0.568
$\text{BERT}_{\text{BASE}}$, AGG (GRU)	67.6	41.5	55.4	0.477	0.705	0.583
SAM-Net (Lv et al., 2019)	64.5	40.9	53.5	0.531	0.719	0.619
$\text{BERT}_{\text{LARGE}}$, MDS	67.4	40.1	54.7	0.542	0.738	0.633
$\text{BERT}_{\text{LARGE}}$, Event triples	66.7	45.0	56.6	0.589	0.719	0.649
$\text{BERT}_{\text{LARGE}}$, AGG (Maxpool)	68.8	46.9	58.6	0.476	**0.648**	0.556
$\text{BERT}_{\text{LARGE}}$, AGG (GRU)	69.2	47.5	59.1	0.483	0.655	0.563
$\text{BERT}_{\text{LARGE}}$, AGG (Maxpool), DPR	70.2	47.0	59.4	0.554	0.728	0.635
$\text{BERT}_{\text{LARGE}}$, AGG (Maxpool), BT	70.0	48.0	59.7	**0.444**	0.662	0.545
$\text{BERT}_{\text{LARGE}}$ ++ (integrated)	**70.3**	**48.4**	**60.1**	0.537	0.650	0.589
Human performance$^{(\alpha)}$	74.6	64.9	71.2	-	-	-
Human performance$^{(\beta)}$	81.3	77.4	79.4	-	-	-

Table 3: **Performance of baseline models on FORECASTQA test set.** "yes/no" refers to yes-no questions, and "multi" to multi-choice questions. We test the closed-book setting, and the constrained open-domain setting, where the accessible articles are limited by t_Q, our time constraint. We use BM25 as the article retriever to select top-10 articles, if not particularly specified. "BT" concatenates the binary encoding of date string to an article encoding before aggregation (see Sec. 6.3 "Ablation on Timestamp Modeling"). Human performance is based on the top-10 retrieved articles (α), and Google Search with the question's time constraint (β).

answerability of our questions by providing gold articles instead of retrieved articles (Sec. 6.3).

Evaluation Metrics. Because forecasting is uncertain, a system's prediction probabilities indicate its confidence answering the question. In addition to accuracy, we consider Brier score (Brier, 1950), which measures the mean squared *error* of probabilities assigned to sets of answer choices (outcomes). Formally, Brier $= \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}(p_{ic} - y_{ic})^2$, where p_{ic} is the probability of prediction; y_{ic} is a label indicator for class c of the instance (1 or 0), N is the number of prediction instances, and C is the number of classes (2 or 4). The highest Brier score is 0 (probability 1 for the correct class, probability 0 else), while the worst possible Brier score is 2 (probability 1 for the wrong class, probability 0 else). A confident model gets low Brier scores.

6.2 Human Performance

To benchmark human performance, seven annotators (computer science graduate students) who were not involved in question generation were asked to answer 150 randomly sampled questions from the test set. We consider two scenarios: 1) annotators are provided with retrieved articles, $\overline{A}$; and 2) annotators can access any article published *before the timestamp* via Google Search. Moreover, as annotators live in the "future" with respect to the timestamp of a question, they might already know the actual answer. To avoid the over-estimation of accuracy, we asked the annotators to not use their "future" knowledge. If they felt this is not possible, we asked them to skip the question. On average, 28.3% of questions are skipped. Given this setup, humans achieve 71.2% and 79.4% accuracy respectively, for the two scenarios when taking a majority vote for each question; we also observed good inter-annotator agreement. The two scenarios are referred as "(α)" and "(β)" in Table 3.

Methods	GRU	Maxpool	MDS
$\text{BERT}_{\text{BASE}}$, TF-IDF	53.2	53.9	51.6
$\text{BERT}_{\text{BASE}}$, DPR	53.7	**54.6**	**54.3**
$\text{BERT}_{\text{BASE}}$, BM25	**55.4**	54.2	52.0
$\text{BERT}_{\text{LARGE}}$, TF-IDF	56.5	55.4	**55.0**
$\text{BERT}_{\text{LARGE}}$, DPR	56.1	**59.4**	54.6
$\text{BERT}_{\text{LARGE}}$, BM25	**59.1**	58.6	54.7

Table 4: **Accuracy with different retrievers:** BM25, TF-IDF, and dense passage retrieval (DPR). We test the retrievers with different aggregators: GRU, Maxpool, and MDS.

6.3 Results and Performance Analysis

Results on the Constrained Open-domain Setting. Table 3 shows the results of baseline methods for comparison. We compare pre-trained language models with different context aggregators and other baselines. The integrated model, $\text{BERT}_{\text{LARGE}}$ ++ shows the best performance in terms of accuracy, while $\text{BERT}_{\text{LARGE}}$ (closed-book) shows the best Brier score. Unlike the accuracy metric, the Brier score penalizes over- and under- confident forecasts (Mellers et al., 2014) — thus the best model under each metric can be different. The marginal differences in performance between the two settings suggest that access to information (text evidence) alone does not solve the forecasting problem. We hypothesize an inability to encode salient relations for forecasting purposes prevents the additional information from proving useful. Among the aggregators in $\text{BERT}_{\text{BASE}}$, the GRU aggregator outperforms other aggregators and summarizers. This suggests that utilizing articles' temporal order helps the reasoning. Overall, baselines fall behind human performance by over 10% points given the same retrieved articles.

Study of Different IR Methods. We further test several retrieval methods: BM25 (Robertson et al., 1995; Qi et al., 2019), TF-IDF (Chen et al., 2017a), and a pre-trained dense passage retriever (DPR) (Karpukhin et al., 2020). As in Table 4, $\text{BERT}_{\text{LARGE}}$ with DPR retriever and the Maxpool aggregator shows the best performance than other combinations. However, DPR does not achieve the best accuracy for all methods. This implies that 1)

Methods / Metrics	GRU		Maxpool	
	ACC (↑)	Brier (↓)	ACC (↑)	Brier (↓)
w/o timestamps	**55.4**	**0.583**	54.2	**0.568**
Pre-pend timestamps	54.2	0.634	54.8	0.599
Binary timestamp encoding	51.1	0.623	**55.6**	0.624
Char-RNN timestamp encoding	54.0	0.640	54.3	0.620

Table 5: **Study on modeling article timestamps (publication dates) in the constrained open-domain setting.** We test several methods for temporal modeling and use $\text{BERT}_{\text{BASE}}$ with two different aggregators: GRU and Maxpool.

Methods / Metrics	Accuracy (↑)			Brier score (↓)		
	yes/no	*multi*	*all*	*yes/no*	*multi*	*all*
Random	48.6	25.3	37.8	0.684	0.827	0.750
Question	66.2	41.5	54.7	0.511	0.715	0.606
Article	73.6	80.7	76.9	0.428	0.263	0.351
Evidence sentence	79.9	89.5	84.4	0.355	0.171	0.269

Table 6: **Answerability study on test set.** Instead of retrieved articles, we provide $\text{BERT}_{\text{BASE}}$ with ground-truth context: a gold article or evidence sentence. We thus convert FORECASTQA to a reading comprehension task and examine the answerability of the questions.

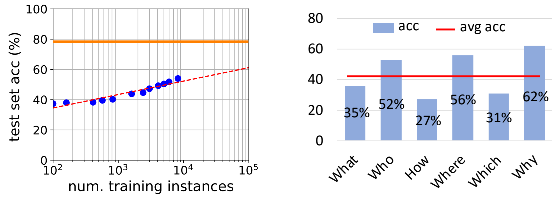
stronger retrieval methods are required to identify useful evidence; 2) complex forecasting abilities may be a bottleneck of current systems.

Ablation on Timestamp Modeling. We conduct an ablation study on modeling time information (publication date) of the retrieved articles, as seen in Table 5. We test: a) pre-pending date string as BERT input, b) using binary encodings of dates[9] and concatenate with article encoding before aggregation, and c) using char-RNN (Goyal and Durrett, 2019) for encoding date string before aggregation[10]. We find that using binary encodings of dates improves the accuracy for the maxpool aggregator. However, the GRU aggregator's accuracy decreases when given date information. We conjecture that our modeling for the time information of each article is not strong enough to help forecasting. We leave more sophisticated modeling for future work.

Answerability of Questions. To validate that the questions in FORECASTQA are indeed answerable, we convert our setup into a machine reading comprehension (MRC) task — find an answer given an assumed appropriate context. We provide the model with a gold article or the evidence sentence (Sec. 4.1). Since pre-trained models have achieved high performance on MRC tasks (Rajpurkar et al., 2016), we expect adequate performance when provided the correct context. As seen in Table 6, we observe that in closed-book setting, BERT is able to beat out a random baseline, but it still does not

[9] https://temporenc.org

[10] Details are described in appendix Sec. E.4



(a) Varying amounts of data. (b) Different question types.

Figure 6: (a) Test accuracy of $\text{BERT}_{\text{BASE}}$ trained with varying amounts of training data, with human performance (79.1%) shown in orange, and (b) development accuracy breakdown by different types of multichoice questions.

perform well; implying our questions are not trivial for BERT, and context is required to answer them correctly. When given the gold article, BERT achieves 76.9% (+22%) and it even performs better (84.4%) given the evidence sentence. This all implies that given the right information, our forecasting questions can be answered correctly.

Study of Data Efficiency. To examine how models might perform with less/more training data, we evaluate $\text{BERT}_{\text{BASE}}$ (closed-book) on the test set, by training it with varying amounts of labeled data. Fig. 6a shows the the resulting "learning curve." We observe the accuracy of the model is "expected" to reach 70%, assuming 100k examples — which is still 9% point lower than human performance.

Results on Different Question Types. We test $\text{BERT}_{\text{BASE}}$ (closed-book) on different question types of multi-choice questions from our development set (Fig. 6b). We find that the accuracy of the model varies across different question types: "*how*" questions are the most difficult to predict while higher accuracy is achieved on "*why*" questions. Also for yes-no questions, the method achieves 69.5% on "*yes*" questions and 62.9% "*no*" questions, indicating that there is no significant bias towards certain type of binary questions.

Error Analysis. We observe 4 main categories of errors produced by the methods in our analysis: (1) retrieving irrelevant articles, (2) incorrect reasoning on relevant evidence, (3) lacking (temporal) common sense, and (4) lacking numerical knowledge. Please refer to Sec. E.7 of appendix for examples and in-depth discussions of these errors.

7 Conclusion

Forecasting is a difficult task that requires every possible advantage to do well. It would be wise to harness this pool of unstructured data for training automatic event forecasting agents. To utilize this form of data for forecasting, we proposed a question-answering task that requires forecasting skills to solve FORECASTQA, and provided the accompanying dataset. Various baseline methods did not perform well, but this is not surprising given the inherent difficulty of forecasting. Our benchmark dataset can benefit future research beyond natural language understanding and hope forecasting performance will be significantly improved.

References

Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. 2015. Icews coded event data. *Harvard Dataverse*, 12.

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017a. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017b. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855, Melbourne, Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963.

Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine.* " O'Reilly Media, Inc.".

Tanya Goyal and Greg Durrett. 2019. Embedding time expressions for deep temporal ordering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4400–4406, Florence, Italy. Association for Computational Linguistics.

Linmei Hu, Juanzi Li, Liqiang Nie, Xiaoli Li, and Chao Shao. 2017. What happens next? future subevent prediction using contextual hierarchical LSTM. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3450–3456. AAAI Press.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018a. Tempquestions: A benchmark for temporal question answering. In *WWW*.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018b. TEQUILA: temporal question answering over knowledge bases. In *Proceedings of the 27th ACM International Conference on Information and*