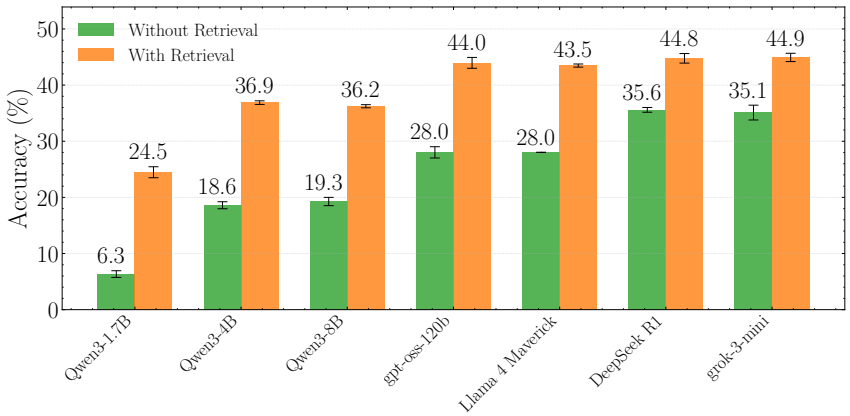


Figure 5: **Retrieval improves accuracy across models.** We use the specialized Qwen3 8B embedding model to retrieve the 5 most relevant chunks (512 tokens) for each question. We take a cautious approach, using articles only until a month before the resolution date.

6 Final Results

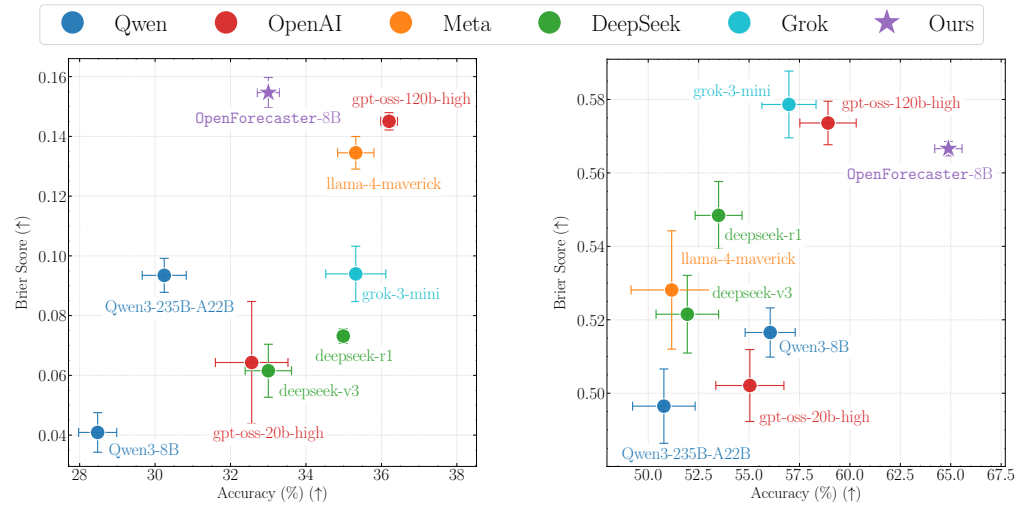
We now present evaluations of our model, `OpenForecaster8B`. To avoid making decisions based on future information, we evaluate on test sets that were not observed until the end.

Open-ended Test Set. Given the lack of *open-ended* forecasting benchmarks that are still "in the future" for our models, we create our own test set with additional steps to ensure high quality. We first use our data creation recipe to generate an initial set of 1,000 questions between May to August 2025 using a stronger model, `o4-mini-high`. We draw from five diverse news sources: Al Jazeera English (global news, based out of Qatar), Time (global news, based out of USA) The Independent (UK focused), Fox News (USA focused), NDTV (India focused), with 200 questions selected from each. The choice of sources was made under the constraint of many established news sources have disallowed crawling of their articles starting 2025. We deliberately use distinct sources from the training set to ensure that our model is learning generalizable forecasting skills, and not source distribution specific biases. Beginning from this initial set of 1000 questions, we perform additional filtering steps to prepare a high-quality test set:

1. We remove any potentially unanswerable questions (noise) by keeping only those which `grok-4.1-fast` could successfully answer with search tool access (85%).
2. To address the issue of late reporting in news outlets, we again use `grok-4.1-fast` with search tool to find the **earliest resolution date** for a given question. This is important to prevent leakage from retrieving articles with the true answer. We retain only those questions with resolution date after May 2025 (64%).
3. Finally, we manually filter the remaining questions to meet our quality checks, resulting in a final test set of 302 questions. We provide more details like news source specific statistics in Section D.3.

External Datasets. We use the FutureX benchmark (Zeng et al., 2025), filtering to non-numeric, English, resolved forecasting questions and evaluating all models with our retrieval. This leaves 86 binary or multiple choice questions, between July to August 2025. For evaluating long-term predictions (without retrieval), we measure consistency metrics on binary questions up to 2028 as proposed by Paleka et al. (2025b), who show they correlate strongly with forecasting performance. Finally, to measure whether our forecasting training generalizes to calibration on standard benchmarks of LLM capabilities, we evaluate without retrieval on a challenging factuality benchmark, SimpleQA (Wei et al., 2024), and popular cross-domain reasoning benchmarks, MMLU-Pro and GPQA-Diamond.

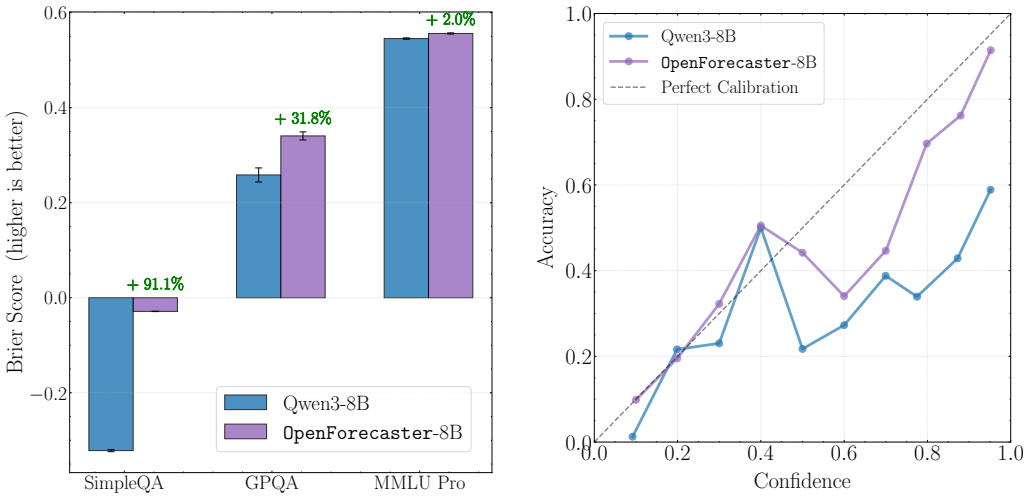
Result 4: Training on our dataset leads to large improvements in forecasting. Figure 6a shows performance of models on our held-out test set of open-ended forecasting questions. On the Brier score (Y axis), the primary metric recommended for forecasting (Tetlock &



(a) Results on `OpenForesight` May-Aug 2025.

(b) Results on FutureX July-Aug 2025.

Figure 6: **Our forecasting training improves accuracy and calibration** both on open-ended questions in our test set, and the external FutureX benchmark. It makes `OpenForecaster8B` competitive with much larger models that have knowledge cutoffs before May 2025.



(a) Brier score on general benchmarks.

(b) Calibration curve on our test set.

Figure 7: Calibration of the models improve significantly after training on `OpenForesight` both on (a) out of distribution benchmarks and (b) on `OpenForesight` test set.

Gardner, 2016), as it measures both accuracy and calibration, `OpenForecaster8B` outperforms even GPT OSS 120B. Our improvements are not merely from calibration, the predictions also become more accurate (X axis), beating Qwen3 235B, but are a bit behind others. Training on `OpenForesight` also improves models from other families like Llama and Gemma as we show in Section B.2. We saw a particularly large (+25% absolute improvement in accuracy) improvement for Llama 3.1 8B Instruct, surpassing the much larger Qwen3-235-A22B. We also show model accuracy by month in Figure 12.

On FutureX, our model has the strongest accuracy by a large margin, even compared to much larger proprietary counterparts. It is close to the best for Brier score as well. Finally, our training leads to more consistent long-term predictions, improving 44% on arbitrage metrics, and 19% on frequentist metrics, with detailed results in Section B.4. Our model also improves on questions from Metaculus prediction market albeit staying behind few larger models which we show in Section B.5.

Result 5: Calibration training for forecasting generalizes to other domains. Figure 7a shows downstream improvements in calibration across SimpleQA, GPQA-Diamond and MMLU-Pro (green text highlighting the relative improvement). This calibration can then be used to reduce hallucinations, for example abstaining on questions the model is not confident about, using simple rules like `if probability < 0.1, replace prediction with "I do not know"`

7 Conclusion

In this paper, we show how to curate data for *scalable training* of *open-ended forecasting*. The results are promising, an 8B model finetuned on our data becomes competitive with proprietary models like GPT-OSS-120B, DeepSeek-R1, and Grok-3-Mini. Calibration improvements from forecasting training generalize out of distribution. A few limitations remain. For example, we only use news to create forecasting questions, which leads to a distributional bias. The news also reports some events late, such as scientific breakthroughs, and this can make such questions easier to "predict" than others in our dataset. This should not affect relative performance comparisons between models though. We also do not consider long-form forecasts, as it is unclear how to grade these. Overall, open-ended forecasting, being a challenging and highly valuable task, offers exciting directions to pursue across research communities. A strong forecaster needs to reason about uncertainty, efficiently seek new information, and make optimal Bayesian updates to its world model, long-standing challenges in the quest for general intelligence. Scaling up end-to-end training of open-ended forecasting systems may lead to emergent improvements in such capabilities. By open-sourcing all our artefacts, we hope to spark more research on this important direction.

Acknowledgments

We thank Douwe Kiela, Alexander Panfilov, Tim Rocktäschel, and Guanhua Zhang for valuable discussions. We thank Maksym Andriushchenko, Arvindh Arun, Alessandro Bifulco, Paras Chopra, and Daniel Paleka for helpful feedback on our draft. We thank CCNews and TheGuardian for providing free access to news articles, Thinking Machines for providing Tinker API research credits, and Contextual AI for letting us test their retrieval system.

References

George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, revised ed. edition, 1976. ISBN 0816211043.

Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. Answer matching outperforms multiple choice for language model evaluation. *arXiv preprint arXiv:2507.02856*, 2025.

Alfred Cowles. Can stock market forecasters forecast? *Econometrica*, 1(3):309–324, 1933.

Hui Dai, Ryan Teehan, and Mengye Ren. Are llms prescient? a continuous evaluation using daily news as the oracle. *arXiv preprint arXiv:2411.08324*, 2024.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training lms to reason about their uncertainty. *arXiv preprint arXiv:2507.16806*, 2025.

Ursula Franklin. *The real world of technology*. House of Anansi, 1999.

Michael M Grynbaum and Ryan Mac. The times sues openai and microsoft over ai use of copyrighted work. *The New York Times*, 27(1), 2023.

Yong Guan, Hao Peng, Xiaozhi Wang, Lei Hou, and Juanzi Li. Openep: Open-ended future event prediction. *arXiv preprint arXiv:2408.06578*, 2024.