mixed setting improve accuracy and Brier, with the *binary+free-form* mix offering the best overall trade-off across testing formats. Our gains by training on binary-only format are consistent with prior work by Turtel et al. (2025b;a). However, we do not arrive at a single unanimous recipe: free-form data is essential for open-ended forecasting, while combining formats appears Pareto-optimal across binary and free-form evaluations. Practically, it seems training on a *mixture* of question styles provides the most robust gains across tasks.

B.2 Varying models and evaluation months

In Figure 10 we observe that while the first few article chunks that are retrieved lead to large improvements, at around five articles, improvements plateau, both on our Qwen3-8B and also other large models like GPT-OSS-120B. Thus, unless otherwise specified, we use 5 articles for all evaluations and training in this work.

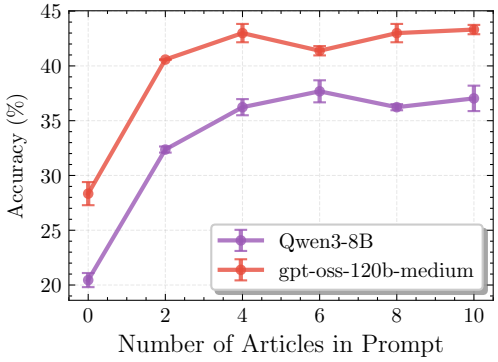


Figure 10: **Improvements from retrieval plateau at $\sim$ 5 chunks.** We show the accuracy of both a large GPT-OSS-120B model and a small Qwen3-8B model which we finetune.

Improvement on non-Qwen models. Our training data `OpenForesight` can be used to improve models across different families. In Figure 11 we show improvements for Llama-3.1-8B-Instruct, Llama-3.2-3B-Instruct and Gemma-3-4B-Instruct. We see particularly large improvements in both accuracy and Brier score for Llama due to both: poor initial performance, but also surprising amenability to RL training with our data as the final performance exceeds much larger models like Qwen3-235B-A22B and DeepSeek-v3.

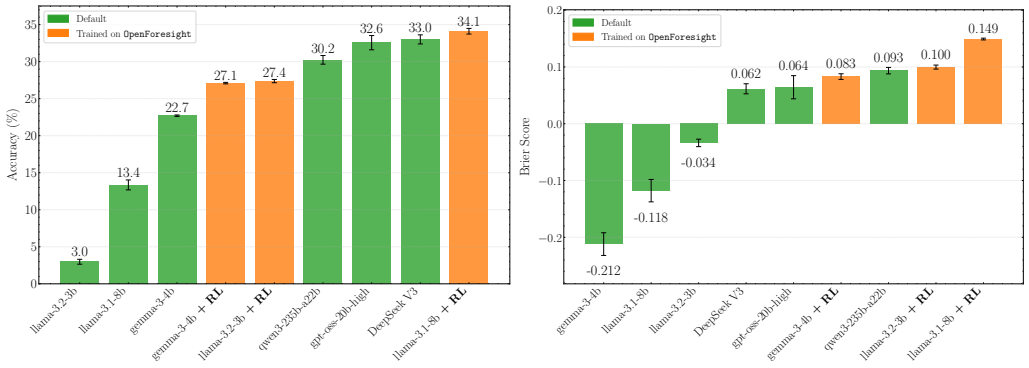


Figure 11: Performance of models from Llama and Gemma family on our test set.

Results over time. As our test set is derived from articles from May to August 2025, so we split the questions by resolution date to get monthly performance of the models. Breaking down by month, our test has 94 questions resolving in May, 85 in June, 76 resolving in July and 47 resolving in August. We have lower number of questions in later months due to our filtering strategy for addressing late reporting in news. We first generated roughly equal number of questions per month and post-hoc filtered the ones whose true resolution date (found using grok-4.1-fast with search) was before May'25.

For monthly performance, our hypothesis is that as we go further into the future, forecasting should become more difficult leading to lower performance. In Figure 12 and Figure 13, we find that the accuracy and brier score of the models indeed drops gradually month-by-

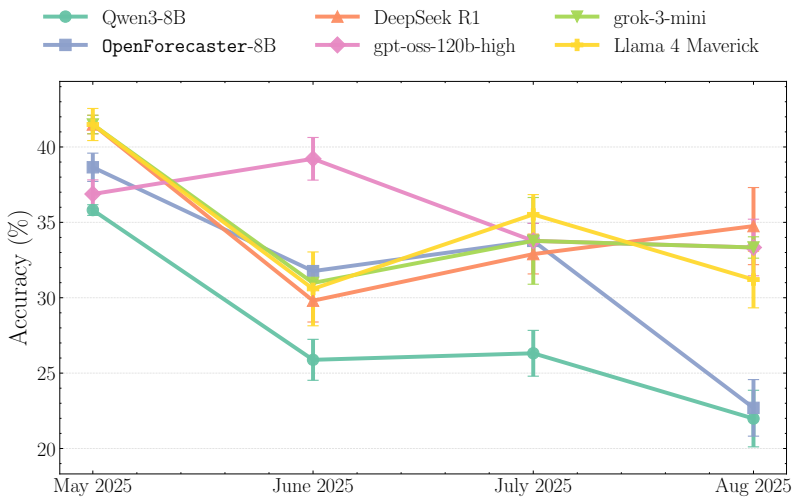


Figure 12: Monthly accuracy of the models on our test set.

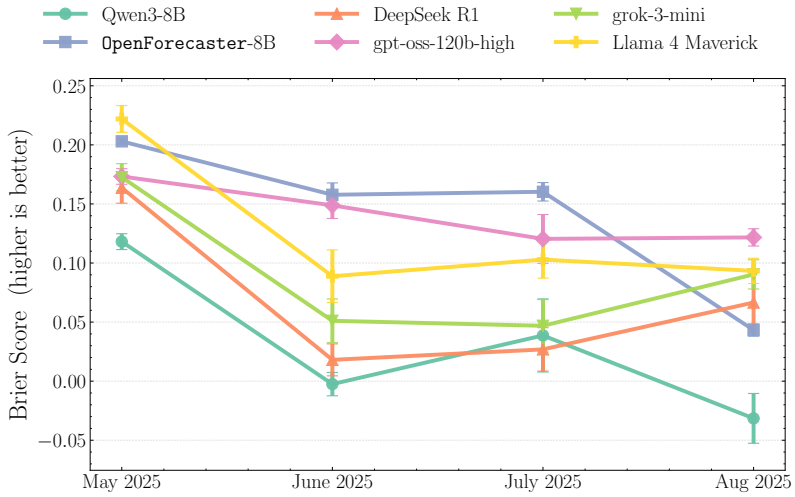


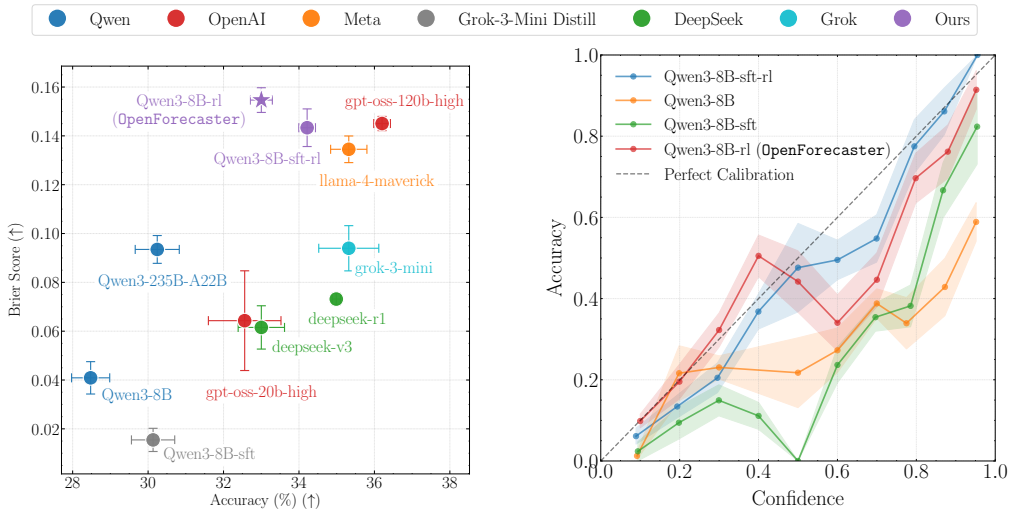
Figure 13: Monthly Brier score of the models

month consistent with our hypothesis. We also find that our trained models are consistently better than the original versions and also better than all other models in Brier score.

B.3 Ablation with Supervised Finetuning

Here we study what would be the benefit if we add a supervised finetuning (SFT) stage in our training process? While we start from the RL trained Qwen3 thinking models, they are far behind proprietary models as shown in Section 5. Several frontier model training reports (Guo et al., 2025) mention using an SFT stage as a warm start before RL. We choose Grok-3-Mini to generate forecasting reasoning traces for SFT, as it has high performance, low cost, and provides the full reasoning trace through the API. Specifically, we construct a dataset of 10,000 questions from *The Guardian* dated January–March 2025, beyond Grok-3-mini's reported knowledge cutoff of June 2024. Obtaining Grok-3-Mini's reasoning traces on this data costed 15 USD. For distillation, we randomly choose the number of articles to put in its prompt (0 to 10) so that the student model can reason with any number of articles.

Finally, we train this distilled checkpoint with GRPO using same data mixture, reward design and configurations as we used in training `OpenForecaster-8B`. We report the results for both just distillation (Qwen3-8B-sft) and RL on the SFT checkpoint (as Qwen3-8B-sft-rl)



(a) Results on our curated test set.

(b) Calibration curve of the models on our test set.

Figure 14: RL training on `OpenForesight` improves the SFT models on accuracy. In particular, the calibration of Qwen-8B-sft-rl model (blue line) is near perfect.

in Figure 14a on our curated test set from May to August 2025. The final model achieves higher accuracy and much better calibration albeit with a slightly lower brier score.

B.4 Consistency Evaluation

Table 2: **Consistency checks before and after RL training**. We report average violation scores and relative changes (negative percentages indicate improvements, positive indicate regressions). The RL-trained model shows improvements in most areas.

	Arbitrage			Frequentist		
Check	Qwen3-8B	`OpenForecaster-8B`	Δ	Qwen3-8B	`OpenForecaster-8B`	Δ
PARAPHRASE	0.030	0.020	-33%	0.157	0.131	-17%
CONSEQUENCE	0.010	0.003	-67%	0.048	0.029	-39%
ANDOR	0.033	0.025	-25%	0.205	0.187	-9%
AND	0.016	0.004	-75%	0.063	0.037	-42%
NEGATION	0.043	0.063	+46%	0.198	0.271	+37%
OR	0.022	0.019	-12%	0.094	0.120	+28%
BUT	0.040	0.027	-31%	0.234	0.202	-14%
COND	0.039	0.033	-15%	0.227	0.222	-2%
CONDCOND	0.036	0.035	-3%	0.256	0.258	+1%
EXPEVIDENCE	0.041	0.034	-18%	0.240	0.195	-19%
Aggregated	**0.031**	**0.026**	**-15%**	**0.172**	**0.165**	**-4%**

Paleka et al. (2025b) release a dataset of long-term forecasting questinos set to resolve up to 2028, showing language models exhibit inconsistencies in their probabilistic predictions. To evaluate consistency, they propose ten consistency checks measuring both arbitrage and frequentist violations.

We evaluate Qwen3-8B and our trained model on the dataset created by Paleka et al. (2025b). We measure performance of the models on all consistency check tuples proposed by them. Table 2 compares the baseline Qwen3-8B with our RL-trained model. The results show improvements across most consistency checks. We observe strong gains in Boolean logic operations (AND: 75% reduction in arbitrage violations, 42% in frequentist) and consequence checks (67% and 39% reductions respectively) but also some regressions, particularly in negation consistency. Overall, our training achieves a 15% reduction in arbitrage violations and 4% reduction in frequentist violations.