

Category	Brier Score ↓		Accuracy ↑	
	Ours	Crowd	Ours	Crowd
Science & Tech	.143 _{.011}	.114 _{.011}	82.2 _{2.7}	84.3 _{2.6}
Healthcare & Biology	.074 _{.015}	.063 _{.020}	93.8 _{4.3}	90.6 _{5.2}
Economics & Business	.198 _{.007}	.147 _{.009}	68.8 _{2.1}	78.3 _{1.9}
Politics & Governance	.172 _{.006}	.145 _{.007}	72.6 _{1.4}	78.2 _{1.3}
Education & Research	.163 _{.024}	.129 _{.024}	80.6 _{6.7}	77.8 _{7.0}
Arts & Recreation	.221 _{.010}	.146 _{.010}	62.4 _{2.5}	76.9 _{2.2}
Security & Defenses	.174 _{.008}	.129 _{.009}	71.0 _{2.1}	78.4 _{1.9}
Sports	.175 _{.004}	.171 _{.005}	73.0 _{1.3}	73.1 _{1.3}
All Categories	.179_{.003}	.149_{.003}	71.5_{.7}	77.0_{.7}

Platform	Brier Score ↓		Accuracy ↑	
	Ours	Crowd	Ours	Crowd
Metaculus	.134 _{.005}	.104 _{.005}	80.3 _{1.2}	86.6 _{1.1}
GJOpen	.193 _{.011}	.157 _{.013}	67.9 _{3.4}	72.6 _{3.2}
INFER	.247 _{.053}	.310 _{.086}	60.0 _{13.1}	53.3 _{13.3}
Polymarket	.172 _{.005}	.127 _{.006}	73.6 _{1.3}	79.9 _{1.1}
Manifold	.219 _{.004}	.200 _{.005}	63.6 _{1.3}	67.9 _{1.3}
All Platforms	.179_{.003}	.149_{.003}	71.5_{.7}	77.0_{.7}

Table 4: **Results of system evaluation** by category (**left**) and by platform (**right**). Subscript numbers are 1 standard error. Averaged across all retrieval dates, our optimal system, as described in [Section 4](#), achieves .179 Brier score (human crowd: .149) and accuracy .715 (human crowd: .770).

5.5%. In comparison with the baseline evaluation ([Section 3.4](#)), our system’s Brier score (.179) significantly outperforms the best baseline model (.208 with GPT-4-1106-Preview)

In prior work, [Zou et al. \(2022\)](#) evaluated their system on the forecasting dataset Autocast, which consists of questions from 3 of the platforms we use: Metaculus, INFER, and GJOpen. They achieved an accuracy of 65.4% compared to a community baseline of 92.8%. [Yan et al. \(2024\)](#) later improved this to 67.9%. Our results ([Table 4](#)) underscore the significant progress we make in automated forecasting—specifically, we achieve a better accuracy (71.5%) even though the questions we consider are harder (with a significantly lower crowd accuracy: 77.0%).

Further detailed results across different platforms and categories can be found in [Table 4](#). Across categories, our system exhibits noticeable variations: on Sports, our system nearly matches the crowd aggregate, and on Environment & Energy, it falls much behind. However, we caution against drawing strong conclusions from subcategories, since the sample size is smaller and variation could be due to noise.

Finally, on the test set, we observe again that our system is well calibrated ([Figure 3c](#)) with RMS calibration error .42 (human crowd: .38). Interestingly, this is not the case in the baseline evaluations ([Section 3.4](#)), where the models are *not* well calibrated in the zero-shot setting ([Figure 3a](#)). Through fine-tuning and ensembling, our system improves the calibration of the base models, without undergoing specific training for calibration.

6.2 System Strengths and Weaknesses

We next seek to understand our system’s strengths and weaknesses. We will investigate these on the validation set, and later use these insights to improve performance on the test set ([Section 6.3](#)).

We find that our system performs best relative to the crowd on the validation set when (1) the crowd is less confident, (2) at earlier retrieval dates, and (3) when it retrieves many articles. Furthermore, we find that our system is well-calibrated.

First, our system significantly outperforms the crowd when the crowd’s predictions express high uncertainty. Specifically, when the crowd’s predictions are between .3 and .7, our Brier score is .199 compared to the crowd’s .246. However, our system underperforms the crowd on questions where they are highly certain, likely because it rarely outputs low probabilities ([Figure 4b](#)). We hypothesize that this stems from our model’s tendency to hedge predictions due to its safety training (see [Figure 17](#) for a qualitative example). Supporting this, our system achieves 7% higher accuracy on questions where the crowd’s prediction is within .05 of 0 or 1, but the Brier score is worse by .04.

Next, our system outperforms the crowd on earlier retrieval dates (1, 2, and 3) but not the later ones (4 and 5). Relative to the crowd, our Brier score improves at a slower rate as questions move towards their

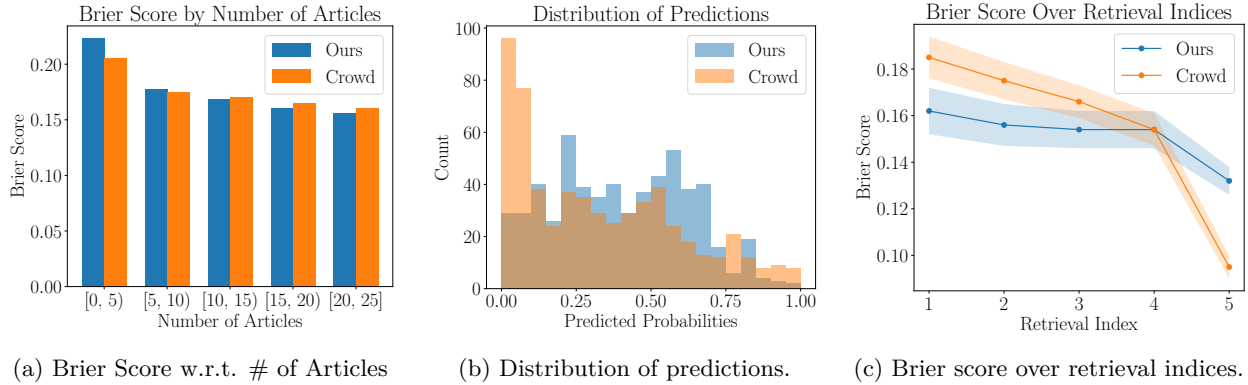


Figure 4: **System strengths.** Evaluating on the validation set, we note: **(a)** When provided enough relevant articles, our system outperforms the crowd. **(b)** For questions where the crowd is unsure (predictions between .3 and .7), we outperform them (Brier score .199 vs. .246). However, the crowd outperforms our system on questions where they are highly confident, e.g. predicting less than .05. **(c)** Our system’s Brier score is better at the earlier retrieval dates. Finally, our system is well-calibrated (c.f. Figure 3b).

resolution (Figure 4c). This may be due to the aforementioned issue: Our model hedges, even as the evidence becomes more decisive.

With respect to retrieval, our system nears the performance of the crowd when there are at least 5 relevant articles. We further observe that as the number of articles increases, our Brier score improves and surpasses the crowd’s (Figure 4a). Intuitively, our system relies on high-quality retrieval, and when conditioned on more articles, it performs better.

Our system is well calibrated on the validation set, with most of the calibration error coming from the system’s underconfidence: predictions near 0 are observed to occur less frequently than anticipated, and similarly, events with predictions close to 1 also occur at a higher rate than the model suggests (Figure 3b).

6.3 System Beats Crowd in the Selective Setting

In real-word forecasting competitions, forecasters do not have to make predictions on every question in the platform at every possible date. Instead, they typically make predictions on questions where they have expertise or interest in and at times that they choose. Therefore, it is natural to leverage our system’s strengths and weaknesses and decide accordingly if we should forecast on a retrieval date k for a question q .

Leveraging the insights from Section 6.2, we outperform the crowd by making selective forecasts. Specifically, we report the performance when forecasting only under the conditions identified in Section 6.2:

1. *Forecasting only on questions when the crowd prediction falls between .3 and .7.* Here, our system attains a Brier score of **.238** (crowd aggregate: **.240**). This comprises 51% of forecasts and 56% of questions.
2. *Forecasting only on earlier retrieval dates (1, 2, and 3).* Our system’s Brier score in this setting is **.185** (crowd aggregate: **.161**). This comprises 66% of forecasts and 100% of questions.
3. *Forecasting only when the retrieval system provides at least 5 relevant articles.* Under this condition, our system’s Brier score is **.175** (crowd aggregate: **.143**). This makes up 84% of forecasts and 94% of questions.
4. Under all three conditions, our system attains Brier score **.240** (crowd aggregate: **.247**). This comprises 22% of forecasts and 43% of questions.

The gap in Brier score between our system and the crowd shrinks under each heuristic, except the third one (Table 3). Under the first heuristic, we outperform the crowd by a small margin (.238 vs. .240). This is

Criteria	Brier Score ↓		% Accuracy ↑	
	Ours	Aggregate	Ours	Aggregate
Full System	.179 _{.003}	.146_{.002}	71.5 _{.7}	77.8_{.6}
Fine-tuned GPT-4-0613	.182 _{.002}	.146_{.002}	70.7 _{.7}	77.4_{.6}
Fine-tuned GPT-3.5 & Base GPT-4	.181 _{.002}	.147_{.002}	70.9 _{.7}	77.4_{.6}
Fine-tuned GPT-3.5	.183 _{.002}	.146_{.002}	71.5 _{.7}	77.4_{.6}
Base GPT-4	.186 _{.002}	.148_{.002}	70.6 _{.7}	77.1_{.6}
Base GPT-4; no IR	.206 _{.002}	.150 _{.002}	66.6 _{.7}	76.9 _{.6}

Table 5: **Ablation study results.** The crowd Brier score and accuracy are .146 and 77.0%, respectively. “Aggregate” indicates the weighted average of our system with the crowd prediction. Our full system uses fine-tuned GPT-4-0613 and base GPT-4-1106-Preview (**row 1**). The system yields similar performance with fine-tuned GPT-3.5 (**rows 3–4**). Our system exhibits poorer performance without a fine-tuned reasoning model (**row 5**), and further declines with neither retrieval nor a fine-tuned reasoning model (**row 6**). Subscript numbers represent one standard error. We bold entries that surpass the crowd aggregate.

valuable as our system can be used to complement the crowd’s prediction when there is greater uncertainty. When all three conditions are jointly met, our system beats the crowd significantly (by more than 1.5 standard errors in both Brier score and accuracy).

6.4 System Complements the Crowd

Finally, we show that aggregates of our system with the crowd forecasts outperform either one in isolation.

Combining the system’s predictions with the crowd using a weighted average—4x weight for the crowd, which we find optimal on the validation set—improves the overall Brier score from .149 to .146 on the full test set (Table 3, top row).

Moreover, our system excels under certain criteria (Section 6.2). It is especially useful in these cases to supplement the crowd prediction. We report these results in Table 3 as well, using an unweighted average (instead of the weighted average above). This outperforms the crowd prediction in all cases: For example, the crowd Brier score is .24 when the prediction is between .3 and .7, while the system achieves .237.

Finally, beyond direct score improvements, our system can potentially aid human forecasters by providing effective news retrieval and novel perspectives in reasoning drawn from LM pre-training knowledge. We leave it as a future direction to explore how our system can interactively assist human forecasters.

7 Ablations

We conduct 3 ablation studies. The first validates that our performance is not solely due to the power of GPT-4. The last two show the benefits of our retrieval and fine-tuning methods.

Fine-tuning a less capable model. To demonstrate that our system’s performance does not hinge on the ability of the base model (i.e., GPT-4), we fine-tune GPT-3.5 on all our fine-tuning data (13,253 samples).

We replace fine-tuned GPT-4 in our system with fine-tuned GPT-3.5, and evaluate using the same methodology as in Section 6.1. We find here that our Brier score is only slightly worse: .182 compared to the previous score of .179.

No fine-tuning. To demonstrate the gain from fine-tuning (Section 5.1), we evaluate our optimal system, except we only use base GPT-4-Preview-1106 as the reasoning model.

In this setup, the ablated system achieves a Brier score of .186, which increased on the original score by .007.

Overall, the results suggest that fine-tuning the reasoning model yields a significant boost to our system’s performance.