Appendix

Contents

A Adapting Brier Score to free-form responses

Let $\mathcal{X}$ be the set of open-ended forecasting questions; and $\mathcal{Y}$ the set of short textual answers. Let $x \in \mathcal{X}$ be a resolved forecasting question and $y^\star$ be the ground truth answer (as the question has already resolved). We ask the forecaster to respond with its best guess answer y, and the probability q they assign to that being the true outcome. We evaluate this prediction tuple $<y, q>$ using the Brier score (Mucsányi et al., 2023) but adapt it to our setting. For a K-class outcome space $\mathcal{Y}$ with reported distribution q and true class y^*, the (multi-class) Brier score is

$$S(q, k) = -\sum_{y \in \mathcal{Y}} (q_y - k_y)^2 = -(q_{y^*} - 1)^2 - \sum_{y \neq y^*} q_y^2,$$

where k is the one-hot encoding with $k_{y^*} = 1$. In our open-ended setting, $\mathcal{Y}$ is not predefined but rather its instances are provided by the forecaster. For simplicity, we elicit only a **single guess** y with probability $q \in [0, 1]$ and assume the forecaster's probability is 0 for all other (semantically different) answers $y' \neq y$.[1] Applying the multi-class Brier scoring rule in such a case induces a simplified score:

$$S(q, y, y^*) = \begin{cases} -(q-1)^2 - 0 = -1 + 2q - q^2, & \text{if } y \equiv y^*, \\ -(0-1)^2 - q^2 = -1 - q^2, & \text{if } y \not\equiv y^*. \end{cases}$$

Dropping the constant -1 yields

$$S'(q, y, y^*) = \begin{cases} 1 - (q-1)^2, & \text{if } y \equiv y^*, \\ -q^2, & \text{if } y \not\equiv y^*, \end{cases}$$

which shifts the range from $[-2, 0]$ to $[-1, 1]$ while providing a more natural interpretation: predicting $q = 0$ gives a baseline 0 regardless of y; correct answers receive positive scores, incorrect answers negative scores; and magnitude scales quadratically with confidence. We report S'score (visualized in Figure 8) as the *Brier score* in this paper.

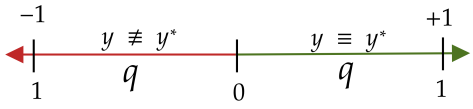


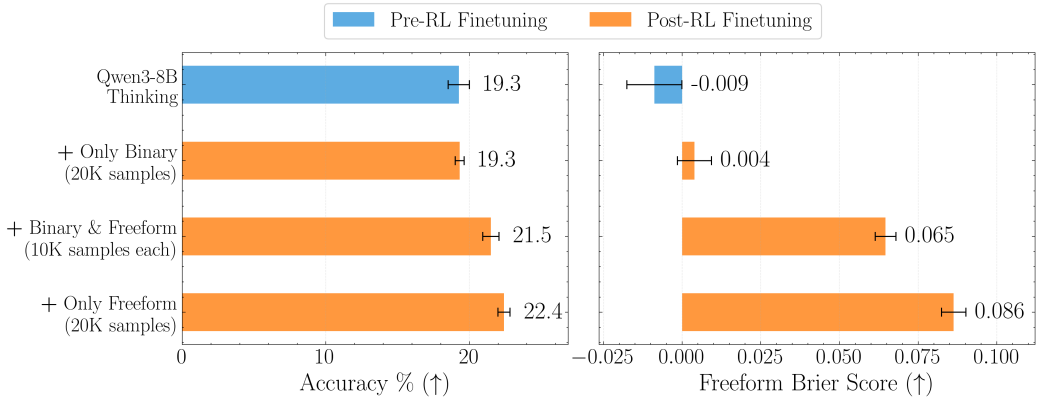
Figure 8: Illustration of the Brier score when adapted to free-form response with answer y and probability q.

Recent work by Damani et al. (2025) shows that this metric is a proper scoring rule, incentivizing both high accuracy and truthful reporting of probability on the answer that seems most likely. However, note that what we call the Brier score here is distinct from the Brier score considered by Damani et al. (2025). Their Brier score is the one traditionally used for evaluating binary outcomes while ours is for free-form responses. Yet, we can show that our Brier score is *same* as the **training reward** considered by them.

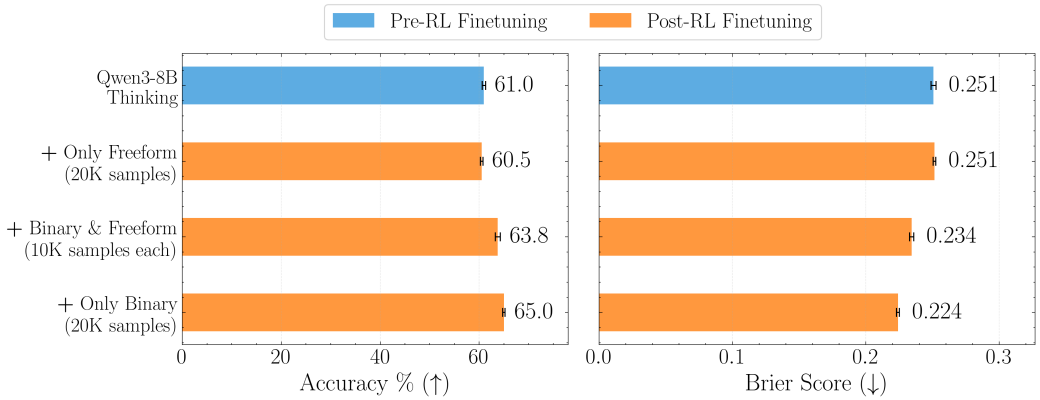
[1] This is technically incorrect to assume as the forecaster may have non-zero probability for guesses other than y. Ideally the forecaster should report all its guesses which have non-zero probability (with the multi-class Brier scoring rule still being applicable) but we leave exploring this direction for future work.

B Additional Results

B.1 Ablation: Using Prediction Market Binary Data



(a) Performance on our **Validation Set** composed of question from TheGuardian news source from July 2025.



(b) Performance on **Metaculus binary** questions resolved in May–July 2025.

Figure 9: **Performance of different data ablations.** We evaluate performance after training on 3 different supervision signals: (i) only binary data (20K samples), (ii) only freeform data (20K samples), and (iii) both binary and freeform data (10K samples each) for data-matched comparison. (a) Accuracy and freeform Brier score of the initial and post-RL model on our Validation Set from July 2025. (b) Accuracy and binary Brier score of initial and post-RL model on volume-filtered binary questions resolved between May and July 2025 on Metaculus. *We find training on binary questions hurts performance on open-ended forecasting, but is necessary to retain performance on binary prediction market questions.*

We ablate supervision type with `Qwen3-8B` using three size-matched settings (Figure 9). For *binary-only*, we curate **20K** resolved markets from Manifold, volume-filtered to ensure engagement; because many markets resolve slowly, this set spans the past five years. For *free-form only*, we use **20K** pipeline-generated, usable questions from Forbes articles. For the *binary+free-form mix*, we take **10K** Manifold + **10K** Forbes questions to keep total examples constant. The goal is to isolate which *learning signal*—binary resolution vs. open-ended outcome specification—most effectively trains calibrated forecasters under identical compute and token budgets.

On the free-form test set (Fig. 9 Left), post-RL performance improves most with *free-form only* supervision (Accuracy 19.3% → 22.4%; Free-form Brier −0.009 → 0.086). Mixing binary and free-form also helps (Brier 0.065), whereas *binary-only* yields minimal gains on free-form evaluation (Brier 0.004). On Metaculus (binary) (Fig. 9 Right), both *binary-only* and the