Scattered Hypothesis Generation for Open-Ended Event Forecasting

He Chang[♠], **Zhulin Tao** [♠*], **Lifang Yang**[♠], **Xianglin Huang**[♠], **Yunshan Ma**[♢],
[♠]Communication University of China, [♢]Singapore Management University,
{hechangcuc,yanglifang,huangxl}@cuc.edu.cn
taozhulin@gmail.com, ysma@smu.edu.sg

Abstract

Despite the importance of open-ended event forecasting for risk management, current LLM-based methods predominantly target only the most probable outcomes, neglecting the intrinsic uncertainty of real-world events. To bridge this gap, we advance open-ended event forecasting from pinpoint forecasting to *scatter forecasting* by introducing the proxy task of hypothesis generation. This paradigm aims to generate an inclusive and diverse set of hypotheses that broadly cover the space of plausible future events. To this end, we propose SCATTER, a reinforcement learning framework that jointly optimizes inclusiveness and diversity of the hypothesis. Specifically, we design a novel hybrid reward that consists of three components: 1) a *validity reward* that measures semantic alignment with observed events, 2) an *intra-group diversity reward* to encourage variation within sampled responses, and 3) an *inter-group diversity reward* to promote exploration across distinct modes. By integrating the validity-gated score into the overall objective, we confine the exploration of wildly diversified outcomes to contextually plausible futures, preventing the mode collapse issue. Experiments on two real-world benchmark datasets, i.e., OpenForecast and OpenEP, demonstrate that SCATTER significantly outperforms strong baselines. Our code is available at https://github.com/Sambac1/SCATTER.

1 Introduction

Event forecasting plays a critical role in risk management, public policy, and strategic decision-making by enabling proactive responses to future socio-economic and geopolitical developments rather than reactive measures (Lee et al., 2025; Keivabu, 2019; Zou et al., 2022). With recent advances in large language models (LLMs) (DeepSeek-AI, 2025; OpenAI, 2023) and post-training techniques, LLM-based event forecasting has emerged as a promising approach. Existing research (Deng et al., 2024; Chang et al., 2024; Schoenegger et al., 2025) typically falls into two primary paradigms: discriminative forecasting and open-ended forecasting. The former (Wildman et al., 2025; Karger et al., 2025; Yang et al., 2025) typically formulates forecasting as a closed-set decision problem such as multiple-choice questions or binary predictions. In contrast, open-ended forecasting (Wang et al., 2025; Zeng et al., 2025) allows LLMs to generate free-form natural language predictions conditioned on historical context. Owing to its capacity to model complex event dynamics and provide detailed generative insights, open-ended event forecasting has recently garnered substantial attention.

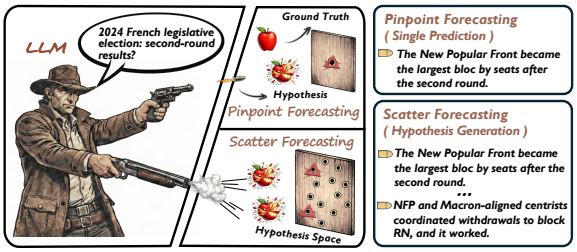

Figure 1: The analogy between shooting and event forecasting: pinpoint forecasting (single prediction) versus scatter forecasting (hypothesis generation).

However, existing open-ended event forecasting methods (Guan et al., 2024; Zhang et al., 2024; Yu et al., 2025b) largely adhere to a single prediction paradigm. As shown in the analogy of Figure 1, this paradigm resembles pinpoint shooting with a single bullet, where the forecasting process yields only one most probable outcome. However, in the real world, the possible future developments of an ongoing event often involve multiple plausible events rather than a single outcome. Consequently, pinpoint forecasting, which restricts outputs to a

[1]Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 17288–17304

narrow subset of plausible futures, fails to capture the intrinsic uncertainty of real-world events [1]. To bridge this gap, we introduce *scatter forecasting*. Analogous to scatter shooting, where each shot covers an area (i.e., a set of points), scatter forecasting produces a set of outcomes, each representing a hypothesis about the future event. Accordingly, open-ended event forecasting can be reformulated as the proxy task of *hypothesis generation*: given a certain context, an LLM-based forecasting model generates a set of distinct hypotheses that maintain both *inclusiveness* and *diversity*, thereby covering the space of plausible future events.

Although hypothesis generation is promising, simultaneously ensuring both inclusiveness and diversity of generated hypotheses remains challenging. A straightforward approach (Schoenegger et al., 2024; Halawi et al., 2024; Guan et al., 2024) is to prompt LLMs to generate multiple non-redundant hypotheses; however, most existing LLMs, primarily trained via supervised fine-tuning, struggle to consistently produce distinct yet valid long-tail events (Kirk et al., 2024; Kandpal et al., 2023). Alternatively, reinforcement learning (RL) (Sutton and Barto, 1998; Ouyang et al., 2022a; Schulman et al., 2017), such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024; Yu et al., 2025a), appear to be a promising solution. Nevertheless, optimizing for both inclusiveness and diversity is inherently conflicting. On the one hand, emphasizing inclusiveness tends to favor conservative hypotheses and is prone to mode collapse (Yue et al., 2025), ultimately reducing the output to a single outcome. On the other hand, emphasizing diversity, in contrast, often yields noisy gradients and unstable optimization, thereby hampering convergence (Yao et al., 2025). This tension is further exacerbated by the irreversibility of real-world events and the scarcity of counterfactual data, which hinder the acquisition of high-quality training data (Shumailov et al., 2024).

To address these challenges, we propose **SCATTER**: a reinforcement learning framework for scattered hypothesis generation. Based on GRPO, SCATTER jointly optimizes inclusiveness and diversity with a hybrid reward that confines exploration to contextually plausible futures. Given the absence of objective gold standards and the sparsity of observable real-world events, we first introduce a validity reward based on embedding similarity. This serves as a coarse-grained yet effective semantic proxy to align generated hypotheses with available factual evidence. Conditioned on validity-gated score, we design two diversity rewards: an intra-group reward to encourage variation within individual samples and an inter-group reward to promote exploration across distinct plausible modes and mitigate mode collapse. By down-weighting diversity rewards for non-factual hypotheses, our approach prevents the reward hacking triggered by trivial novelty, thereby anchoring the learning process toward a diverse yet plausible hypothesis space. We conduct extensive experiments on two real-world benchmark datasets OpenForecast (Wang et al., 2025) and OpenEP (Guan et al., 2024), and the results demonstrate that our method significantly outperforms strong baselines. Our contributions are summarized as follows:

- We formulate open-ended event forecasting as a *hypothesis generation* problem, advancing from single prediction to scatter forecasting.
- We propose **SCATTER**, a reinforcement learning framework that jointly optimizes inclusiveness and diversity via a hybrid reward design.
- Empirical results demonstrate that SCATTER significantly outperforms strong baseline methods.

2 Related Work

2.1 LLM-Based Event Forecasting

Event forecasting (Arrow et al., 2008; Ma et al., 2023b,a) aims to predict future outcomes from historical context and has traditionally been approached with structured statistical and neural methods. More recently, LLMs have been integrated directly into forecasting pipelines (Chang et al., 2025; Li et al., 2024; Su et al., 2024). Broadly, prior work (Halawi et al., 2024; Jin et al., 2021; Li et al., 2026) falls into two paradigms: discriminative forecasting and generative forecasting. Discriminative approaches (Chang et al., 2024; Karger et al., 2025) cast forecasting as closed-set prediction (e.g., binary or multiple-choice). While competitive on constrained question types, these methods inherently restrict forecasts to a predefined outcome space. Generative approaches (Wang et al., 2025; Zhang et al., 2024), in contrast, treat forecasting as open-ended prediction, allowing LLMs to produce free-form descriptions of future events conditioned on context. However, most existing methods (Zeng et al., 2025; Guan et al., 2024)

[1] Further evidence can be seen in Section 5.4.

primarily emphasize dataset construction and baseline evaluation pipelines, with modeling objectives still largely restricted to generating a single prediction per query. In practice, real-world dynamics are complex and multi-modal, and a single deterministic prediction cannot capture the range of plausible future trajectories. These considerations motivate reframing open-ended event forecasting from single prediction to hypothesis generation.

2.2 Post-Training for LLMs

Post-training methods (Ding et al., 2023; Chu et al., 2025), including supervised fine-tuning (SFT), reinforcement learning (Rafailov et al., 2023) and preference-based optimization (Ouyang et al., 2022b), are widely used to steer pretrained LLMs toward desired downstream behaviors. The rapid advancement of reinforcement learning (Guo et al., 2025), especially GRPO, has garnered significant attention, particularly in domains such as mathematics and coding. However, while RL enhances sampling efficiency towards correct paths, current training paradigms rarely elicit fundamentally new reasoning patterns, rendering models prone to mode collapse (Yue et al., 2025). This tendency is particularly detrimental in open-ended forecasting, where practitioners require a diverse set of credible scenarios rather than a single canonical prediction. Although recent research (Chen et al., 2025; Zhou et al., 2025) has begun to investigate exploration within RL, it remains confined to verifiable domains. These approaches are insufficient for open-ended forecasting because they rely on deterministic verification signals (e.g., compilers or solvers) to prune search spaces, signals that are inherently absent in forecasting tasks where ambiguity prevails and *correctness* is probabilistic rather than binary.

3 Preliminary

3.1 Problem Formulation

We reformulate open-ended event forecasting as a hypothesis generation task to better accommodate the intrinsic uncertainty of the real world. Formally, let $\mathcal{D} = \{(\mathcal{C}_i, \mathcal{Q}_i, \mathcal{G}_i)\}_{i=1}^{N}$ denote a dataset of N samples. For each sample, the input context $x = (\mathcal{C}, \mathcal{Q})$ comprises the historical background $\mathcal{C}$ and the specific forecasting question $\mathcal{Q}$, while $\mathcal{G}$ represents the ground-truth outcome. See Appendix A.8 for a detailed example. We aim to train an LLM policy π_θ to serve as the event forecaster. Given a query context x, the policy samples a set of responses $\{y_k\}_{k=1}^{K}$, where each response $y_k = \{h_{k,1}, \ldots, h_{k,M}\}$ contains M hypotheses describing plausible future events.

3.2 Optimization with GRPO

We optimize the policy π_θ using GRPO. Unlike standard PPO, which requires a parametric value function, GRPO utilizes the mean reward of a sampled group as the baseline, reducing computational overhead. Specifically, for each context x, we sample a group of G responses $\{y_1, \ldots, y_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$. The optimization objective is defined as follows:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x,\varepsilon}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\min\Big(\rho_i\hat{A}_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)\hat{A}_i\Big) - \beta\mathbb{D}_{\text{KL}}(\pi_\theta||\pi_{\text{ref}})\Bigg], \tag{1}$$

where $\rho_i(\theta) = \frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}$ is the importance sampling ratio. The advantage $\hat{A}_i$ is computed by normalizing the rewards within the group:

$$\hat{A}_i = \frac{r(y_i, x) - \mu_r}{\sigma_r + \delta}, \tag{2}$$

where $r(y_i, x)$ is the reward for the response y_i, and μ_r, σ_r denote the mean and standard deviation of the rewards within the sampled group, respectively. δ is a small constant for numerical stability.

4 Our Approach: SCATTER

We present our proposed approach SCATTER, which is a hybrid reward mechanism that aims to maximize diversity while maintaining inclusiveness, illustrated Figure 2. Specifically, SCATTER consists of a validity reward and a validity-gated diversity mechanism. This architecture employs the validity signal to ensure factual alignment, while the gate regulates intra-group and inter-group diversity rewards to prevent reward hacking.

4.1 Validity Reward

To quantify the semantic correctness and alignment of the generated hypotheses, we define a validity score $q_{k,i}$ relative to ground truth set $\mathcal{G}_x$. Unlike deterministic tasks (e.g., math or code), open-ended event forecasting is not directly verifiable due to

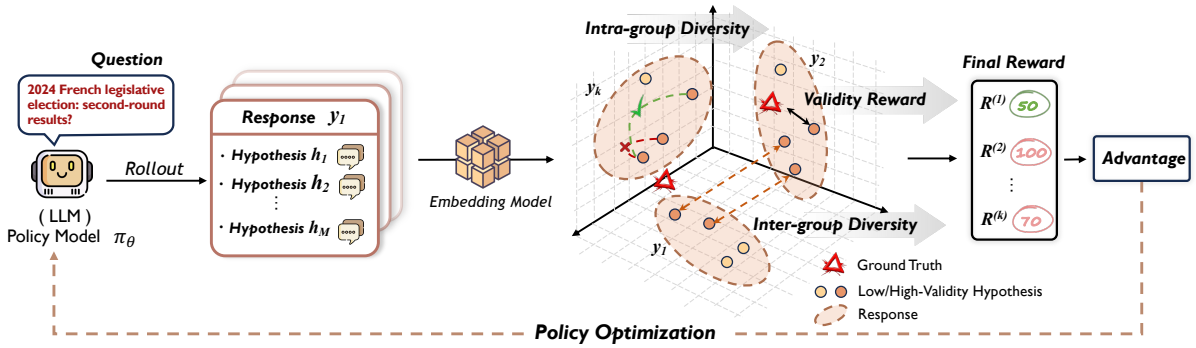

Figure 2: Overall framework of SCATTER. A policy model (LLM) samples multiple hypothesis sets (rollouts) for a given query. These hypotheses are mapped into an embedding space to evaluate *validity* (alignment with ground truth), *intra-group diversity* within each response and *inter-group diversity* across different responses. The resulting rewards are used for policy optimization, guiding the model to produce inclusive and diverse hypothesis sets.

free-form ground truth. Therefore, we use a pre-trained text embedding model to encode each hypothesis $h_{k,i}$ into a dense vector representation. Let $\mathbf{h}_{k,i}$ denote the embedding of i-th hypothesis in the k-th sampled response. We compute $q_{k,i}$ as the maximum cosine similarity between $h_{k,i}$ and the embeddings of ground truth events $\mathbf{g}$.

$$q_{k,m} = \max_{\mathbf{g} \in \mathcal{G}x} \cos(\mathbf{h}_{k,i}, \mathbf{g}). \tag{3}$$

This maximization implies a recall-oriented evaluation strategy: a hypothesis is considered valid if it aligns with at least one plausible future, mirroring the one-to-many nature of open-ended forecasting.

We then aggregate hypothesis-level validity scores to obtain a response-level validity reward by averaging over the M hypotheses:

$$R^{(k)}_{\text{validity}} = \frac{1}{M} \sum_{i=1}^{M} q_{k,i}. \tag{4}$$

While $R^{(k)}_{\text{validity}}$ promotes semantic correctness, optimizing this objective in isolation often causes the policy π_θ to converge to a single mode, resulting in mode collapse.

4.2 Intra-Group Diversity Reward

To incentivize diversity among generated hypotheses while preserving validity, we introduce a intra-group diversity reward. Specifically, we calculate the pairwise cosine dissimilarity between hypotheses, weighted by their respective validity scores:

$$R^{(k)}_{\text{intra}} = \frac{2}{M(M-1)} \sum_{i<j} \Big[\big(1 - \cos(\mathbf{h}_{k,i}, \mathbf{h}_{k,j})\big) \sqrt{q_{k,i}\, q_{k,j}} \Big]. \tag{5}$$

Here, the term $1 - \cos(\cdot, \cdot)$ encourages geometric separation in the embedding space, while $\sqrt{q_{k,i} q_{k,j}}$ serves as a soft constraint. Crucially, this constraint penalizes pairs involving low-validity hypotheses, preventing the model from inflating diversity with semantically irrelevant outliers. As a result, optimization favors a hypothesis set that is both inclusive and diverse.

4.3 Inter-Group Diversity Reward

Beyond local diversity, global exploration is essential to prevent the policy from collapsing into a single mode across different sampling groups. We define an inter-group diversity reward to quantify the unique contribution of each response relative to the global hypothesis pool.

Importance Reweighting. We first construct a probability distribution over hypotheses within response k to prioritize high-validity hypotheses. The normalized importance weight $\omega_{k,i}$ is defined as:

$$\omega_{k,i} = \frac{q_{k,i}}{\sum_{j=1}^{M} q_{k,j} + \epsilon}. \tag{6}$$

This ensures that the divergence metrics focus on the separation of plausible futures rather than distinguishing between noise.

Weighted Asymmetric Chamfer Distance. To measure the directional diversity from response k to response l, we employ a weighted variant of the Asymmetric Chamfer Distance (Borgefors, 1988; Wu et al., 2021). Specifically, we calculate the expected minimum distance from the salient hypotheses in k to the nearest neighbor in response

l:

$$D(k \to l) = \sum_{i=1}^{M} \omega_{k,i} \left(1 - \max_{j} \cos(\mathbf{h}_{k,i}, \mathbf{h}_{l,j})\right). \tag{7}$$

Intuitively, $D(k \to l)$ represents the coverage gap: it is maximized when the high-validity hypotheses in k have no semantic counterparts in l.

Validity-Gated Diversity Reward We aggregate the diversity of response k against all other responses $l \neq k$ via a leave-one-out average:

$$S_{\text{raw}}^{(k)} = \frac{1}{G-1} \sum_{l \neq k} D(k \to l). \tag{8}$$

However, diversity alone is insufficient, as irrelevant hypotheses can inflate diversity without improving coverage. We therefore apply a validity-gated modulation, scaling the raw diversity reward by the response-level validity:

$$R_{\text{inter}}^{(k)} = S_{\text{raw}}^{(k)} \sqrt{\frac{1}{M} \sum_{i=1}^{M} q_{k,i}}. \tag{9}$$

This gating mechanism ensures that the model is rewarded only for discovering novel and high-validity regions of the semantic space.

4.4 Final Reward

We aggregate the three components into a single composite scalar reward:

$$R^{(k)} = R_{\text{validity}}^{(k)} + R_{\text{intra}}^{(k)} + R_{\text{inter}}^{(k)}. \tag{10}$$

This reward directly operationalizes our objective: guiding the model to produce a compact set of valid hypotheses that collectively approximate the true future distribution. Subsequently, we compute per-hypothesis advantage from the reward and optimize via gradient ascent on the advantage-weighted log-likelihood.

4.5 Forecasting

During the inference phase, we employ a stochastic decoding strategy to explore the future outcome space. For a given input x, the policy π_θ generates a response y_k containing a sequence of M hypotheses. To capture the intrinsic uncertainty, we repeat this process to sample K independent responses, thereby constructing an inclusive and diverse hypothesis set.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate our models on two real-world datasets, OpenForecast (Wang et al., 2025) and OpenEP (Guan et al., 2024). Our model is trained on OpenForecast and evaluated under two settings: in-domain testing on the same dataset, and out-of-domain evaluation on OpenEP to verify its cross-domain generalization. Moreover, we construct a hard subset in both datasets consisting of test samples mispredicted by GPT-4o-mini across all sampling rounds. To mitigate information leakage, we chronologically repartition OpenForecast, using 09/2023 and 01/2024 as cut-off dates to construct the training, validation, and test splits. More details about datasets are provided in Appendix A.1.

Evaluation Metrics. We evaluate the inclusiveness and diversity of hypotheses using embedding-based soft matching metrics: *SoftPass@K*, measuring whether any ground-truth hypothesis is hit, short as SP@K, *SoftRecall@K*, measuring the proportion of ground-truth hypotheses recalled, short as SR@K, and *ValidRatio@K*, measuring the ratio of unique and valid hypotheses, short as VR@K, where K denotes the number of sampling rounds. Detailed formulations are provided in Appendix A.2. While embedding-based metrics may introduce marginal noise, they offer a reliable and efficient proxy. Therefore, we also report an LLM-based Pass@K, which leverages GPT-5o-mini[2] to directly verify correctness. Notably, for $K = 1$, we reported the mean and standard deviation averaged over 16 rounds.

Implementation Details. We conduct our experiments on two open-source base models: Qwen2.5-3B-Instruct (Yang et al., 2024) and Llama3.2-3B-Instruct[3]. We adopt Qwen3-Embedding-4B (Zhang et al., 2025) as the embedding model for both training and evaluation. We benchmark our method against three primary baselines: 1) GPT-4o-mini[4], serving as the non-finetuned backbone; 2) Standard SFT (Ding et al., 2023); and 3) Standard GRPO (Shao et al., 2024), which is optimized solely via a validity reward. Notably, both the GRPO baseline and our proposed method utilize a

[2]https://platform.openai.com/docs/models/gpt-5-mini

[3]https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/

[4]https://platform.openai.com/docs/models/gpt-4o-mini

two-stage training strategy: an SFT warm-up (3k instances) to enforce format alignment, followed by RL fine-tuning (10k instances). To ensure parameter efficiency, all fine-tuning is implemented via LoRA (Hu et al., 2022). Regarding generation parameters, we standardize the output to 10 hypotheses per round across 16 sampling rounds. More details regarding training configuration can be found in Appendix A.5.

5.2 Main Results

Overall Performance. Table 1 presents the overall performance of SCATTER against baseline methods (Base, SFT) and the standard RL baseline (GRPO) across two distinct backbones. SCATTER equipped with Qwen2.5-3B-Instruct achieves state-of-the-art results. This performance indicates that our hybrid reward mechanism successfully promotes the inclusiveness and diversity of the generated hypotheses. To rigorously assess robustness against more complex scenarios, we further evaluate performance on a curated hard subset, defined as the specific collection of test samples where GPT-4o-mini to predict the correct outcome. More detailed information is provided in Appendix A.1. As detailed in Table 2, our method maintains a significant performance advantage on these challenging cases, demonstrating enhanced effectiveness even when facing tasks of increased difficulty. On the standard dataset, the base model achieves performance comparable to GPT-4o-mini but exhibits a significantly higher validity ratio. Surprisingly, on the hard subset, the base model even outperforms GPT-4o-mini. We attribute this anomaly to the alignment tax (Ouyang et al., 2022b; Lin et al., 2024) inherent in safety-aligned commercial models; while GPT-4o-mini has stronger reasoning capabilities, extensive pre-training and post-training alignment bias the model towards high-probability events. It is noteworthy that the standard GRPO baseline occasionally achieves higher *Pass@1* scores. However, this comes at a severe cost: GRPO suffers from mode collapse, with *ValidRatio@16* dropping to near-zero levels. This observation suggests that without explicit validity constraints, standard GRPO exploits the reward signal by aggressively optimizing for inclusiveness, producing "high-scoring" but semantically repetitive and linguistically broken outputs.

Domain Generalization. Notably, SCATTER maintains robust performance even under out-of-domain settings, suggesting superior generalization capabilities under distribution shifts. We attribute this to our explicit optimization objective, which effectively balances hypothesis diversity with semantic validity. Interestingly, we observe that post-training methods on Llama3.2-3B-Instruct exhibit a performance regression compared to the backbone on standard subsets. We posit that Llama-3.2 exhibits a heightened sensitivity to the exploration noise inherent in reinforcement learning, which subsequently undermines its instruction-following stability. This is empirically supported by the significantly lower *ValidRatio@16* observed in Llama-based experiments, suggesting that the propensity for generating invalid responses increases when incentivized to explore a diverse output space.

5.3 Ablation Study

Reward Design. To investigate the individual contributions of our proposed components, we compare SCATTER against two variants: 1) *-inter*, which removes the inter-group diversity reward, and 2) *-intra*, which excludes the intra-group diversity constraint. Table 3 summarizes the results. We observe that the full SCATTER framework achieves the most robust balance between performance and generalization. First, regarding the intra-group diversity component, its removal leads to a catastrophic drop in response validity, particularly in the out-of-domain setting. This indicates that enforcing local diversity is essential for maintaining the semantic quality of hypotheses. Second, while the *-inter* variant occasionally exhibits higher pass scores on the in-domain dataset, this performance is deceptive. It comes at the cost of significantly lower validity and, crucially, reduced generalization on the out-of-domain benchmark. This suggests that the inter-group diversity acts as a global regularizer, preventing the model from overfitting to in-domain patterns with high-entropy but low-validity outputs, thereby ensuring robust adaptation to unseen domains.

Validity Gated Score. We conduct an ablation study to isolate the contribution of the validity aggregation strategy. Specifically, we assess whether the validity gate mechanism improves the ability to balance inclusiveness and diversity. We compare the full model against three variants: **Vanilla**, which excludes the validity-gated score; **Mean**, which utilizes the mean of the scores; and **Min**, which adopts the minimum score. As shown in

Model	OpenForecast				OpenEP			
	Pass@1/Pass@16	SP@1/SP@16	SR@1/SR@16	VR@16	Pass@1/Pass@16	SP@1/SP@16	SR@1/SR@16	VR@16
GPT-4o-mini	$7.55^{\pm 0.73}$/25.44	$9.39^{\pm 0.80}$/22.81	$4.96^{\pm 0.37}$/12.76	4.23	$18.82^{\pm 1.44}$/32.22	$6.87^{\pm 1.06}$/17.78	$3.06^{\pm 0.81}$/9.28	4.08
Llama3.2-3B-Instruct								
Base	$4.40^{\pm 0.72}$/21.71	$9.40^{\pm 0.94}$/27.19	$4.85^{\pm 0.53}$/15.42	5.73	**18.26**$^{\pm 1.76}$/**33.33**	$7.29^{\pm 1.57}$/18.89	$2.97^{\pm 0.76}$/9.72	5.67
+ SFT	$4.50^{\pm 1.05}$/20.39	$7.10^{\pm 0.66}$/21.27	$3.64^{\pm 0.45}$/11.72	**7.74**	$6.81^{\pm 2.00}$/25.56	$9.10^{\pm 1.85}$/30.00	$2.37^{\pm 0.54}$/11.43	**6.92**
+ GRPO	**11.90**$^{\pm 1.05}$/28.29	$13.72^{\pm 0.74}$/21.05	$7.08^{\pm 0.42}$/11.61	0.71	$2.85^{\pm 1.30}$/5.56	$11.11^{\pm 1.52}$/20.00	$3.14^{\pm 0.70}$/6.35	0.62
+ SCATTER	$7.73^{\pm 0.80}$/**31.36**	**17.45**$^{\pm 0.81}$/**42.32**	**9.47**$^{\pm 0.55}$/**25.64**	6.54	$7.36^{\pm 2.29}$/25.56	**15.69**$^{\pm 2.63}$/**36.67**	**4.67**$^{\pm 1.03}$/**15.30**	4.90
Qwen2.5-3B-Instruct								
Base	$5.96^{\pm 1.01}$/25.88	$6.95^{\pm 0.58}$/20.61	$3.67^{\pm 0.26}$/10.72	5.07	**15.00**$^{\pm 2.22}$/32.22	$4.93^{\pm 1.71}$/17.78	$2.51^{\pm 0.95}$/10.20	5.56
+ SFT	$5.51^{\pm 0.46}$/23.68	$8.55^{\pm 0.91}$/24.56	$4.48^{\pm 0.45}$/13.85	7.10	$5.00^{\pm 2.19}$/20.00	$6.46^{\pm 1.72}$/21.11	$2.02^{\pm 0.71}$/8.38	5.52
+ GRPO	**11.84**$^{\pm 0.98}$/25.66	$17.60^{\pm 0.66}$/27.85	$9.70^{\pm 0.46}$/15.94	0.87	$5.49^{\pm 1.73}$/13.33	$19.65^{\pm 1.51}$/26.67	**6.92**$^{\pm 0.53}$/10.58	0.76
+ SCATTER	$8.55^{\pm 0.87}$/**30.26**	**17.61**$^{\pm 1.12}$/**41.89**	**9.77**$^{\pm 0.59}$/**24.29**	**9.14**	$9.65^{\pm 2.71}$/**35.56**	**16.25**$^{\pm 2.75}$/**38.89**	$4.88^{\pm 0.93}$/**17.59**	**8.21**

Table 1: Main results. Best scores for each backbone are **bolded** and underlined for the second best. The green superscript indicates the standard deviation.

Model	OpenForecast-Hard (N=335)				OpenEP-Hard (N=61)			
	Pass@1/Pass@16	SP@1/SP@16	SR@1/SR@16	VR@16	Pass@1/Pass@16	SP@1/SP@16	SR@1/SR@16	VR@16
GPT-4o-mini	$0.00^{\pm 0.00}$/0.00	$4.65^{\pm 0.59}$/14.93	$1.76^{\pm 0.35}$/6.75	4.10	$0.00^{\pm 0.00}$/0.00	$2.56^{\pm 1.29}$/4.92	$0.92^{\pm 0.50}$/3.28	4.03
Llama3.2-3B-Instruct								
Base	$0.95^{\pm 0.55}$/10.15	$6.16^{\pm 0.97}$/20.00	$2.55^{\pm 0.43}$/9.43	5.56	$1.23^{\pm 1.08}$/9.84	$3.69^{\pm 1.36}$/9.84	$1.27^{\pm 0.47}$/3.83	5.75
+ SFT	$2.05^{\pm 0.71}$/13.13	$4.79^{\pm 0.90}$/16.72	$1.86^{\pm 0.45}$/7.70	**7.71**	$4.41^{\pm 2.37}$/16.39	$7.07^{\pm 2.51}$/24.59	$1.86^{\pm 0.79}$/8.69	**6.68**
+ GRPO	**6.55**$^{\pm 0.98}$/21.19	$9.27^{\pm 0.54}$/15.22	$3.67^{\pm 0.35}$/6.99	0.75	$0.51^{\pm 0.76}$/1.64	$11.27^{\pm 1.52}$/19.67	$3.05^{\pm 0.39}$/5.23	0.62
+ SCATTER	$3.75^{\pm 0.75}$/**23.28**	**13.15**$^{\pm 0.96}$/**34.33**	**6.14**$^{\pm 0.65}$/**19.09**	6.58	**5.94**$^{\pm 2.24}$/**16.39**	**15.27**$^{\pm 2.37}$/**27.87**	**4.65**$^{\pm 0.76}$/**11.44**	4.71
Qwen2.5-3B-Instruct								
Base	$1.47^{\pm 0.43}$/10.75	$4.68^{\pm 0.54}$/14.63	$1.71^{\pm 0.31}$/5.74	4.93	$2.56^{\pm 1.42}$/6.56	$1.84^{\pm 1.82}$/9.84	$0.85^{\pm 0.80}$/5.33	5.13
+ SFT	$2.16^{\pm 0.43}$/13.43	$5.45^{\pm 0.88}$/17.61	$2.35^{\pm 0.53}$/8.95	6.99	$1.23^{\pm 1.08}$/4.92	$4.41^{\pm 1.50}$/14.75	$1.37^{\pm 0.51}$/4.92	5.44
+ GRPO	**6.36**$^{\pm 0.77}$/17.01	$12.95^{\pm 0.85}$/21.19	$5.68^{\pm 0.50}$/10.56	0.87	$2.15^{\pm 1.61}$/6.56	$16.60^{\pm 1.73}$/21.31	$4.99^{\pm 0.49}$/7.91	0.74
+ SCATTER	$4.12^{\pm 0.79}$/**22.39**	**13.40**$^{\pm 1.07}$/**34.93**	**6.54**$^{\pm 0.62}$/**17.92**	**8.99**	**6.66**$^{\pm 2.05}$/**21.31**	$13.42^{\pm 3.90}$/**31.15**	$4.39^{\pm 1.18}$/**15.44**	**8.15**

Table 2: Performance on the subsets of hard samples. Best scores for each backbone are **bolded**, and second best are underlined. The green superscript indicates the standard deviation.

Model	OpenForecast				OpenEP			
	Pass	SP	SR	VR	Pass	SP	SR	VR
Llama3.2-3B-Instruct								
SCATTER	31.36	42.32	25.64	**6.54**	**25.56**	**36.67**	**15.30**	**4.90**
-inter	**34.43**	**44.74**	**25.91**	4.36	22.22	33.33	14.42	3.97
-intra	31.14	42.11	25.36	2.64	23.33	30.00	12.45	1.93
GRPO	28.29	21.05	11.61	0.71	5.56	20.00	6.35	0.62
Qwen2.5-3B-Instruct								
SCATTER	30.26	**41.89**	24.29	**9.14**	**35.56**	38.89	17.59	**8.21**
-inter	**33.11**	41.01	**24.44**	4.74	33.33	**40.00**	**18.84**	4.78
-intra	33.11	41.45	24.14	4.19	32.22	40.0	18.7	4.14
GRPO	25.66	27.85	15.94	0.87	13.33	26.67	10.58	0.76

Table 3: Performance comparison *w.r.t.* various reward designs. Pass, SP, SR, and VR are based on $K = 16$.

Table 4, the complete SCATTER method demonstrates the most robust overall performance. On OpenEP, it achieves state-of-the-art results across all accuracy metrics. Similarly, on OpenForecast, SCATTER leads in both *SoftPass@16* and *SoftRecall@16*. In contrast, while Vanilla attains the highest validity ratio, its significantly lower performance on accuracy-oriented metrics indicates that optimizing for validity alone, without effective aggregation, compromises the correctness of the generated hypotheses. Although SCATTER-Min shows competitive performance on OpenForecast, the full SCATTER framework provides a superior

Model	OpenForecast				OpenEP			
	Pass	SP	SR	VR	Pass	SP	SR	VR
Vanilla	19.3	24.56	13.84	**17.77**	11.11	15.56	6.9	**15.7**
Mean	11.09	41.24	24.08	9.54	26.67	27.78	11.59	8.49
Min	**31.80**	41.67	23.90	8.21	18.89	23.33	9.41	7.17
SCATTER	30.26	**41.89**	**24.29**	9.14	**35.56**	**38.89**	**17.59**	8.21

Table 4: Performance comparison *w.r.t.* various validity gated score. Pass, SP, SR, and VR are based on $K = 16$.

balance between accuracy and diversity.

5.4 Model Study

Impact of Number of Sampling Round K. Figure 3 illustrates the performance scaling of Qwen2.5 across varying sampling rounds ($K \in [1, 16]$). Additional results regarding Llama3.2 can be found in Appendix A.7. In terms of inclusiveness, SCATTER demonstrates superior scaling properties. Unlike GRPO, which saturates at high K, SCATTER scales effectively and significantly outperforms backbone and SFT models on *SoftPass@k*. Crucially, this gain in diversity does not come at the expense of quality; while the *ValidRatio@k* for GRPO suffers a precipitous decline toward near-zero values, SCATTER consistently retains the highest validity density. This confirms that our method effectively leverages expanded

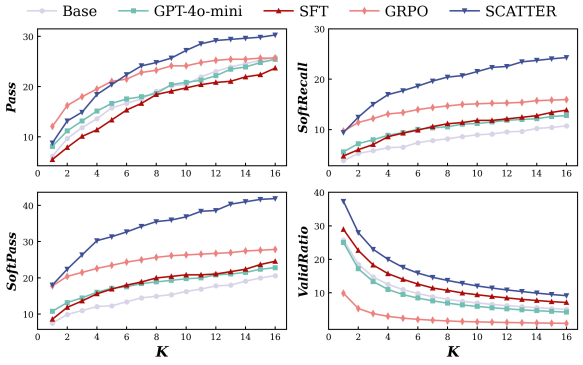

Figure 3: Performance scaling *w.r.t.* sampling rounds K. Comparison of SCATTER against baselines on OpenForecast using Qwen2.5-3B-Instruct as Base model.

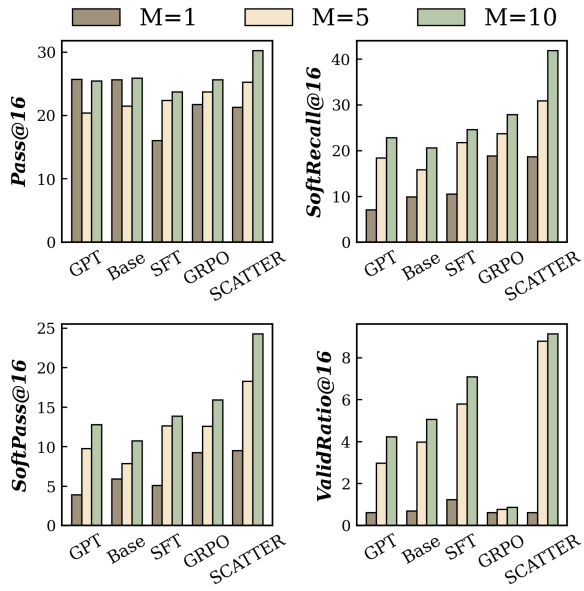

Figure 4: Performance scaling *w.r.t.* the number of hypotheses M per round.

compute budgets to uncover a broader spectrum of semantically valid hypotheses without drifting into plausible-but-incorrect generation modes.

Impact of Number of Hypothesis M. Figure 4 illustrates the impact of varying $M \in \{1, 5, 10\}$ on model performance. SCATTER exhibits the most robust scaling properties among all compared methods. SCATTER maintains a steep upward trajectory. Notably, at $M = 10$, SCATTER achieves significant margins over the strongest baseline. This confirms that our method can effectively leverage larger M values to enhance semantic coverage and accuracy. However, baselines such as GPT and the base model exhibit limited improvement or saturation as M increases. Notably, a performance dip occurs when transitioning from deterministic ($M = 1$) to stochastic decoding. This decline is attributed to the introduction of noise and local optima, which small sample sizes fail to mitigate until a sufficient number of hypotheses are generated.

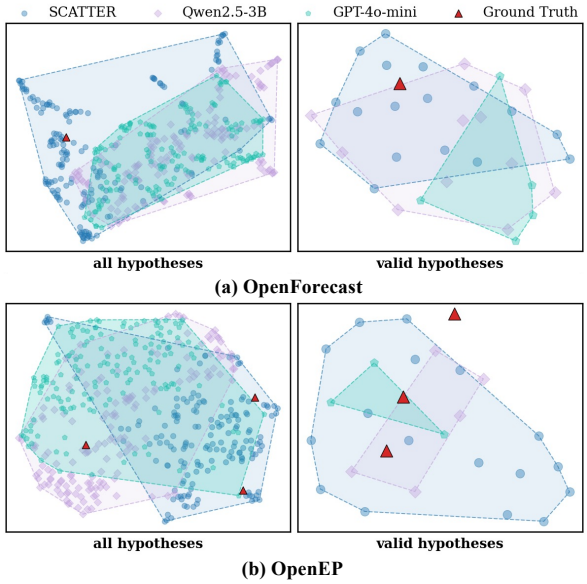

Figure 5: UMAP visualization of the semantic manifold of the generated hypotheses on two datasets.

Case Study. To provide a qualitative perspective on the mechanism behind SCATTER's superior performance, Figure 5 visualizes the semantic manifold of generated hypotheses via UMAP. SCATTER demonstrates a robust balance between semantic exploration and ground truth alignment. In the distribution of all hypotheses, SCATTER shows significantly broader coverage than baselines. More importantly, the distribution of valid hypotheses indicates that this exploration is constructive. While SCATTER shows broader coverage overall, the significantly more uniform distribution of its valid hypotheses is particularly notable. This indicates constructive exploration, demonstrating that our method effectively confines exploration to contextually plausible futures. Specific examples of generated hypotheses are provided in the Appendix A.8.

6 Conclusion

In this paper, we address the limitations of the prevailing single-prediction paradigm in open-ended event forecasting and advocate for a shift toward *scatter forecasting*. By reformulating the task as hypothesis generation, we aim to better capture the intrinsic uncertainty of real-world developments. To realize this, we propose SCATTER, a reinforcement learning framework designed to simultaneously optimize the inclusiveness and diversity of generated hypotheses. Extensive experiments on

the OpenForecast and OpenEP benchmarks demonstrate that our approach significantly outperforms existing baselines, producing hypotheses that are both diverse and high-quality.

Limitations

In this work, we identify several limitations to be addressed in future research. First, due to computational constraints and API costs, our experiments are primarily conducted on 3B-parameter models using LoRA fine-tuning, and our evaluation of closed-source commercial LLMs is restricted to GPT-4o-mini. Second, we set the maximum number of sampling rounds (K) to 16. Given the inherent uncertainty of LLMs, minor statistical fluctuations may occur across different runs; however, these marginal deviations do not compromise the validity of our core experimental conclusions. Finally, although the datasets we adopt contain multiple ground-truth references, they still capture a limited subset of the complexity found in real-world events. To address this gap, we will extend our framework to label-free settings in the future work. Additionally, AI assistants were used solely for language editing and polishing, without affecting the research content, methodology, or conclusions.

7 Acknowldegements

We thank the anonymous reviewers for their valuable comments. This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

References

Kenneth J Arrow, Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D Nelson, and 1 others. 2008. The promise of prediction markets.

Gunilla Borgefors. 1988. Hierarchical chamfer matching: A parametric edge matching algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.*, 10(6):849–865.

He Chang, Jie Wu, Zhulin Tao, Yunshan Ma, Xianglin Huang, and Tat-Seng Chua. 2025. Integrate temporal graph learning into llm-based temporal knowledge graph model. *CoRR*, abs/2501.11911.

He Chang, Chenchen Ye, Zhulin Tao, Jie Wu, Zhengmao Yang, Yunshan Ma, Xianglin Huang, and Tat-Seng Chua. 2024. A comprehensive evaluation of large language models on temporal event forecasting. *CoRR*, abs/2407.11638.

Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. 2025. SEED-GRPO: semantic entropy enhanced GRPO for uncertainty-aware policy optimization. *CoRR*, abs/2505.12346.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. 2025. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *ICML*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Songgaojun Deng, Maarten de Rijke, and Yue Ning. 2024. Advances in human event modeling: From graph neural networks to language models. In *KDD*, pages 6459–6469. ACM.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat. Mac. Intell.*, 5(3):220–235.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. 2024. Approaching human-level forecasting with language models. In *NeurIPS*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net.

Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. 2021. Forecastqa: A question answering challenge for event forecasting with temporal text data. In *ACL/IJCNLP (1)*, pages 4636–4650. Association for Computational Linguistics.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *ICML*, Proceedings of Machine Learning Research, pages 15696–15707. PMLR.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip Tetlock. 2025. Forecastbench: A dynamic benchmark of AI forecasting capabilities. In *ICLR*. OpenReview.net.

Risto Conte Keivabu. 2019. Philip tetlock and dan gardner: Superforecasting: The art and science of prediction. *Sociologickỳ časopis/Czech Sociological Review*, 55(3):406–409.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of RLHF on LLM generalisation and diversity. In *ICLR*. OpenReview.net.

Sang-Woo Lee, Sohee Yang, Donghyun Kwak, and Noah Y. Siegel. 2025. Advancing event forecasting through massive training of large language models: Challenges, solutions, and broader impacts. *CoRR*, abs/2507.19477.

Haoxuan Li, He Chang, Yunshan Ma, Yi Bin, Yang Yang, See-Kiong Ng, and Tat-Seng Chua. 2026. Thinktank-me: A multi-expert framework for middle east event forecasting. In *WWW*, pages 8541–8544. ACM.

Haoxuan Li, Zhengmao Yang, Yunshan Ma, Yi Bin, Yang Yang, and Tat-Seng Chua. 2024. Mm-forecast: A multimodal approach to temporal event forecasting with large language models. In *ACM Multimedia*, pages 2776–2785. ACM.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. 2024. Mitigating the alignment tax of RLHF. In *EMNLP*, pages 580–606. Association for Computational Linguistics.

Yunshan Ma, Chenchen Ye, Zijian Wu, Xiang Wang, Yixin Cao, and Tat-Seng Chua. 2023a. Context-aware event forecasting via graph disentanglement. In *KDD*, pages 1643–1652. ACM.

Yunshan Ma, Chenchen Ye, Zijian Wu, Xiang Wang, Yixin Cao, Liang Pang, and Tat-Seng Chua. 2023b. Structured, complex and time-complete temporal event forecasting. *CoRR*, abs/2312.01052.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022a. Training language models to follow instructions with human feedback. In *NeurIPS*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022b. Training language models to follow instructions with human feedback. In *NeurIPS*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP/IJCNLP (1)*, pages 3980–3990. Association for Computational Linguistics.

Philipp Schoenegger, Peter S. Park, Ezra Karger, Sean Trott, and Philip E. Tetlock. 2025. Ai-augmented predictions: LLM assistants improve human forecasting accuracy. *ACM Trans. Interact. Intell. Syst.*, 15(1):4:1–4:25.

Philipp Schoenegger, Indre Tuminauskaite, Peter S Park, Rafael Valdece Sousa Bastos, and Philip E Tetlock. 2024. Wisdom of the silicon crowd: Llm ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, 10(45):eadp1528.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross J. Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nat.*, 631(8022):755–759.

Jing Su, Chufeng Jiang, Xin Jin, Yuxin Qiao, Tingsong Xiao, Hongda Ma, Rong Wei, Zhi Jing, Jiajun Xu, and Junhong Lin. 2024. Large language models for forecasting and anomaly detection: A systematic literature review. *CoRR*, abs/2402.10350.

Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement learning - an introduction*. Adaptive computation and machine learning. MIT Press.

Gemma Team. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

Zhen Wang, Xi Zhou, Yating Yang, Bo Ma, Lei Wang, Rui Dong, and Azmat Anwar. 2025. Openforecast: A large-scale open-ended event forecasting dataset. In *COLING*, pages 5273–5294. Association for Computational Linguistics.

Jack Wildman, Nikos I. Bosse, Daniel Hnyk, Peter Mühlbacher, Finn Hambly, Jon Evans, Dan Schwarz, and Lawrence Phillips. 2025. Bench to the future: A pastcasting benchmark for forecasting agents. *CoRR*, abs/2506.21558.

Tong Wu, Liang Pan, Junzhe Zhang, Tai Wang, Ziwei Liu, and Dahua Lin. 2021. Density-aware chamfer distance as a comprehensive metric for point cloud completion. *CoRR*, abs/2111.12702.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Qingchuan Yang, Simon Mahns, Sida Li, Anri Gu, Jibang Wu, and Haifeng Xu. 2025. Llm-as-a-prophet: Understanding predictive intelligence with prophet arena. *CoRR*, abs/2510.17638.

Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, and Kay Chen Tan. 2025. Diversity-aware policy optimization for large language model reasoning. *CoRR*, abs/2505.23433.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, and 16 others. 2025a. DAPO: an open-source LLM reinforcement learning system at scale. *CoRR*, abs/2503.14476.

Zi Yu, Shaoxiang Wang, Guozheng Li, Yu Zhang, and Chi Harold Liu. 2025b. Forestcast: Open-ended event forecasting with semantic news forest. In *EMNLP (Findings)*, pages 12667–12681. Association for Computational Linguistics.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *CoRR*, abs/2504.13837.

Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, Mingren Yin, Zhenwei Zhu, Tianle Cai, Zehui Chen, Jiecao Chen, Yantao Du, Xiang Gao, Jiacheng Guo, Liang Hu, Jianpeng Jiao, and 11 others. 2025. Futurex: An advanced live benchmark for LLM agents in future prediction. *CoRR*, abs/2508.11987.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *CoRR*, abs/2506.05176.

Zhihan Zhang, Yixin Cao, Chenchen Ye, Yunshan Ma, Lizi Liao, and Tat-Seng Chua. 2024. Analyzing temporal complex events with large language models? A benchmark towards temporal, long context understanding. In *ACL (1)*, pages 1588–1606. Association for Computational Linguistics.

Yujun Zhou, Zhenwen Liang, Haolin Liu, Wenhao Yu, Kishan Panaganti, Linfeng Song, Dian Yu, Xiangliang Zhang, Haitao Mi, and Dong Yu. 2025. Evolving language models without labels: Majority drives selection, novelty promotes variation. *CoRR*, abs/2509.15194.

Dataset	Split	#Questions	Time Span
OpenForecast	Train	32596	1950-02 ∼ 2023-08
	Valid	654	2023-09 ∼ 2023-12
	Test	456	2024-01 ∼ 2024-12
	Hard-Test	335	2024-01 ∼ 2024-12
OpenEP	Test	90	2024-07 ∼
	Hard-Test	61	2024-07 ∼

Table 5: Dataset statistics for OpenForecast and OpenEP. #Questions denotes the number of samples.

Model	Knowledge Cutoff
GPT-4o mini	Oct. 2023
Llama-3.2-3B-Instruct	Dec. 2023
Qwen2.5-3B-Instruct	Dec. 2023

Table 6: Knowledge cutoff dates of the employed LLMs.

Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. 2022. Forecasting future world events with neural networks. In *NeurIPS*.

A Appendix

A.1 Datasets Details

We evaluate our method using two real-world datasets: OpenEP (Guan et al., 2024) and OpenForecast (Wang et al., 2025). For OpenEP, since each answer corresponds to a span in a real news article, we re-extract answer spans to ensure evaluation reliability. Refer to Appendix A.9 for the specific prompts utilized. Notably, we filter out time and location prediction queries due to their high uncertainty and task misalignment. For OpenForecast, we focus on the short-term forecasting subset, which features precise timestamps and allows for multiple valid answers. o mitigate information leakage, we re-partitioned the OpenForecast dataset based on the knowledge cutoff date of the employed LLM. The knowledge cutoff dates are listed in Table 6, and detailed statistics for both datasets are provided in Table 5.

A.2 Evaluation Details

To rigorously assess the inclusiveness and diversity of the generated hypotheses, we employ a hybrid evaluation framework combining embedding-based metrics with LLM-based verification. We define the dataset-level metric as the average of sample-level scores over N test samples:

$$\text{Metric@}K = \frac{1}{N}\sum_{n=1}^{N}\text{Score}_n. \tag{11}$$

Model	OpenForecast			OpenEP		
	SP@1/SP@16	SR@1/SR@16	VR@16	SP@1/SP@16	SR@1/SR@16	VR@16
Base	$1.43^{\pm 0.36}$/6.14	$0.82^{\pm 0.28}$/3.81	24.45	$0.49^{\pm 0.68}$/3.33	$0.15^{\pm 0.24}$/1.00	25.69
+ SFT	$2.12^{\pm 0.39}$/8.77	$1.14^{\pm 0.24}$/4.42	22.03	$1.04^{\pm 0.92}$/6.67	$0.33^{\pm 0.40}$/2.75	19.38
+ GRPO	$6.35^{\pm 0.38}$/11.62	$3.33^{\pm 0.21}$/5.94	2.39	$6.88^{\pm 1.53}$/15.56	$1.86^{\pm 0.35}$/4.41	2.26
+ SCATTER	$6.21^{\pm 0.63}$/18.64	$3.38^{\pm 0.40}$/9.85	28.41	$4.38^{\pm 1.64}$/17.78	$1.02^{\pm 0.39}$/5.29	26.01

Table 7: Cross-embedding evaluation results of Qwen2.5-3B-Instruct using all-MiniLM-L6-v2. The green superscript indicates the standard deviation.

Model	OpenForecast			OpenEP		
	SP@1/SP@16	SR@1/SR@16	VR@16	SP@1/SP@16	SR@1/SR@16	VR@16
Base	$23.68^{\pm 0.93}$/46.71	$11.63^{\pm 0.69}$/26.40	1.04	$35.14^{\pm 3.64}$/67.78	$14.60^{\pm 1.66}$/37.72	1.28
+ SFT	$23.73^{\pm 0.69}$/52.19	$11.39^{\pm 0.48}$/30.00	2.08	$30.49^{\pm 3.45}$/70.00	$13.49^{\pm 1.61}$/39.98	1.56
+ GRPO	$23.99^{\pm 1.02}$/41.01	$12.24^{\pm 0.66}$/24.08	0.73	$26.74^{\pm 2.47}$/46.67	$11.16^{\pm 1.07}$/21.37	0.71
+ SCATTER	$32.11^{\pm 1.25}$/64.04	$16.63^{\pm 0.76}$/41.87	3.18	$40.35^{\pm 3.53}$/73.33	$19.97^{\pm 2.08}$/42.10	3.28

Table 8: Cross-embedding evaluation results of Qwen2.5-3B-Instruct using gemma-300m. The green superscript indicates the standard deviation.

The specific definitions for each metric are detailed below.

SoftPass@K. This metric measures the generation success rate, indicating whether the model produces *at least one* hypothesis that meets the semantic requirement. Formally, we define the sample-level score as:

$$\text{SoftPass}_n = \mathbb{I}\left(\sum_{k=1}^{K}\sum_{i=1}^{M}\mathbb{I}\left(S_{k,i} > \tau_{sp}\right) \geq 1\right), \tag{12}$$

where $\tau_{sp} = 0.8$ is the similarity threshold, $\mathbb{I}(\cdot)$ is the indicator function. Here, $S_{k,i} = \max_{g \in \mathcal{G}x} cos(\mathbf{h}_{k,i}, \mathbf{g})$ is the maximum similarity between the i-th hypothesis in the k-th round and any ground truth $g \in \mathcal{G}_x$

SoftRecall@K. SoftRecall assesses the semantic coverage of the ground truth space. It quantifies the proportion of ground truth embeddings that are successfully retrieved by the generated hypothesis set. The score is defined as:

$$\text{SoftRecall}_n = \frac{1}{|\mathcal{G}x|}\sum_{g \in \mathcal{G}x}\mathbb{I}\left(S_{h,g} > \tau_{sr}\right), \tag{13}$$

where $\tau_{sr} = 0.8$ serves as the semantic similarity threshold for a successful retrieval and $S_{h,g}$ denote the cosine similarity between the generated hypothesis and the ground truth.

ValidRatio@K. To quantify diversity, this metric measures the proportion of valid, non-redundant hypotheses. The score ValidRatio_n is defined as:

$$\text{ValidRatio}_n = \frac{1}{K \cdot M}\sum_{k=1}^{K}\sum_{i=1}^{M} V_{k,i}, \tag{14}$$

where $V_{k,i} \in \{0, 1\}$ serves as the validity indicator determined by our diversity filtering mechanism. The indicator $V_{k,i}$ is computed dynamically to enforce distinctiveness: for the initial round ($i = 1$), we apply greedy intra-round filtering; for subsequent rounds ($i > 1$), candidates are filtered against the history of all valid hypotheses from previous rounds. Additionally, we enforce a minimum semantic relevance by requiring $S_{k,i} \geq \tau_{valid}$, where $\tau_{valid} = 0.4$.

Pass@K. Complementing the embedding-based metrics with strict verification, we utilize GPT-5o-mini as an external oracle. Pass@K is defined as the proportion of problems where at least one hypothesis is verified as logically correct by the LLM. The detailed evaluation prompt is provided in Appendix A.9.

A.3 Cross-embedding Evaluation

To evaluate performance across different embedding models, we use two alternative embedding models, all-MiniLM-L6-v2 (Reimers and Gurevych, 2019) and gemma-embedding (Team, 2024), which are widely adopted and architecturally distinct from the embedding model used for reward computation. Specifically, all-MiniLM-L6-v2 is a lightweight Sentence-BERT model optimized for general semantic similarity, while

Method	Qwen3-Embedding-4B		all-MiniLM-L6-v2	
	SP@1/SP@16	SR@1/SR@16	SP@1/SP@16	SR@1/SR@16
Threshold = 0.70				
GRPO	$44.78^{\pm 0.95}/59.43$	$27.75^{\pm 0.59}/38.60$	$16.13^{\pm 0.87}/28.51$	$8.37^{\pm 0.60}/15.50$
SCATTER	$47.81^{\pm 1.15}/78.51$	$29.84^{\pm 0.80}/56.79$	$18.00^{\pm 0.95}/46.05$	$9.74^{\pm 0.56}/26.61$
Threshold = 0.80				
GRPO	$11.84^{\pm 0.98}/25.66$	$17.60^{\pm 0.66}/27.85$	$6.35^{\pm 0.38}/11.62$	$3.33^{\pm 0.21}/5.94$
SCATTER	$8.55^{\pm 0.87}/30.26$	$17.61^{\pm 1.12}/41.89$	$6.21^{\pm 0.63}/18.64$	$3.38^{\pm 0.40}/9.85$
Threshold = 0.90				
GRPO	$3.29^{\pm 0.45}/6.36$	$1.57^{\pm 0.18}/2.79$	$1.40^{\pm 0.36}/2.19$	$0.89^{\pm 0.24}/1.35$
SCATTER	$2.73^{\pm 0.38}/9.87$	$1.32^{\pm 0.19}/4.78$	$0.66^{\pm 0.33}/4.82$	$0.34^{\pm 0.18}/2.57$

Table 9: Impact of similarity thresholds on SP@16 and SR@16 for OpenForecast. The green superscript indicates the standard deviation.

Method	Qwen3-Embedding-4B	all-MiniLM-L6-v2
Threshold = 0.40		
GRPO	0.66	1.05
SCATTER	2.20	9.85
Threshold = 0.50		
GRPO	0.72	1.50
SCATTER	5.15	17.20
Threshold = 0.60		
GRPO	0.87	2.39
SCATTER	9.14	28.41

Table 10: Impact of similarity thresholds on VR@16 for OpenForecast.

Hyperparameter	Value
Training Batch Size	480
PPO Mini-Batch Size	480
PPO Micro-Batch Size (per GPU)	32
Total Epochs	6
Group Rollout Size (G)	5
Learning Rate	3×10^{-5}
KL Loss	True
KL Coefficient (β)	0.001
LoRA	True
LoRA Rank (r)	64
LoRA Alpha (α)	32
Max Prompt Length	1024
Max Response Length	512
Generation Temperature	1.0

Table 11: Training configurations for GRPO and SCATTER.

gemma-embedding is built upon the Gemma foundation model family, providing a different embedding architecture and training paradigm. As reported in Table 7 and Table 8, SCATTER consistently outperforms GRPO and other baselines across both alternative embedding spaces. Notably, the relative performance margins remain stable despite the change of evaluator, demonstrating that the effectiveness of SCATTER does not depend on a specific embedding model. These results indicate that the observed improvements are robust and not artifacts of evaluator–trainer embedding alignment.

A.4 Impact of similarity thresholds

We further perform a comprehensive sensitivity analysis over multiple similarity thresholds. Specifically, we vary the matching threshold $\tau_{sr}, \tau_{sp} \in \{0.70, 0.80, 0.90\}$ for SP@16 and SR@16, and $\tau_{valid} \in \{0.40, 0.50, 0.60\}$ for VR@16. The detailed results are reported in Table 9 and Table 10. Across all threshold configurations, SCATTER consistently outperforms the GRPO baseline. These results demonstrate that the observed improvements are stable and not dependent on a specific similarity cutoff.

A.5 Training Configuration

We benchmark our method against four baselines: gpt-4o-mini, the non-finetuned base model, standard SFT, and standard GRPO. To ensure a fair evaluation, we enhance the prompt of base model and gpt-4o-mini with specific hypothesis constraints (see Appendix A.9). For the SFT baseline, we curate a dataset of 3k instances via knowledge distillation from GPT-4o-mini; to guarantee high-quality supervision, we employ a rewrite-and-replace strategy where the ground truth is stylistically adapted to replace the most semantically similar generated candidate. Notably, both the GRPO baseline and our method adopt a two-stage training paradigm:

Method	OpenForecast			OpenEP		
	SP@1/SP@16	SR@1/SR@16	VR@16	SP@1/SP@16	SR@1/SR@16	VR@16
Base	$9.33^{\pm 1.14}/23.90$	$4.69^{\pm 0.66}/13.04$	4.44	$7.01^{\pm 1.74}/18.89$	$3.34^{\pm 1.11}/9.62$	4.09
+ SFT	$10.35^{\pm 0.86}/28.95$	$5.34^{\pm 0.56}/16.31$	8.57	$4.44^{\pm 2.19}/23.33$	$1.36^{\pm 0.69}/9.83$	6.82
+ GRPO	$18.16^{\pm 0.55}/24.56$	$9.41^{\pm 0.32}/13.37$	0.64	$14.10^{\pm 0.94}/17.78$	$4.96^{\pm 0.53}/6.42$	0.62
+ SCATTER	$17.65^{\pm 1.32}/43.64$	$9.46^{\pm 0.70}/25.48$	10.95	$7.36^{\pm 2.18}/27.78$	$2.36^{\pm 0.94}/10.99$	8.88

Table 12: Performance comparison on the Qwen3-7B-Instruct backbone. The green superscript indicates the standard deviation.

an initial SFT warm-up on these 3k instances to enforce output formatting, followed by the respective RL fine-tuning stages on 10K randomly sampled instances. The detailed configuration is shown in Table 11. During inference, we employ a rollout temperature of 0.7, nucleus sampling with $p = 0.8$, top-k sampling with $k = 20$, and a maximum response length of 512 tokens.

A.6 Impact of Model Size

We further conduct experiments on Qwen2.5-7B-Instruct. As shown in Table 12, despite the stronger base capability of the larger backbone, SCATTER continues to provide consistent improvements over the corresponding baselines, indicating that the proposed reward design remains effective as model capacity increases. We acknowledge that evaluation on 70B-scale models would be valuable future work; however, due to computational and API constraints, such experiments are currently infeasible. Nevertheless, the consistent gains observed across both 3B and 7B backbones suggest that SCATTER's reward formulation is largely backbone-agnostic and benefits from increased model capacity. Notably, the 7B model performs slightly worse than the 3B model on OpenEP. This may be due to the relatively small size of OpenEP, which can increase variance for larger models, as well as potential cross-domain overfitting since training is conducted on OpenForecast while evaluation is performed on OpenEP. These observations are preliminary, and a more systematic cross-scale and cross-domain analysis is left for future work.

A.7 Impact of Sampling Rounds on Llama

The performance scaling of llama3.2-3B-Instruct across varying sampling rounds on OpenForecast are shown in Figure 6.

A.8 Case Study

We proceed to illustrate an example from OpenEP, detailing the question, background, and generated

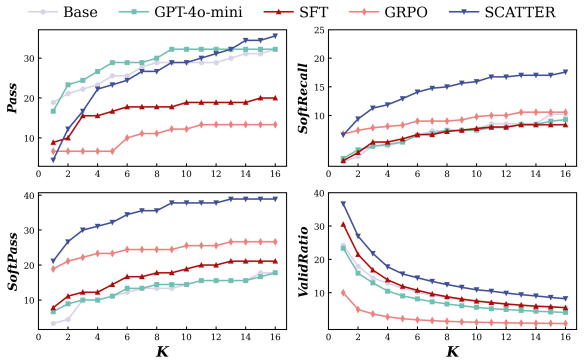

Figure 6: **Impact of sampling budget K on generation performance.** Comparison of SCATTER against baselines on OpenForecast using llama3.2-3B-Instruct as Base Model.

hypotheses from a randomly sampled round as follows:

Question & Background

Question:
2024-07-05: What key developments can be expected in the dialogue between Viktor Orbán and Vladimir Putin regarding Ukraine's situation?

Background:
Hungarian Prime Minister Viktor Orbán recently visited Kyiv for the first time since the onset of Russia's full-scale invasion in 2022, sparking significant discussions with Ukrainian President Volodymyr Zelensky. During his visit, Orbán proposed prioritizing a ceasefire to expedite peace negotiations, a suggestion that contrasts with the ongoing European support strategy for Ukraine, which leans towards military aid. This proposal was set against the backdrop of Orbán's contentious relationship with EU policies due to his close ties with Russian President Vladimir Putin, whom he is slated to meet soon after his trip to Kyiv.

While in Ukraine, Orbán emphasized the urgency of addressing the war, which he termed as Europe's "most important issue," suggesting that a ceasefire could accelerate peace talks. However, Zelensky maintained a firm stance, appreciating Hungary's humanitarian efforts but dismissing the ceasefire approach, which aligns with his previous rejections of similar proposals from other European leaders. Orbán's visit coincides with Hungary's recent assumption of the EU Council's rotating presidency, adding layers of diplomatic complexity given his government's contentious EU relations and internal democratic criticisms.
Orbán's outreach to Ukraine also included discussions on improving bilateral relations and economic cooperation, despite his history of blocking EU efforts to extend more robust support to Kyiv. This diplomatic maneuver occurs as Orbán gears up for strategic dialogues in upcoming high-profile European meetings, where Ukraine's geopolitical status and EU relations will be key topics. The Hungarian leader's actions and statements during this critical visit highlight the delicate balance of regional alliances, the contentious dynamics within the EU, and the ongoing challenges in the peace process amidst Europe's most severe security crisis in decades.

Ground Truth:
[Viktor Orbán met Russian President Vladimir Putin.,
Orbán said Europe's "war policy" is hurting its people and economy.,
The meeting with Putin was an unsanctioned attempt to resolve the conflict in Ukraine.]

Hypotheses GPT-4o-mini

["Orbán may advocate for a stronger Russian influence in Ukraine during discussions with Putin.",
"Putin could propose increased military support for Hungary as a counter to EU sanctions.",
"Orbán might suggest a new economic partnership with Russia to bypass EU restrictions.",
"Putin could express discontent with Orbán's ceasefire proposal, urging a more aggressive stance.",
"Orbán may negotiate for energy supplies from Russia to stabilize Hungary's economy amid EU tensions.",
"Putin might offer concessions to Hungary in exchange for Orbán's support in the EU.",
"Orbán could push for a united front with Russia against Western sanctions.",
"Putin may highlight Hungary's importance as a strategic ally in Central Europe.",
"Orbán could seek assurances from Putin regarding Hungary's security amidst EU pressures.",
"Putin might leverage Orbán's EU presidency to influence European policies favorably towards Russia."]

Hypotheses Qwen2.5-3B-Instruct

[Orbán may propose a joint statement condemning NATO expansion.,
Putin could insist on troop withdrawal from eastern Ukraine territories.,
Orbán suggests increased economic sanctions against Russia as leverage.,
Putin might counter with demands for Ukrainian territorial integrity.,
Orbán advocates for a neutral buffer zone along the Ukrainian-Russian border.,
Putin proposes direct military assistance to Kyiv in exchange for political concessions.,
Orbán calls for a temporary truce to focus on humanitarian aid and reconstruction.,
Putin insists on a full-scale ceasefire to stabilize the region immediately.,
Orbán seeks to restart stalled peace talks with renewed commitment.,
Putin demands immediate recognition of separatist regions in Ukraine.]

Hypotheses SCATTER

<answer>
["Ukrainian President Volodymyr Zelensky and Hungarian Prime Minister Viktor Orbán engage in a second summit in Berlin, Germany.,
Orbán proposes a renewed push for a ceasefire in Ukraine amid ongoing tensions.,

Zelensky maintains a firm stance against any ceasefire agreements, citing EU support.,
The Hungarian government assumes leadership of the EU Council's presidency starting July 5, 2024.,
Orbán emphasizes the urgency of addressing the war, positioning it as Europe's most critical issue.,
A news conference occurs in Budapest, Hungary, following Orbán's return from Kyiv.,
Orbán and Zelensky agree on the need for improved bilateral relations and economic cooperation., The EU Council votes to extend sanctions against Russia, potentially complicating future negotiations.,
Orbán's visit leads to a shift in public opinion within Hungary regarding the war and EU support., A third summit takes place in Brussels, Belgium, featuring Orbán and Zelensky alongside other European leaders.]
</answer>

A.9 Prompts

Prompt Templates for Post-Training

You are an expert in event forecasting. Given a forecasting question (Forecasting Question) and its background information (Background), your task is to generate 10 possible hypotheses.

Here is the Background:
{background}
Here is the Forecasting Question:
{question}

Prompt Templates for Evaluation

You are an expert in event forecasting. Your task is to evaluate whether the Model Predictions correctly answers the Question, based on the Ground Truth (Correct Answer) provided. The answer should be true if at least one item in Ground Truth is contained within the Model Predictions, otherwise false.

Here is the Question:
{query}
Here is the Model Predictions:
{model_output}
Here is the Ground Truth:
{ground_truth}

Prompt Templates for the Backbone Model

You are an expert in event forecasting. Given a forecasting question (Forecasting Question) and its background information (Background), your task is to generate possible hypotheses.

Guidelines:
- Validity: Each hypothesis must be directly relevant to the Forecasting Question.
- Coverage: The set of hypotheses must collectively span different possible outcomes. Each hypothesis should be distinct from the others, avoiding redundancy or near-paraphrases edge cases.

Here is the Background:
{background}
Here is the Forecasting Question:
{question}

Constraints:
- Output exactly 10 new hypothesis.
- Each hypothesis should be specific and concise ($\leq$ 25 words).
- Do NOT reference or rely on any information dated after {date}.
- Output must follow the JSON schema below and contain no extra text.

Example output format:
["Prediction 1", "Prediction 2", "Prediction 3","Prediction 4",...,"Prediction 9","Prediction 10"]

Prompt Templates for Answer Generation

You are an expert in event forecasting. You are given a list of Ground Truth events (Ground Truth) and a list of model-generated hypotheses (Hypotheses). Your task is to rewrite each Ground Truth into a new hypothesis, following the same style, length, and tone as the Hypotheses.

Guidelines:
- Preserve the factual meaning of each Ground Truth, but phrase it in the same stylistic form as the Hypotheses
- Hypotheses are typically concise, forward-looking statements ($\leq$ 25 words).
- Maintain neutrality and avoid adding explanations, reasoning, or extra context.
- Ensure the wording is similar in style to the provided Hypotheses (e.g., same verb tense, tone, and abstraction level).
- Each Ground Truth must correspond to exactly one new hypothesis.

Here is the Hypotheses:
{hypotheses}
Here is the Ground Truth:
{ground_truth}

Constraints:
- Do NOT invent new events beyond the Ground Truth.
- Output must be a JSON array of hypotheses, with one entry per Ground Truth, in the same order as the Ground Truth list.

Output JSON schema:
[the number of Ground Truth here]

Prompt Templates for Answer Extraction

You are an expert in event forecasting. Your task is to convert a raw, unstructured answer paragraph into a structured list of concise "ground truth" statements. Given a Background, Question and Answer, extract the core factual points solely from the "Answer" text that directly answer the question.

Guidelines:
- Remove conversational filler (e.g., "Indeed," "On the other hand," "According to a survey").
- Each statement should be concise (aim for $\leq$25 words) but preserve key entities and numbers.
- Statements must be derived strictly from the "Answer" section. Do NOT include or infer information found only in the "Background", even if it provides context. Do NOT invent new events.
- Each Ground Truth must directly relate to answering the specific Question provided.

Here is the Background:
{background}
Here is the Question:
{question}
Here is the Answer:
{answer}

Output JSON schema:
["Ground Truth 1","Ground Truth 2",...]