Yan, Q., Seraj, R., He, J., Meng, L., and Sylvain, T. (2024). Autocast++: Enhancing world event prediction with zero-shot ranking-based context retrieval. In *International Conference on Learning Representations (ICLR)*.

Zhang, Y., Chen, X., and Park, D. (2018). Formal specification of constant product (xy= k) market maker model and implementation. *White paper*.

Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Dou, Z., and Wen, J.-R. (2024). Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

Zou, A., Xiao, T., Jia, R., Kwon, J., Mazeika, M., Li, R., Song, D., Steinhardt, J., Evans, O., and Hendrycks, D. (2022). Forecasting future world events with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Details about Models and Knowledge Accuracy

A.1 Models

We give a list of detailed information of the models we use below. The weights of the open models are available publicly on Hugging Face, and we primarily use Together AI's serving API to access them. All cut-offs are based on official statements.

Model	Source	Open Weights	Knowledge Cut-off	Evaluation Cost
GPT-4-1106-Preview	OpenAI	No	Apr 2023	\$0.01/1K tokens
GPT-4 (GPT-4-0613)	OpenAI	No	Sep 2021	\$0.03/1K tokens
GPT-3.5-Turbo-Instruct	OpenAI	No	Sep 2021	\$0.0015/1K tokens
GPT-3.5-Turbo-1106	OpenAI	No	Sep 2021	\$0.001/1K tokens
Claude-1	Anthropic	No	Dec 2022	\$0.024/1K tokens
Claude-2	Anthropic	No	Dec 2022	\$0.024/1K tokens
Claude-2.1	Anthropic	No	Dec 2022	\$0.024/1K tokens
Llama-2-7B-Chat	Meta	Yes	Sep 2022	\$0.0002/1K tokens
Llama-2-13B-Chat	Meta	Yes	Sep 2022	\$0.00025/1K tokens
Llama-2-70B-Chat	Meta	Yes	Sep 2022	\$0.0009/1K tokens
Mistral-7B-Instruct	Mistral AI	Yes	*Unknown*	\$0.0002/1K tokens
Mistral-8x7B-Instruct	Mistral AI	Yes	*Unknown*	\$0.0002/1K tokens
Mixtral-8x7B-DPO	NousResearch	Yes	*Unknown*	\$0.0002/1K tokens
YI-34B-Chat	01.AI	Yes	June 2023	\$0.000776/1K tokens
Gemini-Pro	Google	No	Early 2023	\$0.0005/1K characters

Table 6: **Overview of the LMs we evaluate**: A breakdown of the LMs used in our study, including their sources, availability of weights, knowledge cut-off dates, and evaluation costs. The evaluation costs of the open-weight models are based on Together AI's pricing. The knowledge cut-off of Gemini-Pro is claimed to be early 2023 ($\sim$ April 2023). We are not aware of the exact knowledge cut-offs of the Mistral series, as it is not publicly reported.

A.2 Testing Potential Leakage from Post-training

GPT-4-1106-Preview and GPT-3.5-Turbo-1106, the two models we use in our system, were released in November, 2023. We find no evidence that the post-training phase leaks further information after their knowledge cut-offs (April, 2023 and January, 2021). As a test, we manually query the model on 20 major events in June, 2023–September, 2023[3], such as "Who won the 2023 Turkish presidential election?". For all 20 questions, both models either claim no knowledge or simply hallucinate.

As a sanity check, we also prompt GPT-4-1106-Preview to answer another 20 questions about events during November, 2022–January, 2023, prior to its knowledge cut-off, such as "Which team won the 2022 FIFA World Cup Final?". The model answers all of them correctly.

A.3 Crowd Predictions

On any given question, each platform computes a community prediction that aggregates all individual forecasts. The prediction is dynamically updated and recorded as the forecasts are made. We source the records directly from the platforms (instead of computing them from scratch using the individual forecasts). For binary questions, we provide more details on the aggregation mechanisms as follows.

- On Metaculus, for a given question, each prediction of a forecaster is marked by t (starting at 1), from their earliest prediction to the latest. The platform computes the crowd prediction of the question by weighted median. The weight of the tth forecast from an individual forecaster is $e^{\sqrt{t}}$, so the more recent

[3] sourced from `https://www.onthisday.com/events/date/2023/`.

forecasts receive higher weights. We remark that the platform also publishes another aggregated forecast called “Metaculus prediction” (which we do not use or compare with in this paper). This differs from the crowd prediction described above and is computed via a proprietary algorithm.

- GJOpen computes the crowd predictions by the mean of the most recent 40% of the forecasts from each forecaster.
- INFER initializes the crowd prediction to be the mean of all individual forecasts. As the question progresses, it reweights the forecasts, for example, by “putting more weight on the forecasts of individuals with the best track record.”[4] Exact details on the aggregation mechanisms are not found on their website.
- Manifold and Polymarket are prediction markets, where the community predictions are the prices (between 0 and 1). The prices are adjusted by their automated market makers, as bets are made. The mechanisms are variants of constant-product market makers (Hanson, 2007; Zhang et al., 2018); see Polymarket (2023); Manifold (2022) for more details.

B Details about Base Evaluations

In this section, we provide experimental details on our baseline evaluations (Section 3.4).

B.1 Evaluation Method

For both zero-shot and scratchpad prompting, we conduct basic prompt optimization by by crafting 5 candidate zero-shot prompts and 4 candidate scratchpad prompts. We evaluate each prompt on the validation set by comparing Brier scores. Specifically, we randomly select 200 questions from the validation set and calculate the mean Brier scores across the 14 LMs under consideration.

- The best zero-shot prompt achieves an average Brier score of 0.246, outperforming the others, which score 0.261, 0.276, 0.279, and 0.252, respectively.
- For scratchpad, all prompts yield similar Brier scores. We observe that potentially due to safety training, models can sometimes refuse to answer forecasting questions by simply claiming “I don’t know”. Therefore, we use the number of “refuse to answer” responses as the deciding metric. The winning scratchpad prompt averages 88 “refuse to answer” responses, while the others average 106, 93, and 94, respectively.

The best zero-shot and scratchpad prompts are shown in Figure 5 and Figure 6. In both prompting styles, models are only provided with the question, background, resolution criterion, and question’s open and close dates (`date_begin` and `date_end`). All the data are sourced from the forecasting platforms and publicly available on the question page to human forecasters. We do no additional news retrieval.

Finally, we use the best prompt of each prompting strategy to forecast on each question in the test set. In Section 3.4, we find that none of the models are naturally good at forecasting. We provide the full results next in Section B.2.

[4] https://www.infer-pub.com/frequently-asked-questions