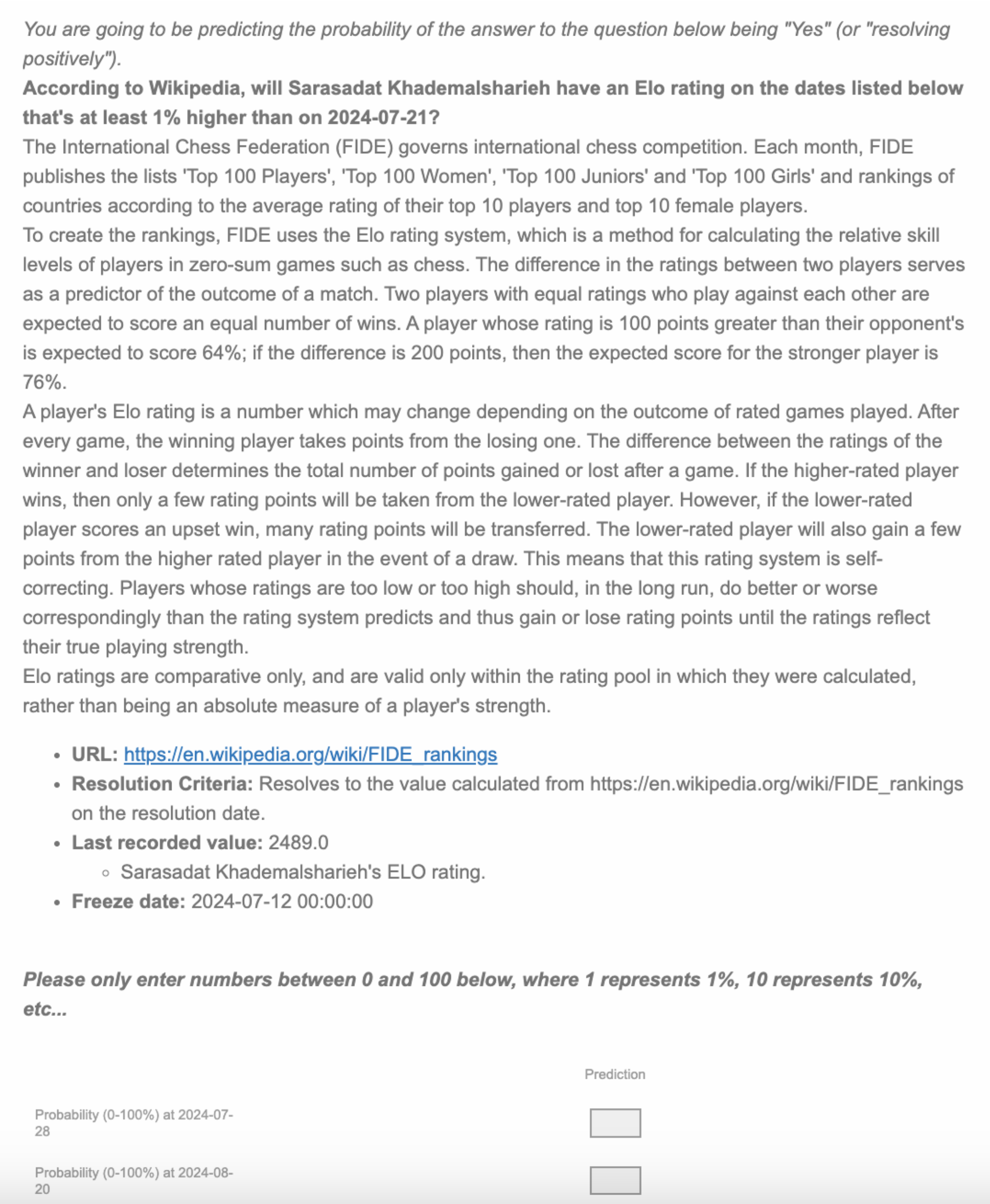

You are going to be predicting the probability of the answer to the question below being "Yes" (or "resolving positively").

According to Wikipedia, will Sarasadat Khademalsharieh have an Elo rating on the dates listed below that's at least 1% higher than on 2024-07-21?

The International Chess Federation (FIDE) governs international chess competition. Each month, FIDE publishes the lists 'Top 100 Players', 'Top 100 Women', 'Top 100 Juniors' and 'Top 100 Girls' and rankings of countries according to the average rating of their top 10 players and top 10 female players.

To create the rankings, FIDE uses the Elo rating system, which is a method for calculating the relative skill levels of players in zero-sum games such as chess. The difference in the ratings between two players serves as a predictor of the outcome of a match. Two players with equal ratings who play against each other are expected to score an equal number of wins. A player whose rating is 100 points greater than their opponent's is expected to score 64%; if the difference is 200 points, then the expected score for the stronger player is 76%.

A player's Elo rating is a number which may change depending on the outcome of rated games played. After every game, the winning player takes points from the losing one. The difference between the ratings of the winner and loser determines the total number of points gained or lost after a game. If the higher-rated player wins, then only a few rating points will be taken from the lower-rated player. However, if the lower-rated player scores an upset win, many rating points will be transferred. The lower-rated player will also gain a few points from the higher rated player in the event of a draw. This means that this rating system is self-correcting. Players whose ratings are too low or too high should, in the long run, do better or worse correspondingly than the rating system predicts and thus gain or lose rating points until the ratings reflect their true playing strength.

Elo ratings are comparative only, and are valid only within the rating pool in which they were calculated, rather than being an absolute measure of a player's strength.

- **URL:** https://en.wikipedia.org/wiki/FIDE_rankings
- **Resolution Criteria:** Resolves to the value calculated from https://en.wikipedia.org/wiki/FIDE_rankings on the resolution date.
- **Last recorded value:** 2489.0
    - Sarasadat Khademalsharieh's ELO rating.
- **Freeze date:** 2024-07-12 00:00:00

Please only enter numbers between 0 and 100 below, where 1 represents 1%, 10 represents 10%, etc...

	Prediction
Probability (0-100%) at 2024-07-28	□
Probability (0-100%) at 2024-08-20	□

Figure 3: An example question generated from a data provider, in this case DBnomics, from the public survey. Two of eight forecast horizons for which we elicited forecasts are included above. The rationale text boxes (one for each forecast horizon) have also been excluded from the screenshot for brevity.

E LLM "ENSEMBLE" BASELINE

E.1 MODELS

To construct an ensemble baseline that includes diverse candidates, we evaluate models using the most recent forecasting dataset containing cross-domain questions with true resolutions from Halawi et al. (2024). We assess models from the following organizations: OpenAI, Mistral AI, Qwen, Google, Anthropic, and Meta. Using the same scratchpad prompting method from Halawi et al. (2024), we then select the top three models: GPT-4o, Gemini-1.5.Pro, Claude-3.5-Sonnet. See Table 16 for the results.

Table 16: Brier Scores from each LLM "crowd" candidate.

Model	Scratchpad
GPT-4o	**0.207 (0.026)**
Llama-3-70b	0.232 (0.020)
Mistral-Large	0.233 (0.026)
Qwen-1.5-110b	0.222 (0.025)
Gemini-1.5-Pro	**0.214 (0.025)**
Claude-3.5-Sonnet	**0.178 (0.025)**
GPT-4-0613	0.222 (0.009)
GPT-4-1106-Preview	0.209 (0.012)
GPT-3.5-Turbo-1106	0.261 (0.010)
GPT-3.5-Turbo-Instruct	0.257 (0.009)
Claude-2	0.219 (0.014)
Claude-2.1	0.215 (0.014)
Gemini-Pro	0.230 (0.007)
Mistral-7B-Instruct	0.243 (0.008)
Mistral-8x7B-Instruct	0.238 (0.010)
Mixtral-8x7B-DPO	0.248 (0.010)
Yi-34B-Chat	0.241 (0.009)
Llama-2-7B	0.264 (0.011)
Llama-2-13B	0.268 (0.008)
Llama-2-70B	0.282 (0.011)

The highlights indicate the models we decide to use. Subscript numbers denote 2 standard errors. Above the divider line are the new models we evaluate, while below the divider line are the models evaluated by Halawi et al. (2024) on the same dataset using the same prompt.

E.2 AGGREGATION METHODS

With the forecasts generated by the top 3 models selected in Section E.1, we then compare performance of 5 aggregation methods: Median, Trimmed Mean, Geometric Mean, and Geometric Mean of Log Odds (Satopää et al., 2014). See Table 17 for the results.

E.3 OTHER DETAILS

LLM Parameters. We set the temperature to 0 and the max output token length to 2000.

F PROMPTS

In this section, we present the following prompts: zero-shot (Figure 4), scratchpad (Figure 5), and three prompts (Figure 6, Figure 7, Figure 8) written by superforecasters from the Forecasting Research Institute used to construct the LLM "ensemble" baseline. Note that the scratchpad with information retrieval prompt is simply the scratchpad with an additional line "We have retrieved

Table 17: Brier Scores from each Aggregation Method.

Aggregation Method	Brier Score
Median	0.197 (0.024)
Trimmed Mean	0.197 (0.024)
Geometric Mean	**0.194 (0.023)**
Geometric Mean of Log Odds	**0.194 (0.023)**

The geometric mean and geometric mean of log odds are highlighted as they achieve the best performance. Subscript numbers denote 2 standard errors.

```
You are an expert superforecaster, familiar with the work of Tetlock and others. Make a prediction of the probability that the question will be resolved as true. You MUST give a probability estimate between 0 and 1 UNDER ALL CIRCUMSTANCES. If for some reason you can't answer, pick the base rate, but return a number between 0 and 1.
Question: {question}
Question Background: {background}
Resolution Criteria: {resolution_criteria}
Question close date: {close_date}
Output your answer (a number between 0 and 1) with an asterisk at the beginning and end of the decimal. Do not output anything else.
Answer:
{{ Insert answer here }}
```

Figure 4: Zero-shot Prompt from Halawi et al. (2024)

the following information for this question: {retrieved_info}" before the instructions begin. For combination questions, we slightly modify each of the above prompts by including the second question's information (Figure 9). Finally, we present the prompt used to determine whether a question is marked as valid or invalid in Figure 10.

G STATISTICAL DETAILS

While our pairwise bootstrapped p-values are precise, our statistical tests ignore one potentially important source of imprecision. We assume that each question is independent of the other questions. In a world where forecasting accuracy is quite correlated within a topic, or where most events are correlated, this may result in us overstating how confidently we can reject the equivalence of different models. We hope to explore this question in future work, but we are somewhat reassured by the fact that our gathering of forecasting questions from diverse domains and sources makes it unlikely that correlated questions would change our interpretation of these results in any meaningful way.

Our statistical tests are significantly more precise than the 95% confidence intervals for each model would imply because accuracy on each question is quite correlated across models. To understand this phenomenon, consider a hypothetical world where Model A outperforms Model B by a constant (ϵ) on each question. No matter how close the performance of Models A and B are (how small ϵ is), and no matter how much variance there is in the accuracy of Model A across questions (which drives the 95% confidence interval surrounding Model A's accuracy), a pairwise bootstrap would show that Model A is more accurate than Model B. This is because forecasts of Model A and Model B are perfectly correlated.

H LEADERBOARDS: TOP 50

We show the best 50 performers on several leaderboards. In Table 18we show the leaderboard for the human question set of 200 standard (non-combination) questions. Table 19 shows the leaderboard for the full LLM question set of 1,000 questions. Finally, in Table 20, we present the human leaderboard with combination questions included where humans provided forecasts on both components of the combination question. We derive human forecasts for these combination questions by treating each component of the combination question as independent.