

Approaching Human-Level Forecasting with Language Models

Danny Halawi*, Fred Zhang*, Yueh-Han Chen*, Jacob Steinhardt
April 8, 2024

Outline

Preliminaries

Baseline

System

Results

Discussion

Preliminaries

- Judgemental forecasting
 - * Forecasters, assign probabilities to future events based on their own judgments.
 - * Make use of available historical data, domain knowledge, Fermi estimates, and intuition.
- Accurately predicting the future is valuable
 - * Governments rely on economic and geopolitical forecasts for decision making.
 - * Companies hire and invest based on forecasts of market conditions (Armstrong, 2001).
 - * In 2020, epidemiological forecasts for COVID-19 prompted national lockdowns across the globe (Adam, 2020).

- Source binary forecasting questions from 5 competitive forecasting platforms

Field	Information
Question	WILL STARSHIP ACHIEVE LIFTOFF BEFORE MONDAY, MAY 1ST, 2023?
Key Dates:	Open Date: 2023-04-17, Close Date 2023-04-30, Resolve Date 2023-04-20
Platform	Metaculus
Resolution	"Yes"
Background	On April 14th, SpaceX received a launch license for its Starship spacecraft...
Resolution Criteria	This question resolves Yes if Starship leaves the launchpad intact and under its own power before 11:59pm ET on Sunday, April 30th.
Community Predictions	[(2023-04-17, 0.725), (2023-04-17, 0.793), ..., (2023-04-18, 0.701), (2023-04-18, 0.698), ..., (2023-04-19, 0.696), (2023-04-19, 0.695), ..., (2023-04-20, 0.733), (2023-04-20, 0.734)]

- Forecaster (human or model) makes probabilistic forecast(s) between two dates: question's *open date* and *end date*

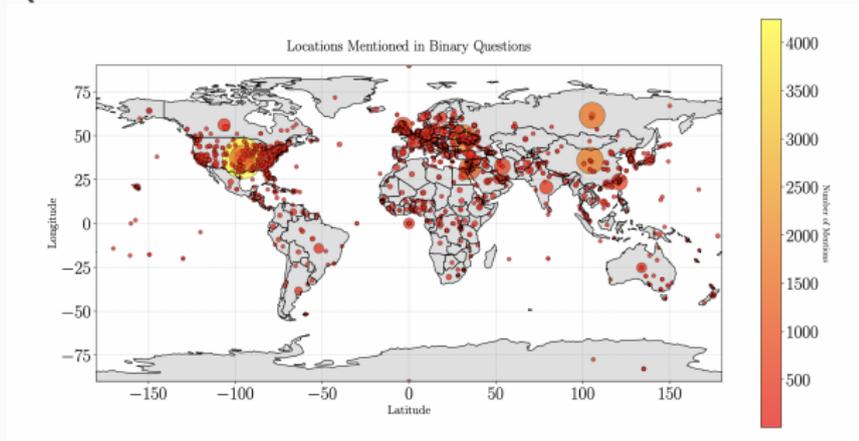
Setup as an ML Benchmark

For questions that are opened after the model's pertaining cut off

- (1) Take date between question's open date and close date:
retrieval date.
- (2) Retrieve information up to the retrieval date.
- (3) Feed information and prompt LM to make a prediction.
- (4) Compute performance using Brier score: $(\text{pred} - \text{outcome})^2$
* $0 \leq \text{pred} \leq 1$, $\text{outcome} \in \{0, 1\}$, unskilled Brier score
= .25
- (5) Compare to Brier score of crowd prediction on the retrieval date
* On Metaculus, crowd prediction is superior to the aggregate of the top 5, 10, ..., 30 forecasters

Dataset

- Train: questions before June 1st, 2023; Test: questions June 1st, 2023 (after pre-training cut off)
- Sample Questions:
 - * Will Russia successfully land on the Moon in August 2023?
 - * Will there be a case of Marburg virus in Algeria by July 28th?
 - * Will Tesla's market cap be $>$ \$1 trillion before July 1, 2023?
- Questions mention a diverse set of locations



Why use LMs for Forecasting?

Question for class:

- What advantages do they have over human forecasters?

Why use LMs for Forecasting?

- They can parse and produce texts rapidly
 - * Provide cheap and timely forecasts.
- They are pre-trained on web-scale data
 - * Endowed with massive, cross-domain knowledge.
- We can elicit their reasonings through prompts
 - * Can understand (partially) the rationale behind the forecast.
- Their knowledge is up to a specific date (cut off).
 - * Can train/test them and on events that have already been resolved, in a way we can't do with humans.

Reality (so far): Autocast++ 2024 (67% vs 92%), Autocast 2022 (65% vs 92%)

Baseline

LMs are not Naturally Good at Forecasting

- 5 platforms, 914 questions
- All questions published after June 1st, 2023 (after all LM's pre-training cut off)
- Zero-shot prompt 14 LMs (with no information retrieval)

You are an expert superforecaster, ...

Question: {question}

Question Background: {background}

Resolution Criteria: {resolution_criteria}

Question open date: {date_begin}

Question close date: {date_end}

Output a prediction (a number between 0 and 1).

Answer: {{ Insert answer here }}

Model	Zero-shot	Scratchpad
GPT-3.5-Turbo	0.237 (0.014)	0.257 (0.009)
GPT-3.5-Turbo-1106	0.274 (0.016)	0.261 (0.010)
GPT-4 (GPT-4-0613)	0.219 (0.013)	0.222 (0.009)
GPT-4-1106-Preview	0.208 (0.013)	0.209 (0.012)
Llama-2-7B	0.353 (0.020)	0.264 (0.011)
Llama-2-13B	0.226 (0.009)	0.268 (0.008)
Llama-2-70B	0.283 (0.014)	0.282 (0.011)
Mistral-7B-Instruct	0.237 (0.018)	0.243 (0.008)
Mistral-8x7B-Instruct	0.238 (0.018)	0.238 (0.010)
Mixtral-8x7B-DPO	0.260 (0.022)	0.248 (0.010)
Yi-34B-Chat	0.238 (0.012)	0.241 (0.009)
Claude-2	0.220 (0.013)	0.219 (0.014)
Claude-2.1	0.220 (0.013)	0.215 (0.014)
Gemini-Pro	0.243 (0.019)	0.230 (0.007)

- Crowd Brier score: **.149**

System

Example

Question: Will Humza Yusaf be the winner of the leadership race for the next First Minister of the Scottish Parliament?

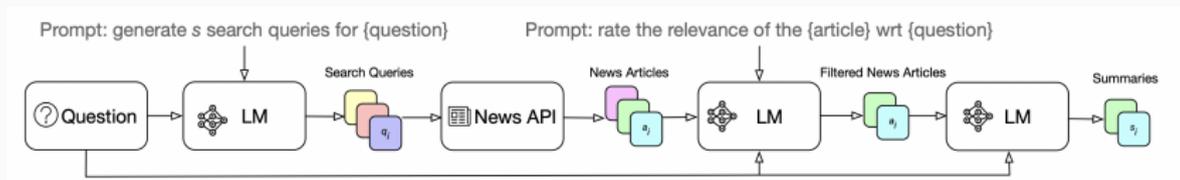
Open: 2023-03-01; **Close:** 2023-03-26; **Resolve:** 2023-03-27

Background: Nicola Sturgeon, stated her intention to step down from the role of First Minister of Scotland. There are three candidates in the running: Kate Forbes, Ash Regan, and Humza Yousaf.

Resolution Criteria: This will resolve as Yes if Humza Yousaf wins the leadership race to succeed Nicola Sturgeon.

Retrieval Date: 2023-03-14

Retrieval



Retrieval (Query Expansion)

Prompt: Generate brief search queries.. to gather information on Google that could influence the forecast.

Model Output:

Humza Yousaf leadership race odds

Scottish National Party leadership candidates

Scottish Parliament First Minister predictions

Scottish National Party election updates

Retrieval (Sub-question Expansion)

Prompt: Generate brief search queries.. Start off by writing down sub-questions. Then use your sub-questions to help steer the search queries.

Model Output:

Sub-questions

- Who are his competitors?
- What do the polls say?
- Scandals/controversies affecting Humza/his competitors?
- What is the opinion of SNP members on Humza's candidacy?
- How does his policy platform compare to his competitors'?
- What endorsements has he received?

Retrieval (Relevancy Ranking & Summarization)

<https://www.msn.com/en-gb/news/uknews/snp-leadership-contest-too-close-to-call-party-members-poll-finds/ar-AA18btaC>

Prompt: Rate relevancy from 1-6 given criteria.

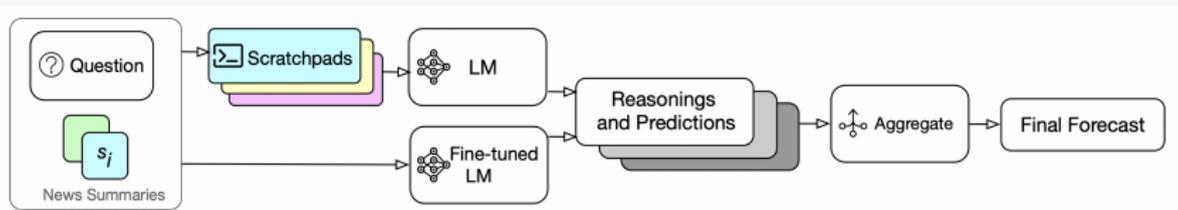
Model Output:

Thoughts: The article provides information... This information is directly relevant.. Rating: 5

Prompt: Condense article to 100 words.. Keep information relevant to question.

Model Output: ...Yousaf has 31%, Forbes 25%, Regan 11%, and 32% are undecided among 515 party members. Despite controversies over Forbes' religious views ... With two TV debates and several hustings ahead.. Preferences among decided voters are 46% for Yousaf, 37% for Forbes, and 17% for Regan,...

Reasoning



Reasoning (Base GPT-4)

Rephrased and Expanded Question: Will Humza Yousaf emerge victorious... Considering the contest's dynamics, ...

Reasons Why the Answer Might Be No: Kate Forbes' Lead: Polls have shown... Strength: Strong. Public Criticism of Yousaf's Record: Humza Yousaf has faced criticism.. Strength: Moderate....

Reasons Why the Answer Might Be Yes: Strong Endorsements: Humza Yousaf has received significant endorsements... Strength: Strong. Experience: Yousaf's long tenure... Strength: Moderate...

Aggregated Considerations: Considering the factors above, ...

Initial Probability (Prediction): Given the mixed factors... a balanced view would suggest a probability closer to the middle

Evaluation of Confidence Levels: The probability is conservative...

Final Prediction: Given all considerations, ... **0.55**

Crowd prediction: 0.52, Outcome: 1

Fine-Tuned (GPT-4)

Based on the information provided, ...

1. Decision Tree Outline: ...

Yousaf Wins: – High SNP Member Support – Strong Establishment Backing – Successful Campaign Messaging – Effective Mobilization of Undecided Voters – Positive Reception of Independence Strategy – High Voter Turnout among Supporters

Yousaf loses: ...

2. Discussion on Wildcards: ...

3. Branch Probability Analysis: ...

– Initial Probability (Prediction): 0.65

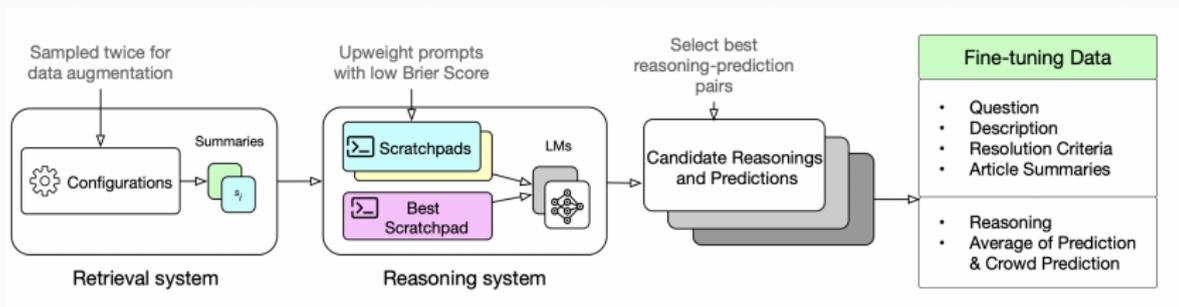
– Adjustment Thoughts:...

4. Final Answer: 0.6

Final Prediction

- 3 predictions from base GPT-4
- 3 predictions from fine-tuned GPT-4
- final prediction is trimmed mean of the 6 forecasts

Fine-tuning



Results

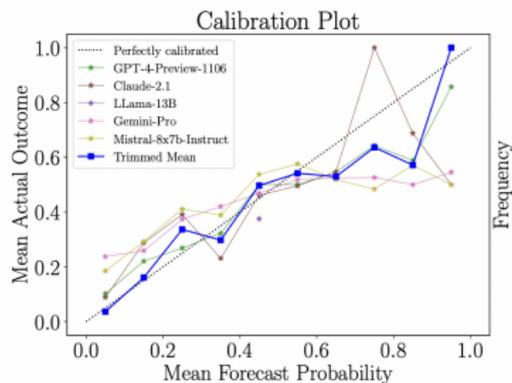
Results

- 5 Platforms, 914 questions
- All questions published after June 1st, 2023 (after all LM's pre-training cut off)
- Up to 5 retrieval dates for each question (3848 forecasts)

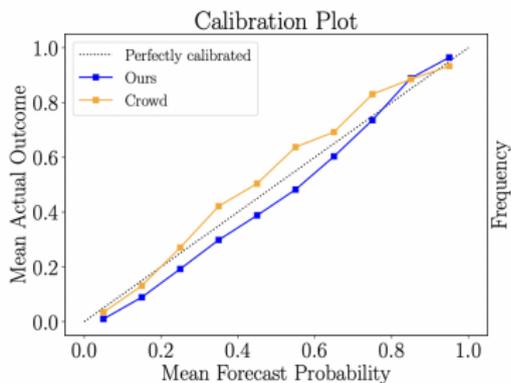
- Brier Score: **0.179 (ours)**, **0.149 (crowd)**
- Compare to prior work

Metric	Accuracy	
	System	Crowd
Ours	71.5%	77%
Autocast (Zou et al., 2022)	65.4%	92.8%
Autocast++ (Yan et al., 2024)	67.9%	92.8%

Calibration



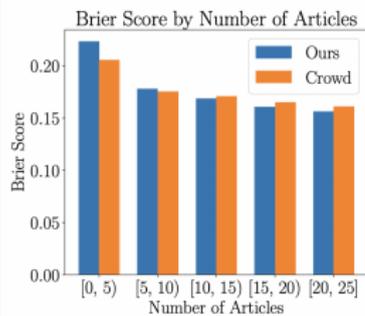
(a) Calibration of Base Models on Test



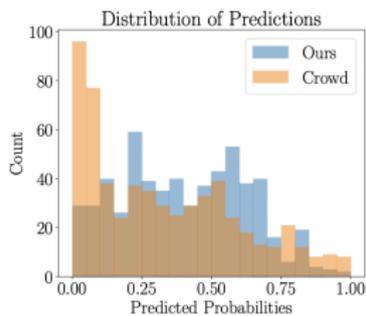
(b) Calibration of System on Test

- RMS Calibration Error: .42 (ours), .38 (crowd)

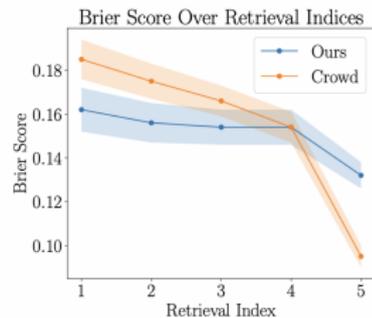
Model Strengths (Analysis on Validation Set)



(a) Brier Score w.r.t. # of Articles



(b) Distribution of predictions.



(c) Brier score over retrieval indices.

Selective Forecasting & Aggregation

- Selectively forecast where our system excels

Criteria	Brier Score ↓			% Accuracy ↑			% Data Retained ↑	
	Ours	Crowd	Aggregate	Ours	Crowd	Aggregate	Forecasts	Questions
All Questions	.179 _{.003}	.149 _{.003}	<u>.146_{.002}</u>	71.5 _{.7}	77.0 _{.7}	<u>77.8_{.6}</u>	100%	100%
Crowd Uncertain	.238_{.004}	.240 _{.003}	<u>.233_{.002}</u>	58.1 _{1.3}	58.3 _{1.3}	<u>60.2_{1.2}</u>	51%	56%
Early Retrieval	.186 _{.003}	.162 _{.004}	<u>.159_{.003}</u>	70.0 _{.9}	74.4 _{.9}	<u>75.0_{.8}</u>	84%	100%
5+ Articles	.175 _{.003}	.142 _{.003}	<u>.140_{.002}</u>	72.3 _{.8}	77.7 _{.7}	<u>78.7_{.7}</u>	84%	94%
All Criteria	.240_{.005}	.247 _{.004}	<u>.237_{.003}</u>	58.0_{1.7}	54.2 _{1.7}	56.6_{1.7}	22%	43%

- Average of our prediction with crowd prediction (aggregate) is better than either.

Ablations

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

Discussion

Path to AI Superforecasters

- Iterative self-supervision.
 - (1) fine-tune model on good predictions
 - (2) use fine-tuned model to generate predictions
 - (3) collect good predictions, repeat.
- Lots of data in the wild.
 - * stock market, sports, historical events, etc.
 - * prompt newer LM to come up with questions that older LM doesn't know
- Domain-adaptive training.
 - * Model's knowledge in a category negatively correlates with Brier score.
- LMs get better at forecasting naturally.
 - * GPT-4-1106 (2023) is better than GPT-4-0613 (2021) which is better than GPT-3.5 (2021)

Forecasting as a Benchmark for LMs

- Strong benchmark for LM capabilities
 - Model needs to synthesize information (live & pre-training knowledge), reason, and remain calibrated.
- Important for safety
 - e.g., forecast a nuclear crisis, pandemic, or AI catastrophe
- If you forecast on live questions, there's no concern of pre-training leakage. (nobody knows the label until resolution)

Repository:

https://github.com/dannyallover/llm_forecasting

Demo:

https://github.com/dannyallover/llm_forecasting/blob/main/notebooks/demo/system_demo.ipynb