Automating Arb's Big Three Report

Jonny Spicer, Sage

Summary

Frontier reasoning models show promising results for extracting and scoring forecasts from texts containing relatively explicit predictions. Attempting to score such predictions with an objective quantitative measure, such as Brier score, is inappropriate, however subjective measurements of predictive accuracy such as those used in Arb's "Big Three" report can still be informative.[1]
Our relatively naive prompt was able to extract >70% of predictions found by human readers, with at least part of the missing 30% being explained by the model choosing to cluster multiple statements into a single prediction where manual evaluators had separated into distinct, related predictions.
We scored these predictions using a similar rubric to Arb, and arrived at the same score for >60% of predictions. For predictions with differing scores, a rationale is provided by the model.
We anticipate that there are many low-hanging fruit with regards to improving the delta between our analysis and the Arb report. Furthermore, given natural differences in subjective opinion, we assert that our analysis may already be valuable in its current form.

Context

In 2022, Holden Karnofsky commissioned Arb Research to write a report expanding on his own work attempting to evaluate the track record of futurists. The rationale was that if futurists were able to make accurate predictions about 50 years in the future, then one could weight the opinions of forecasters more highly, particularly on the subject of superintelligent AI systems, and whether we might be living in the most important century.
Arb's methodology involved many manually steps, and made use of MTurk workers. Given the whole project took around 8 person-weeks, it would be very expensive to analyse similar track records with the same level of rigour. I was interested in understanding whether the dramatic improvements in publicly available LLMs since 2022 could allow for parts of the project to be automated, thereby making it feasible to cheaply evaluate other influential thinkers in a similar manner.
I only looked at automating extraction of predictions from text and scoring them, and not automating data collection - Arb notes in their report that the latter took 5 person-weeks. I would anticipate there being some small speed-ups available here with LLMs, but it generally still being a time consuming process. There are other corpuses that I believe would be worth evaluating the predictive accuracy of that would be in a readier state to analyse than that of Asimov, Clarke and Heinlein. I used broadly the same scoring rubric as the Arb report, to allow for easy comparison of the results.

[1] Although that is not without some debate; see Dan Luu's post with his own analysis of the predictions of futurists, as well as his responses to various similar analyses.

Methodology

I initially used GPT-4o to extract predictions from the text The Arb report data conveniently provides links to some of the sources; given I am most familiar with Asimov's work, I chose to focus on his predictions. One of the linked sources is a New York Times article, and sending API calls related to the text often threw a `content_filter_error`, even after I manually removed references to "thermonuclear war" which I thought may have caused the issue - now I suspect this may be due to their ongoing lawsuit[2]. There is an archived version of an article Asimov wrote in the Toronto Star that appeared to have no such issues, so I used that.

The major issue with 4o was its inconsistency; even with `temperature` set to 0 and a `seed` provided, the number of predictions extracted varied from 6 to 17 between runs. I considered breaking the document into chunks, but it wasn't particularly long in the first place, and it was preferable to keep as much context as possible. Fortunately o1 was significantly more consistent, varying between 12 and 17 predictions per run with `reasoning effort` set to high. These calls took roughly two minutes each. The model was also instructed to only extract predictions related to technology in some way, in order to allow direct comparison with Arb's results.

I then scored each prediction using the same system as Arb; 2 points for "unambiguously right", 1 for "ambiguous or near miss" and 0 for "unambiguously wrong". I didn't attempt to evaluate the prediction category or difficulty - I would expect the former to be relatively straightforward but the latter to be trickier. There was no meaningful difference between 4o and o1. I tweaked the prompt a few times to try to bring its scores more in line with the Arb ones (generally our scores were harsher), and anticipate that if I used more examples I would be able to achieve ~90% identical scoring.

Results

Overview

Total Arb predictions	Total Sage predictions	Matched predictions	Matched predictions with identical scores	Arb predictions not in Sage set	Sage predictions not in Arb set
14	15	10	7	4	0

Arb predictions and Sage predictions were matched using an LLM call; there are some duplicates where multiple Sage predictions are matched with the same Arb prediction (in part because the Arb predictions are generally vaguer than the Sage ones. I encouraged the model to make its predictions as specific as possible).

[2] Interestingly I was able to get a legitimate response when making the exact same query through the chat interface.

There was some slight variation in results between runs; there are various techniques that could be applied in order to reduce this, but I didn't think they were worth the time in order to prove this concept. The results above are from a single run that I feel is representative.

Predictions in Arb set but not Sage

- Fertility tending towards replacement
- Active anti-fertility policies in place (non-birth control)
- Space station
- Return to the moon "in force"

My hypothesis for the model not catching the first two is that it ruled them to be unrelated to technology, and for the second two that it clustered them with other, similar predictions about space that it did catch.

Differences in scoring

See output data for examples of differences in scoring, including rationales for both Sage and Arb.

Proposed next steps

The results above suggest that frontier reasoning models are able to extract predictions from text with similar accuracy of humans, when the text in question has somewhat explicit predictions. My greatest uncertainties about trying to scale this approach up are:

1. Will it have similar success on texts where predictions are implicit?
2. Will anyone find the results valuable?
    a. What scoring rubric would people find valuable and trust LLMs to have implemented with sufficient accuracy?

Feedback

I am particularly interested in the following questions:

1. Which track record(s) would you find valuable to have evaluated in a similar way to Asimov, Clarke and Heinlein's, as in the Arb report?
2. What would you want to see from an LLM-based evaluation that would give you confidence that the results are meaningful and accurate?