

Can Artificial Intelligence Forecast Treatment Effects in Economics?

Eva Vivalt*

October 28, 2026

Abstract

I use LLMs to forecast treatment effects of research studies pre-registered on the AEA RCT Registry and the Open Science Framework. I focus on studies whose results were first publicly released in 2026 and elicit forecasts from models whose training or knowledge cut-off dates were before the relevant set of papers was released. First, LLM models extract forecasting questions from the pre-registrations or pre-analysis plans. These questions elicit forecasted treatment effects; not all resolve in the subsequent papers, but those that do can be used to consider forecast accuracy. Forecasts were elicited for 889 treatment effects from 39 papers with results disclosed in January 2026 or later, or 673 from 33 papers after screening. Forecasts were elicited in raw units but I focus on those that could be standardized and therefore compared across papers. Currently, the median absolute forecast error is 0.054 SD for the best model, the average absolute error is 0.093 SD, and the MSE is 0.0218. The AI forecasts perform substantially better than benchmarks forecasting 0 effect or forecasting the mean effect of other papers' results in that time period. Overall, these results suggest considerable scope for forecasts to be of value in informing policy and research.

JEL codes: C53, C45, C93.

*University of Toronto eva.vivalt@utoronto.ca

1 Introduction

How well can we predict the results of a study before its findings are known? Decision-makers regularly draw on existing evidence to anticipate the effects of a treatment, but often high-quality evidence does not exist and they need to make their best guess. Large language models offer a way to generate predictions from descriptions of research projects before they resolve, however, the accuracy of these models is not known. Research on human forecasts documents they contain predictive power as well as systematic errors ([DellaVigna and Vivalt, 2025](#); [Dreber et al., 2015](#); [Camerer et al., 2016](#)), and recent evidence establishes that AI can also predict effects in social science experiments ([Ashokkumar et al., 2026](#)). This paper shows that LLMs can forecast the magnitudes of treatment effects across a broad range of economic interventions using pre-registrations and pre-analysis plans. Their forecasts are informative not only about expected effects, but also about uncertainty: predictive intervals have approximately their stated coverage, and wider intervals identify effects that are harder to predict.

This paper considers recent projects registered on the AEA RCT Registry and the Open Science Framework (OSF). The pre-registrations provide descriptions of planned trials in economics. I link these registrations to their earliest identified public results. I restrict attention to those with papers first shared in 2026, in order to use results released after the reported knowledge cutoff of GPT-5.5 in December 2025 and look at subsets from February 2026 and May 2026 in order to include still more recent models like Opus 4.8, Fable 5, and GPT-6 Astra. More capable models, as measured by the Epoch Capability Index (ECI), seem to do a better job forecasting than older models, though the sample sizes available after their knowledge cut-off dates are correspondingly smaller. In total, there were 889 forecasting targets for standardized treatment effects based on results for 39 papers. 673 of these, from 33 papers, passed some screening quality checks, and these form the main analysis sample.

The median absolute forecast error for the most recent model (GPT-6 Astra, on a sample of 339 forecasts from 20 papers released since May 2026) is 0.054 SD, with a mean of 0.093 and a MSE of 0.0218. For comparison, using the leave-one-paper-out mean within a time period to forecast effects in the remaining paper would produce a median absolute error of 0.092 SD, a mean of 0.126, and a MSE of 0.0463. Another possible point of comparison is [Vivalt \(2020\)](#), which collected results

from 635 impact evaluations across 20 development interventions. Across 47 intervention-outcome combinations in that paper with at least 3 papers, a leave-one-out meta-analysis would predict the results of another study with a median absolute forecast error of 0.112 SD, mean of 0.182, and MSE of 0.102.¹ While this sample of papers is different, and in principle a mixed model that included useful explanatory variables could do better at forecasting, this is nonetheless a large gap in forecasting performance. AI has been doing better than humans on other types of forecasting tasks, such as forecasting geopolitical events ([Forecasting Research Institute, 2026](#)), and the models are only improving, so there seems to be considerable scope for AI forecasts to usefully inform decisions that would benefit from estimates of treatment effects.²

A growing literature uses LLMs to simulate human responses and economic behavior ([Argyle et al., 2023](#); [Horton, 2023](#)). [Ashokkumar et al. \(2026\)](#) use simulated responses to predict experimental treatment effects, studying primarily survey experiments in psychology and political science, alongside a set of megastudies. This paper examines forecasts across a broad range of economic interventions, asking models to predict treatment effect magnitudes from pre-registration materials.

This paper makes four main contributions. First, it shows that LLMs can directly forecast economic treatment effects with small errors across a range of interventions and settings. Second, it compares these forecasts against simple prediction benchmarks. Third, it illustrates that AI forecasts are not only good at estimating the mean effect, but are also pretty well-calibrated in their 10th–90th percentile predictive intervals and in “knowing when they do not know”: wider intervals identify effects that are harder to predict. Finally, it shows that more accurate forecasts are obtained for more detailed protocols.

These results imply that AI forecasts can be useful for informing priors in policy and research. They also suggest that to be most useful, more information should be shared. Finally, while this work suggests a role for forecasts, it also indirectly provides a perhaps sobering assessment of current economics research: if results can already be forecast so well, what is gained from research that only modestly overturns priors? Should we not be pursuing more ambitious projects?

¹Weighting intervention-outcome groups equally and observations equally within groups.

²RCT evidence may remain the gold standard, however, RCT evidence is frequently not available to decision-makers, so AI forecasts only need to improve on human forecasts in this case.

2 Data and forecast construction

I construct a dataset linking public descriptions of planned research with resultant reported treatment effects. Figure A.1 summarizes the steps from identifying papers to constructing the evaluation sample.³ The following sections describe each of these stages in more detail.

2.1 Study identification and source materials

I draw on three sources of study descriptions: public pre-analysis plans posted on the AEA RCT Registry, public pre-registrations without public pre-analysis plans posted on the AEA RCT Registry, and public pre-registrations on OSF. Registrations without a public pre-analysis plan are considered usable when their descriptions provide enough information for a LLM to formulate a numerical treatment effect forecasting question.⁴

I restrict attention to outcome families whose earliest documented public findings are in 2026, based on systematic web searches, and record the date of first public release.⁵ To search for results, registries are searched for linked results papers; targeted searches of working paper repositories are implemented, such as searches of NBER papers that first came out in 2026; and public materials on the internet are searched.

This approach identified 89 candidate records, of which 49 papers satisfied the paper-level requirements concerning when results were seemingly first made publicly available and whether pre-registration information was sufficiently detailed to enable forecasting treatment effects that could later be standardized.⁶ The 49 candidates at this stage comprise 21 papers with public AEA pre-analysis plans, 25 with AEA registrations without public plans, and three with OSF pre-registrations.

These papers' protocols must still support sufficiently detailed questions that can be matched

³Source documents, questions, extracted results, matching decisions and forecast responses underlying these steps are preserved in a replication package, though the responses to the prompts are non-deterministic.

⁴Registrations without public pre-analysis plans were screened for sufficient information to generate numerical treatment effect forecasting questions.

⁵It is possible that some results may have made it into the training data for a particular model regardless. I do two things to mitigate this concern. First, another LLM looked for papers where a large share of treatment effects are exactly or nearly-exactly estimated and inspected every instance (such as cases where treatment effects are near 0 and forecast to be near 0). Second, I perform sensitivity analyses to dropping each paper in turn.

⁶Forecasts were elicited in raw units; information on standard deviations could come from either the pre-registration documents or the resultant paper.

with results and standardized. The resultant sample is predominantly experimental. Of the 33 papers in the screened January-onward sample, 31 use some form of randomized design. The two nonrandomized studies, both found in OSF, examine school phone bans and SNAP sugary-drink restrictions using difference-in-differences designs. The predominance of randomized control trials (RCTs) in the corpus has several benefits. For example, it may mitigate the influence of selective reporting (Brodeur et al., 2020; Vivalt, 2019), and if a paper later reports results with slightly different controls than initially specified, it is unlikely to substantively affect treatment effect estimates.

2.2 Constructing forecasting questions

The pre-registration and pre-analysis plan materials were used to construct forecasting questions without access to the resultant papers. Claude Opus 5 was used to initially propose questions, while GPT-6 Astra was used to review Claude’s outputs for coverage, consistency, duplicate questions, units, and clarity. Explicit conventions were used to determine how to treat questions that were almost fully specified but for some detail. The research package retains the prompts, submitted materials, responses and dispositions.

Questions specify the intervention, treatment arm, outcome, sample, time period, unit, and specification (if provided in the pre-registration documents). Questions are not restricted to primary outcomes but also include heterogeneous treatment effects and results reported at different time periods. If a pre-registration is unclear on some dimensions, such as not reporting the time period, it is not excluded on that basis, however, each question requires an identifiable intervention or treatment contrast, substantive outcome, relevant population, and numerical scale; any permitted conventions for unspecified details are stated explicitly.

Using these broad criteria, 9,922 questions were extracted from 44 papers. Most, however, could not be connected to unique numerical results in the papers, as described in the next section.

2.3 Extracting and matching reported results

A separate numerical extraction step extracted values from the identified papers. The intervention, treatment arm, outcome, sample, time period, unit and specification, where available, were

logged and the closest match selected between the extracted results and each forecasting question. Forecasting questions and results were matched 1:1, with duplicates removed. Exact matches in the content between a forecasting question and a result were distinguished from close but inexact matches. If a result was reported in a way that flipped the sign from the original forecasting question, that was also identified and results flipped before analysis. Results were extracted in the way they were reported (e.g., a treatment coefficient and standard error vs. a treatment group mean, standard deviation, and number of observations, along with a control group mean, standard deviation, and number of observations). Results were also read off of figures, where applicable, and a flag was added to indicate if a result was extracted in this way.

Additional review rounds checked the identified results against the supplied papers and assessed correspondence to the forecasting question. A “match score”, on a 0–100 scale, provides a continuous measure of the degree of match between the description in the pre-registration and the paper for each specific question.

Overall, 1,866 matched question–result pairs were produced from 43 papers.

2.4 Reporting units and standardization

Forecasts were elicited in “raw” units where feasible and converted to comparable standardized effects. A compatible baseline outcome standard deviation was preferred for standardizing where available; otherwise, a documented compatible control-group, pooled, follow-up, or other outcome SD was used. For binary outcomes, the SD was constructed from reference probabilities.

Of the 1,866 matched pairs from 43 papers, 889 questions from 39 papers yielded usable standardized results.

2.5 Forecast elicitation and eligibility

Forecasting requests contain the study protocols, forecasting questions and reporting instructions, but no results, and are implemented through the API as a closed-book exercise without any web search/browsing/retrieval/or coding tools enabled. The prompts request a mean forecast and distributional quantiles. The main analyses use xhigh reasoning effort for GPT-5.5 and GPT-6 Astra and max effort for the Anthropic models, with a 64,000-token output cap. GPT-5.1 uses high, its

maximum available effort, with a 32,000-token cap (Appendix Table B.1). Additional checks verify information such as units and time periods.

This paper uses GPT-5.1, GPT-5.5, Opus 4.6, Opus 4.8, Fable 5 and GPT-6 Astra. Papers from January onwards are considered for GPT-5.1, GPT-5.5 and Opus 4.6, papers from February onwards are used for Opus 4.8 and Fable 5, and papers from May for Astra, per provider-reported model cutoffs (Table A.1).

For the analysis, a question must have a correspondence score of at least 60 and must not have a demonstrated major mismatch according to two eligibility screens. A major mismatch requires evidence that the result answers a substantively different question regarding the intervention, treatment arm, outcome, sample, time period, or unit. The correspondence screen excludes 193 of the 889 targets. I also exclude an outcome family if numerical or qualitative findings on it were publicly documented before the start of the model’s evaluation window.⁷ Applying these criteria leaves 673 questions from 33 papers in Jan+, 629 from 30 in Feb+, and 339 from 20 in May+, as shown in Figure A.1. Table A.2 provides further details on the resultant set of papers and forecasting questions and their sources. The 33 papers in the January-onward screened sample include 16 with public AEA RCT Registry pre-analysis plans, 14 with AEA RCT Registry pre-registrations without public pre-analysis plans, and 3 with OSF pre-registrations. The papers retained in the screened analysis sample are listed in Appendix C.

To assess accuracy, I calculate squared errors in SD units, average within each paper, and then weight papers equally. Neither results nor forecasts are winsorized. The 10th and 90th percentiles were also elicited, defining an 80% predictive interval.

3 Results

3.1 The distribution of treatment effects and forecasts

Figure 1 plots distributions of reported treatment effects and forecasts. Each question receives equal weight. For the sake of this figure, questions are oriented so that a positive treatment

⁷For example, in one case, results for different outcomes pertaining to the same intervention were available in companion papers. Outcomes that were clearly related were removed, but it is possible that even outcomes that appear to be different could be correlated. Results dropping particular papers will be presented for robustness.

effect represents a good outcome. There is a long right-hand tail of realized treatment effects and relatively few negative results or forecasts of negative results.

3.2 Forecast accuracy

Table 1 reports the MSE by model and time period. Results for older models are also shown in the Feb+ and May+ subset to enable closer comparison with models only eligible beginning in those time periods, since the samples otherwise differ. MSEs are generally small and declining for more recent models.

The table also shows that the AI forecasts perform well compared to benchmarks of always predicting 0 as a treatment effect or predicting the leave-one-paper-out mean within the sample.⁸ To further benchmark these results, I compare forecasts against the “best guess” that one might make through meta-analysis. In particular, I use the data set in Vivalt (2020), formed from a harmonized set of 20 meta-analyses and systematic reviews, which cover 47 intervention-outcome combinations (e.g., the effects of unconditional cash transfers on enrollment rates). For each intervention-outcome combination, I repeatedly drop one paper and see what the MSE would be if it had been estimated by all the other papers in the set. Results are shown in Appendix D – these leave-one-out results perform much worse than simply predicting 0 or the leave-one-paper-out mean, perhaps reflecting the different sample.

Appendix Figure A.2 shows the distribution of absolute forecast errors. Appendix Figure A.3 plots the MSE of different models against their Epoch Capability Index; while there are only a few models, there do seem to be qualitative improvements in the models over time.

⁸This benchmark uses realized results, rather than only information available when a forecast could have been made, which may advantage it if treatment effects change over time.

Reported effects and forecasts

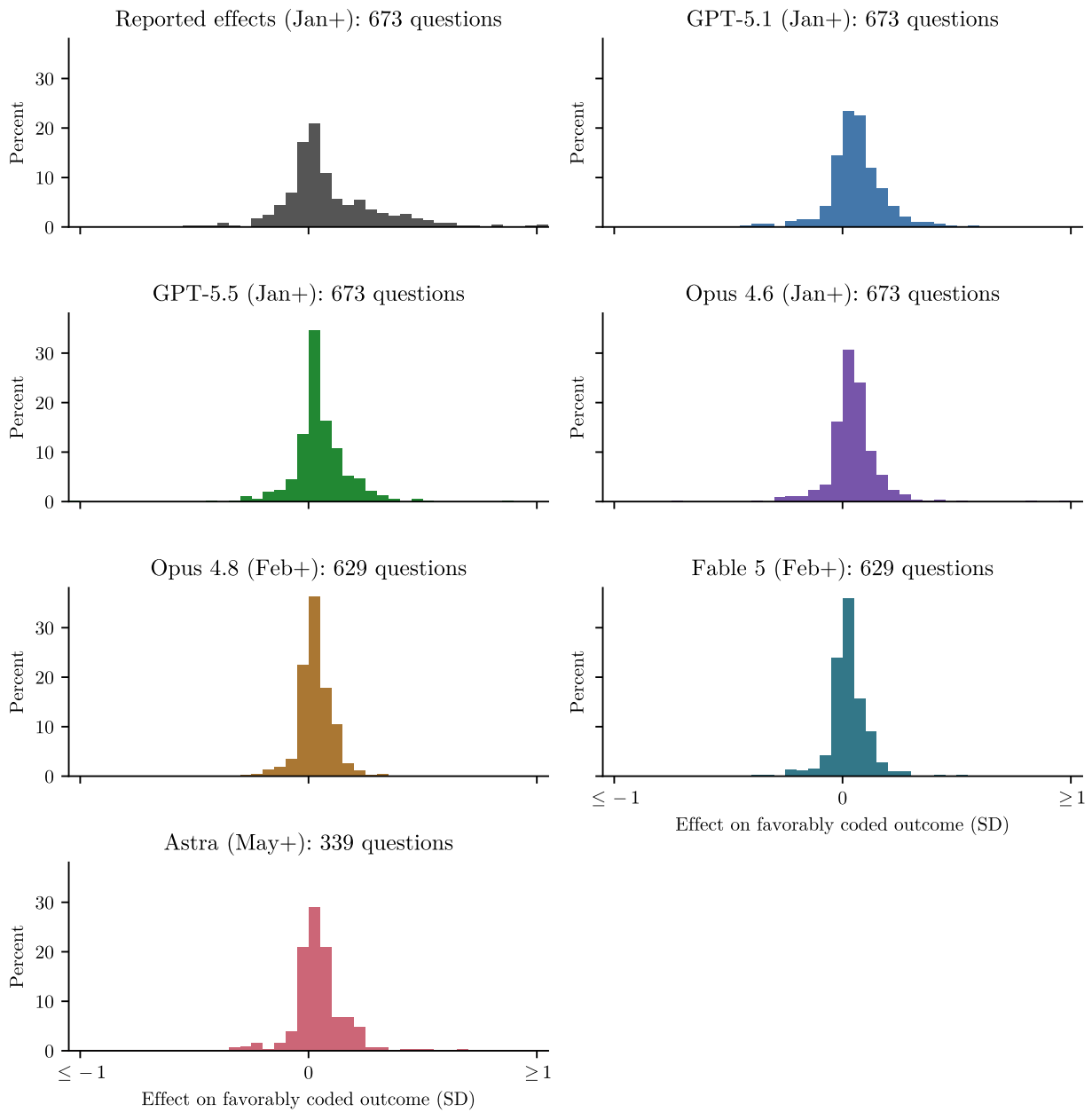


Figure 1: Reported effects and forecasts

Outcomes are oriented so that an increase is considered a positive outcome. Each question has equal weight. The tails are binned in end bins.

Table 1: Forecast MSE by model and period

	Jan+	Feb+	May+
<i>AI models</i>			
GPT-5.1	0.0365	0.0375	0.0286
GPT-5.5	0.0346	0.0352	0.0307
Opus 4.6	0.0369	0.0377	0.0350
Opus 4.8	—	0.0337	0.0308
Fable 5	—	0.0256	0.0236
Astra	—	—	0.0218
<i>Benchmarks</i>			
Always predict zero	0.0597	0.0603	0.0507
Leave-one-paper-out mean	0.0537	0.0539	0.0463
Papers	33	30	20
Questions	673	629	339

Errors are averaged within paper, then equally across papers in each column. Benchmarks are calculated within each time period’s sample.

Figure 2 compares mean reported effects with mean forecasts across bins of forecast magnitude. Forecasts appear to be relatively well-calibrated. Appendix Figure A.4 provides more detail.

Appendix Figure A.5 plots the MSE by paper. The values are clearly driven by a few outlier papers. Appendix Figure A.6 assesses the sensitivity of these results to paper inclusion, and they seem relatively robust. Notably, since the MSE is so small for most papers, even if one were worried a couple of papers may have made it into the training data, this would not substantively influence the aggregate MSE; the MSE is more affected by a few projects with outlier results, without which it would be even more exact.

Re-running the forecasting prompt yields forecasts correlating 97% across identical requests on average (Table A.3). I also assess sensitivity to different prompts at the same effort settings, as described in Appendix B.1, but results do not change much. Finally, sampling variance could be inflating MSEs. Examining this issue with the May-onward Astra model, it appears that about 19% of MSE is accounted for by the mean sampling variance.⁹

⁹This test is restricted to the 245 results with directly reported standard errors.

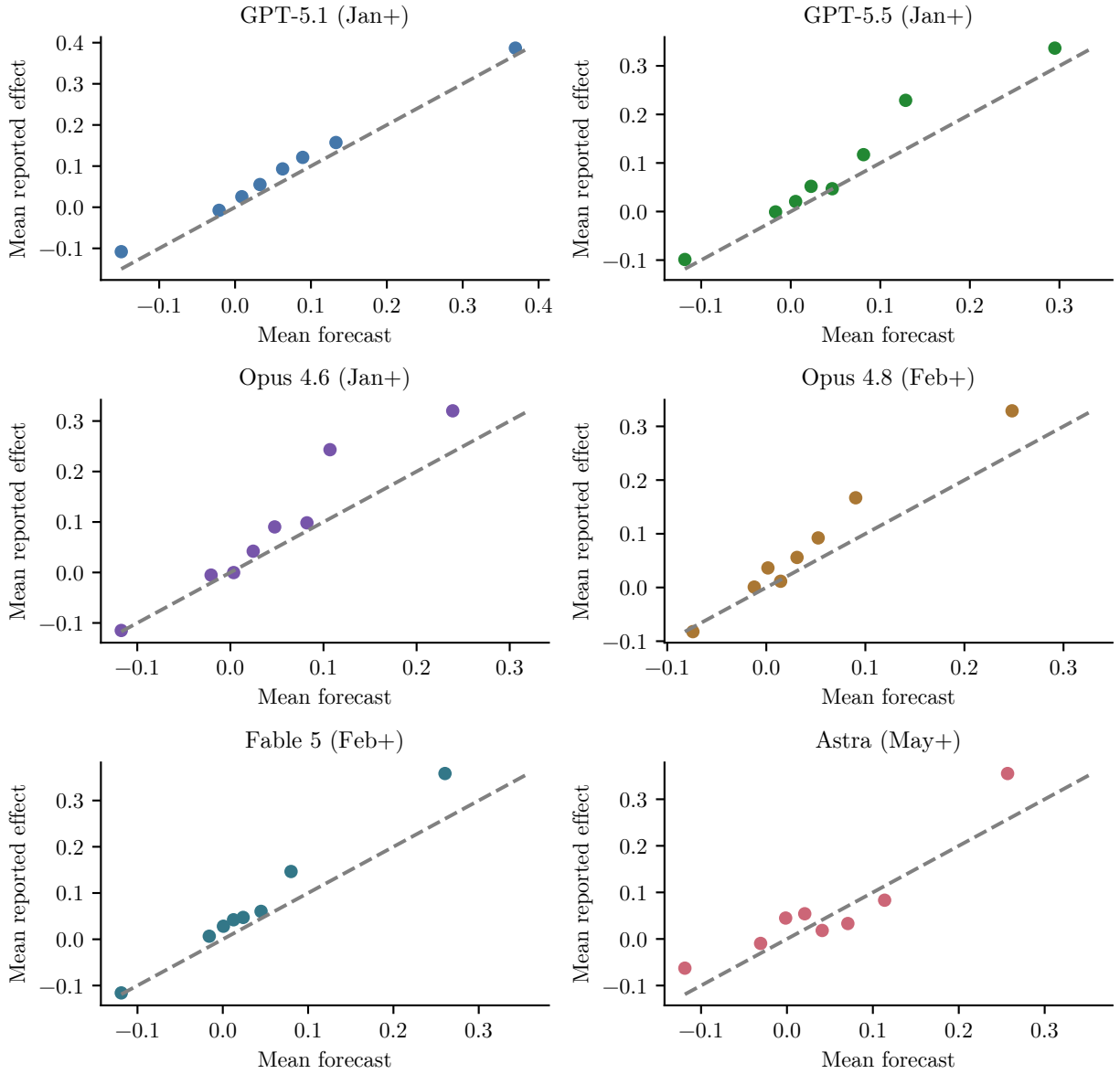


Figure 2: Binned calibration by model

This figure plots the mean forecast against the mean reported effect, both in standard deviations, for each model. Different models use different samples given their different cut-off dates.

3.3 Distributional results

Figure 3 evaluates the 10th–90th percentile intervals against the reported treatment effects. The treatment effects fall within this 80% interval 80.5–82.3% of the time for five models, and 72.7% for GPT-5.1. The second panel of Figure 3 suggests these forecasts of different quantiles could be meaningful in decision-making: the questions with the largest estimates at the 90th percentile are not the same as the questions with the largest estimates at the mean. In particular, questions’ rank according to their 90th percentile estimate has a Spearman correlation of 0.62–0.84 with their rank at the mean, and for a decision-maker trying to select the top 10%, judging by the 90th percentile rather than the mean would change 23.5–55.6% of selections.¹⁰

Figure 4 asks whether wider forecast intervals identify less predictable results. Wider intervals do appear to generally accompany larger absolute forecast errors. This implies that models can be used to gauge which results are less predictable, beyond achieving coverage on average.

¹⁰This abstracts from any differences in the costs of different interventions or value of different outcomes, considering all outcomes in SD units equally.

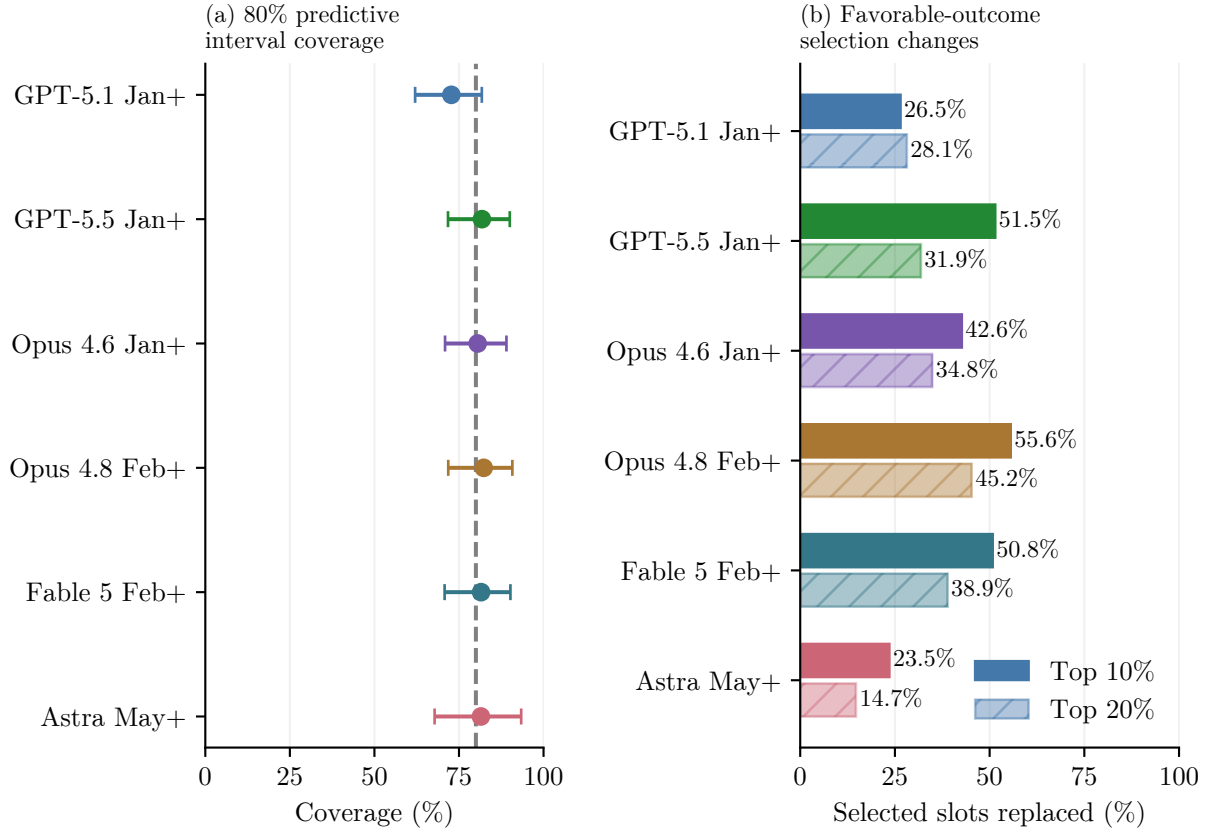


Figure 3: Predictive interval coverage and changes in top-decile and top-quintile selection

Panel (a) shows inclusive coverage of the 10th–90th percentile interval against reported treatment effects, with equal paper weights and 95% intervals from 2,000 bootstrap draws. The dashed line marks 80% coverage. Panel (b) ranks questions by their mean or 90th-percentile forecast in SD units and selects the top 10% or 20% under each ranking. Bars show the percentage of selected slots replaced. Jan+ includes 673 questions in 33 papers; Feb+ includes 629 in 30; May+ includes 339 in 20.

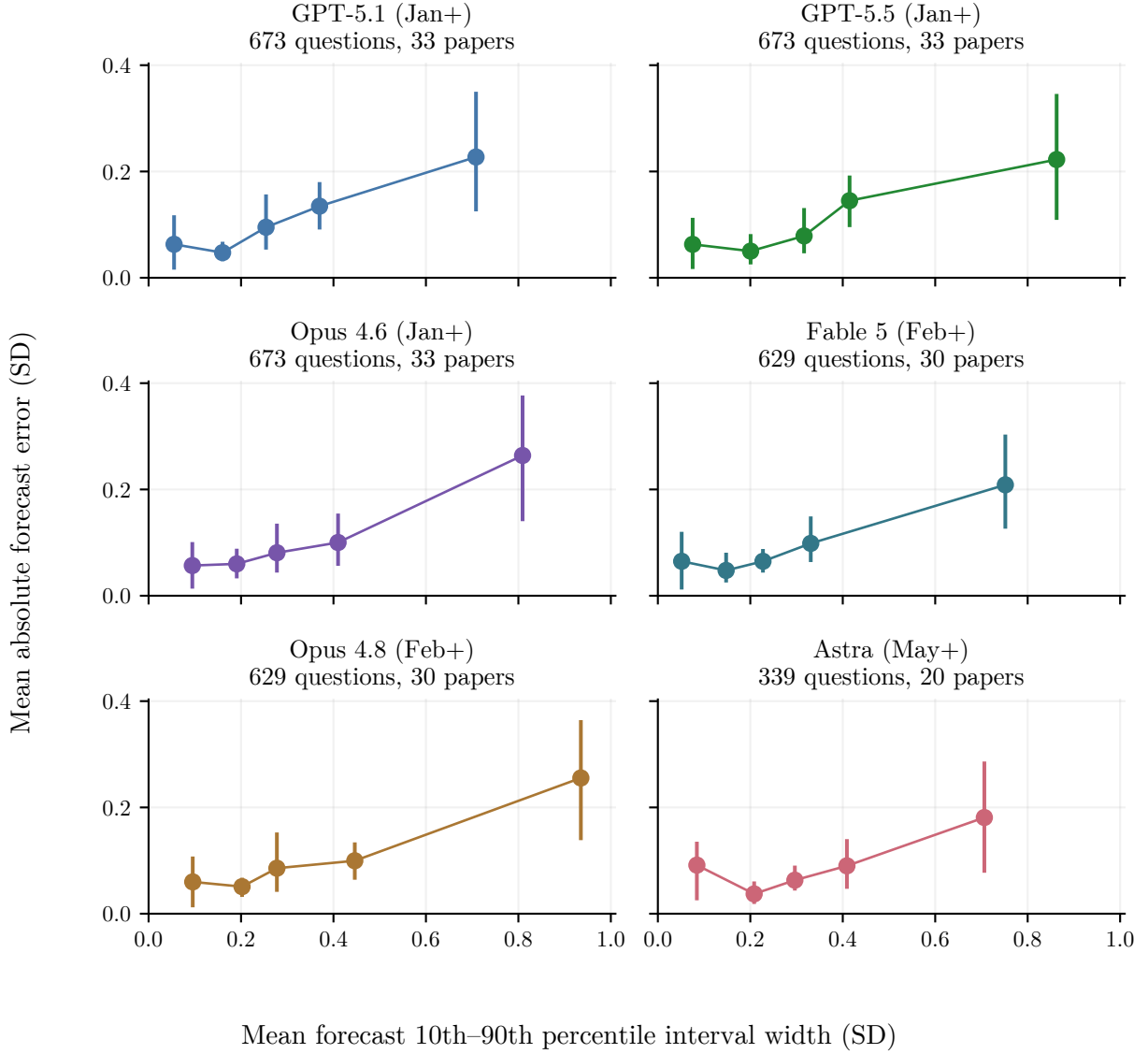


Figure 4: Do wider forecast intervals identify less predictable results?

Questions are binned into five groups of forecast interval width. Questions are weighted equally within a paper and papers are equally weighted before binning. Vertical bars represent 95% confidence intervals from 2,000 bootstrap draws. Errors are absolute differences between the mean forecast and reported effect, in standard deviations. Models' eligible windows differ; comparisons across panels are descriptive. Those with wider interval bins, indicating more uncertainty, appear to have larger forecast errors.

4 Discussion and Limitations

The results describe how accurately LLMs can predict treatment effects from registrations and pre-analysis plans and seem to suggest substantial scope for these forecasts to improve decision-making.

Several limitations remain. First, despite best efforts to ensure that models are not contaminated with results, the exercise would be more credible if done prospectively. Some work is in progress by the author and others to do exactly this. Second, the leave-one-paper-out benchmark uses a simple prediction rule based on other papers’ realized results. Future work could compare AI forecasts with alternative benchmarks. Third, the matching of results to forecasting questions could introduce noise. The findings also apply only to outcomes with reported results that can be standardized and matched sufficiently closely to the pre-registered questions.

Importantly, the sample is selected in ways that could affect results: the focus is on studies with well-documented pre-registrations or pre-analysis plans, but these may be precisely the easiest to forecast. Indeed, studies with longer descriptions (as measured by number of words) tend to be forecast better within both AEA RCT Registry source categories (Figure A.7).¹¹ Figure A.8 illustrates.

Rather than this being a drawback, however, it highlights that forecasts could do even better in the future if researchers wrote fuller descriptions of their studies. With LLMs being used more in research, writing more complete descriptions may become easier. LLMs might both write scientific protocols and assess their treatments’ likely effects.

Despite these limitations, it seems clear that forecasting can already serve as a useful tool - not just for forecasting states of the world like who wins an election or sports game, but also for forecasting causal treatment effects.

Forecasting treatment effects may at first glance seem like a step backwards: the credibility revolution occurred in large part to mitigate the risks of biases that can result in the absence of rigorous evidence. However, given we do not always have results from a RCT, and sometimes one cannot be done, high-quality forecasts can at least improve over the counterfactual of less well-informed and biased individual beliefs. Further, by repeatedly forecasting a variety of research

¹¹There are too few OSF papers to assess this relationship reliably.

results in different contexts, we can hope to learn more about the types of situations under which forecasts perform well. The evidence so far seems to suggest they are generally very accurate, and wider forecast intervals identify effects that are harder to predict. This suggests it may be worth focusing research efforts more on topics where there is more uncertainty in the AI forecasts, where there may be higher value of information to running a trial.

References

- Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. “Out of One, Many: Using Language Models to Simulate Human Samples.” *Political Analysis* 31(3): 337–351. doi:10.1017/pan.2023.2.
- Ashokkumar, Ashwini, Luke Hewitt, Isaias Ghezae, and Robb Willer. 2026. “Large Language Models Can Predict the Results of Social Science Experiments.” *Nature* 656: 115–122. doi:10.1038/s41586-026-10742-x.
- Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. “Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics.” *American Economic Review* 110(11): 3634–3660. doi:10.1257/aer.20190687.
- Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. 2016. “Evaluating Replicability of Laboratory Experiments in Economics.” *Science* 351(6280): 1433–1436. doi:10.1126/science.aaf0918.
- DellaVigna, Stefano, and Eva Vivalt. 2025. “Forecasting Social Science: Evidence from 100 Projects.” NBER Working Paper 34493, revised April 2026. doi:10.3386/w34493.
- Dreber, Anna, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson. 2015. “Using Prediction Markets to Estimate the Reproducibility of Scientific Research.” *Proceedings of the National Academy of Sciences* 112(50): 15343–15347. doi:10.1073/pnas.1516179112.
- Forecasting Research Institute. 2026. “AI Models Have Likely Reached Parity with Superforecasters on ForecastBench.” July 16. <https://forecastingresearch.substack.com/p/ai-models-have-likely-reached-parity>.
- Horton, John J. 2023. “Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?” arXiv:2301.07543, version 1. <https://arxiv.org/abs/2301.07543v1>.
- Vivalt, Eva. 2019. “Specification Searching and Significance Inflation Across Time, Methods and Disciplines.” *Oxford Bulletin of Economics and Statistics* 81(4): 797–816. doi:10.1111/obes.12289.
- Vivalt, Eva. 2020. “How Much Can We Generalize from Impact Evaluations?” *Journal of the European Economic Association* 18(6): 3045–3089. doi:10.1093/jeea/jvaa019. Replication files: doi:10.6084/m9.figshare.12048600.

A Supplementary tables and figures

Table A.1: Provider-reported model cutoffs and evaluation periods

Model	Reported cutoff	Provider terminology	Evaluation period
GPT-5.1	September 30, 2024	Knowledge cutoff	Jan+ 2026
GPT-5.5	December 1, 2025	Knowledge cutoff	Jan+ 2026
Opus 4.6	May 2025; August 2025	Reliable knowledge cutoff; training data cutoff, respectively	Jan+ 2026
Opus 4.8	January 2026	Reliable knowledge cutoff and training data cutoff	Feb+ 2026
Fable 5	January 2026	Reliable knowledge cutoff and training data cutoff	Feb+ 2026
GPT-6 Astra	April 30, 2026	Knowledge cutoff	May+ 2026

Reported cutoff and provider terminology come from Provider documentation for [GPT-5.1](#), [GPT-5.5](#), [Opus 4.6](#), [Opus 4.8](#), [Fable 5](#), and [GPT-6 Astra](#), accessed October 18, 2026. Evaluation period refers to the window over which each model was considered eligible in this paper.

Table A.2: Analysis sample by source

Source	Jan+		Feb+		May+	
	Papers	Questions	Papers	Questions	Papers	Questions
AEA public PAP	16	509	14	466	8	217
AEA registration, no public PAP	14	150	13	149	10	116
OSF	3	14	3	14	2	6
Total	33	673	30	629	20	339

This table summarizes the number of papers and questions available at each time period from each source.

Table A.3: Repeatability of forecasts under identical requests

Model	Period	Papers	Questions	Mean absolute change (SD)	Correlation	Original MSE	Repeat MSE
GPT-5.1	Jan+	33	673	0.0385	0.9319	0.0365	0.0404
GPT-5.5	Jan+	33	673	0.0186	0.9769	0.0346	0.0327
Opus 4.6	Jan+	33	673	0.0143	0.9775	0.0369	0.0361
Opus 4.8	Feb+	30	629	0.0178	0.9746	0.0337	0.0314
Fable 5	Feb+	30	629	0.0188	0.9624	0.0256	0.0249
Astra	May+	20	339	0.0094	0.9892	0.0218	0.0211

This table shows how similar repeat runs of the same forecasting prompts were. The original run is the one used in the main analysis.

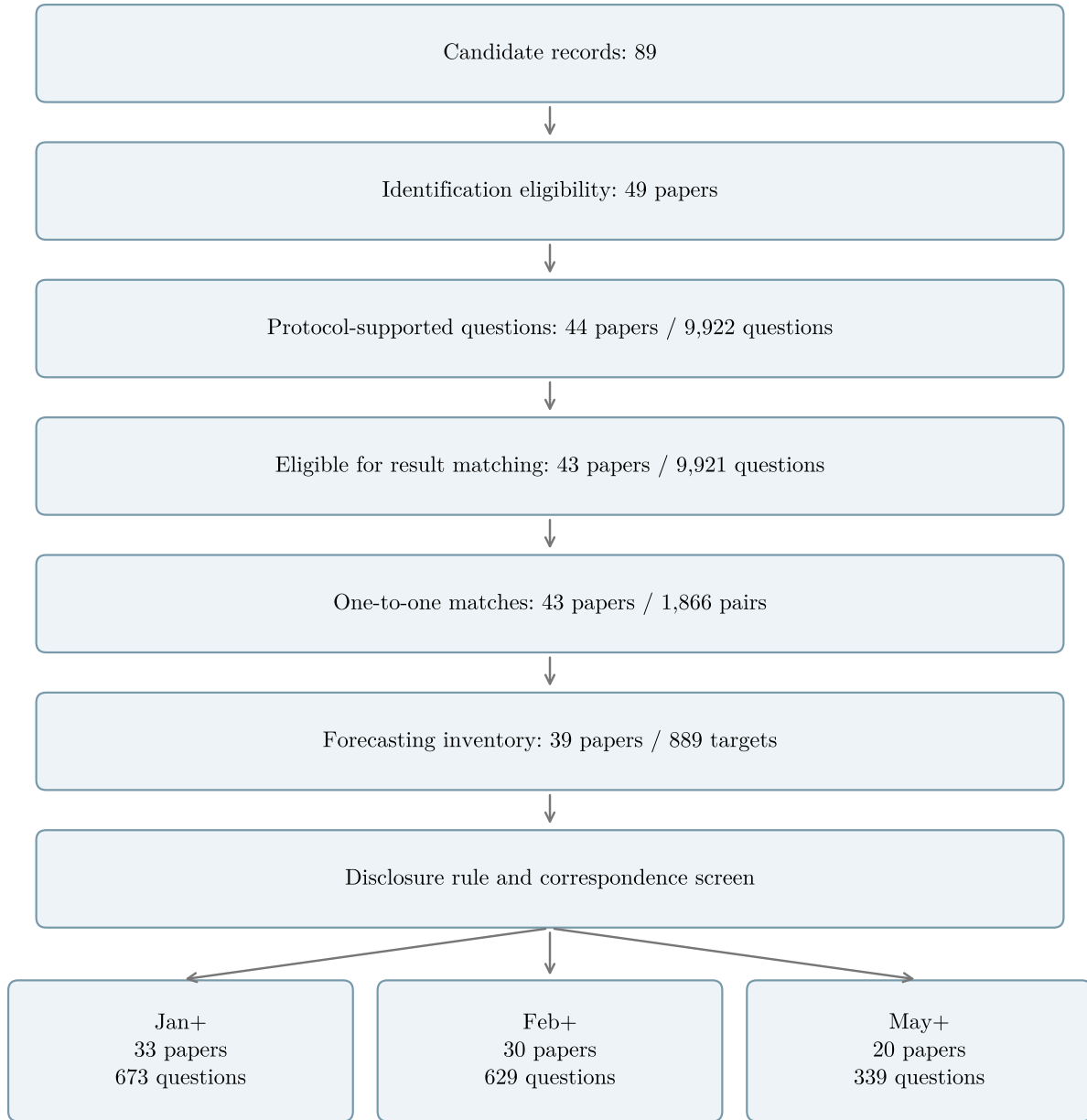


Figure A.1: Sample construction

This flowchart shows how papers and questions were screened at each stage.

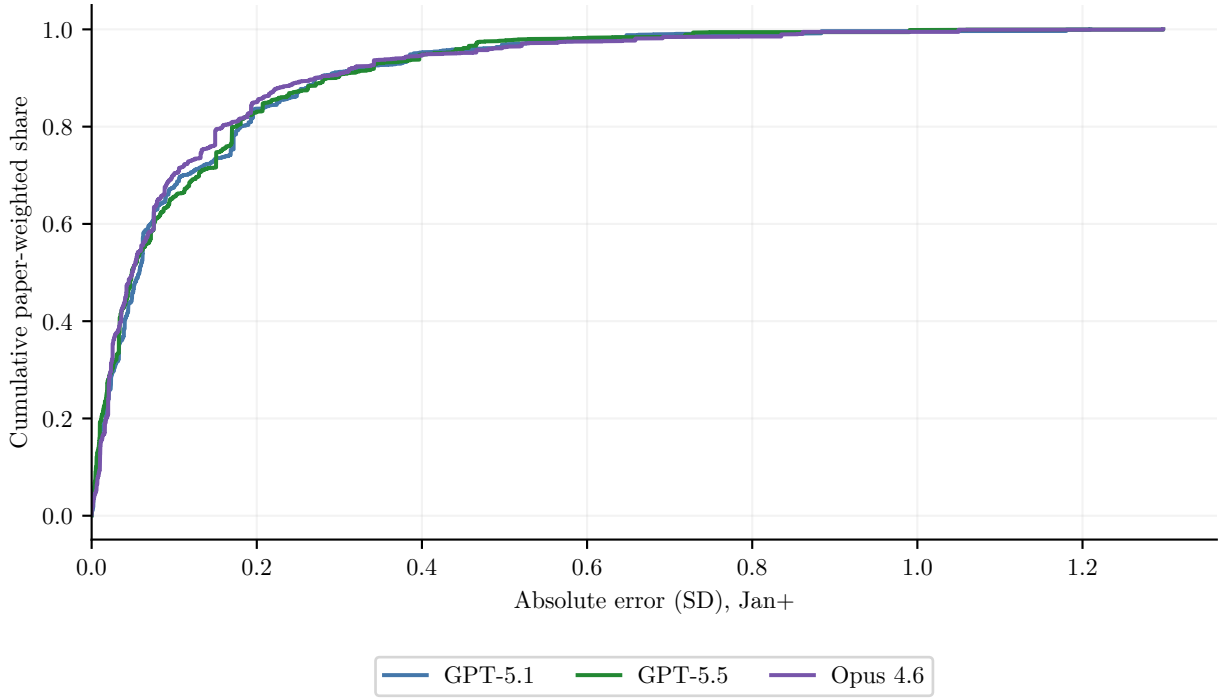


Figure A.2: Distribution of absolute forecast errors

This figure plots the distribution of forecast errors across models for the time period with the most papers / questions, comparing GPT-5.1, GPT-5.5 and Opus 4.6 on all 673 Jan+ questions in 33 papers. Each question is equally weighted within a paper and each paper is equally weighted.

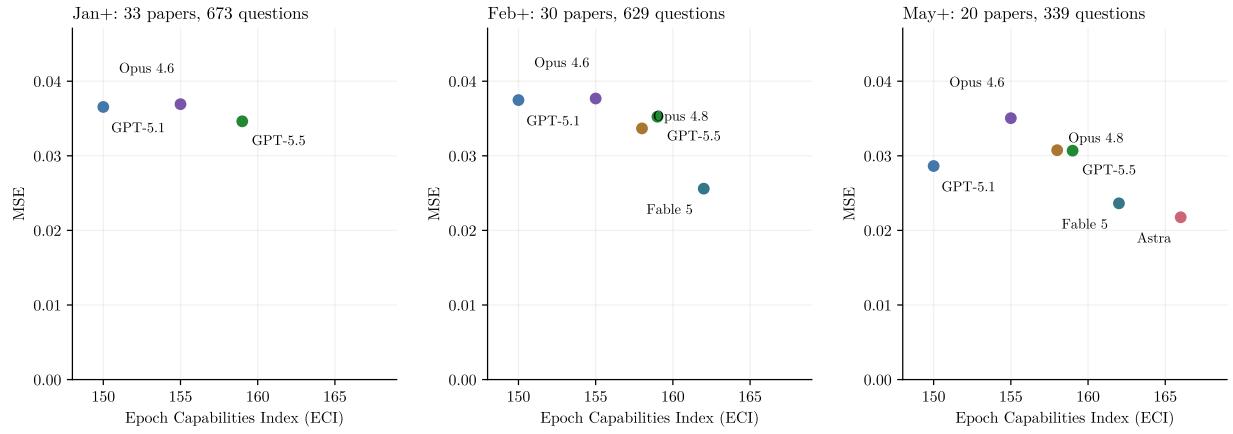


Figure A.3: Forecast MSE and general model capability

This figure shows the performance of different models against their ECI values. Within each panel, all models use identical questions; periods are nested.

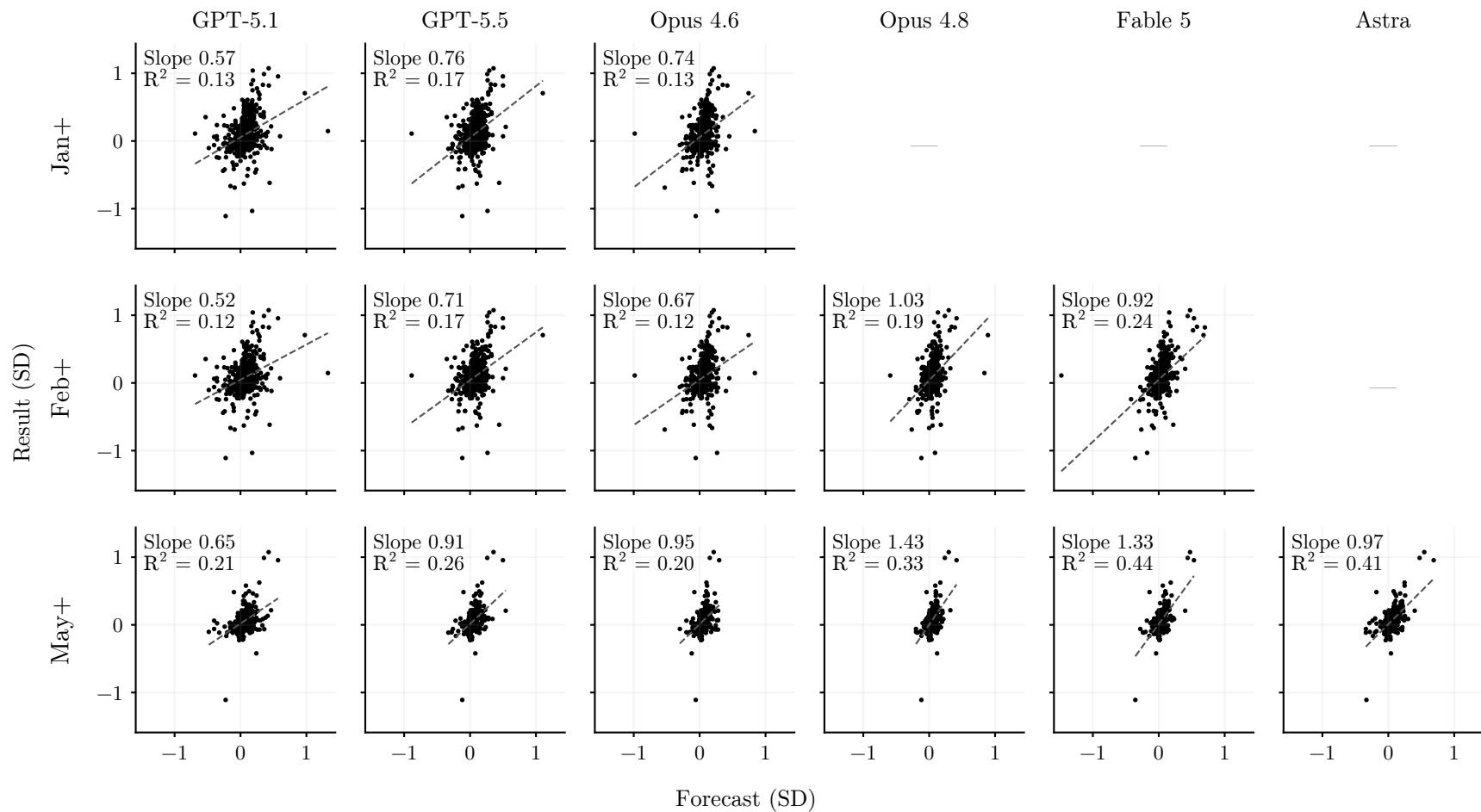


Figure A.4: Forecasts and reported treatment effects

This figure shows scatterplots and lines of best fit between forecasts and results (both in standard deviations). Rows show the Jan+ (673 questions, 33 papers), Feb+ (629 questions, 30 papers), and May+ (339 questions, 20 papers) samples, while columns show different models.



Figure A.5: Forecast accuracy by paper across the Jan+ sample

This figure shows accuracy by paper and model. The sample is partitioned into nonoverlapping disclosure periods. Right arrows indicate the 14 paper-model MSEs above 0.10.

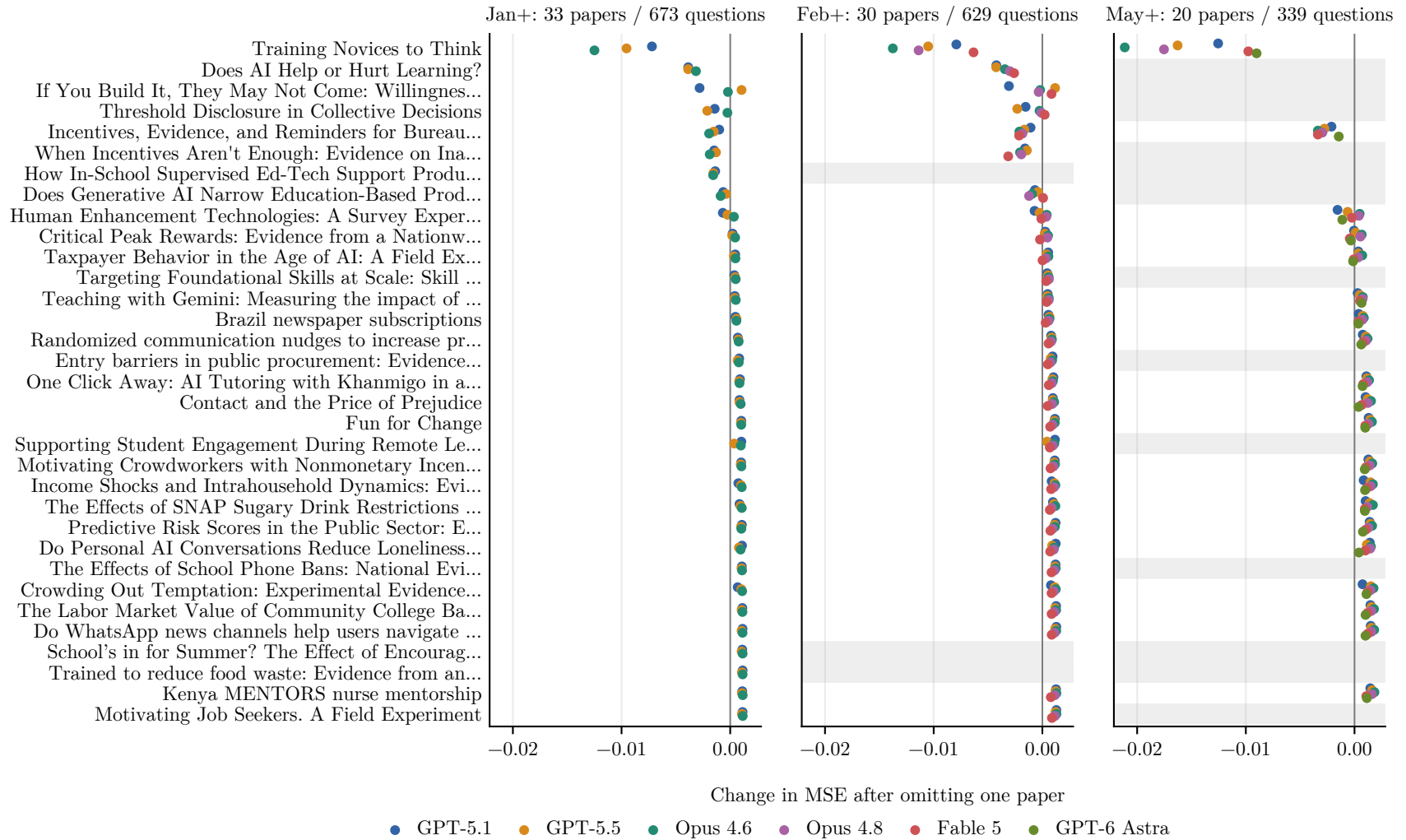


Figure A.6: Sensitivity to omitting one paper

Each point is the change in the MSE after omitting a given paper. Negative values indicate lower MSE after omission. Shaded rows identify papers outside a panel's disclosure period.

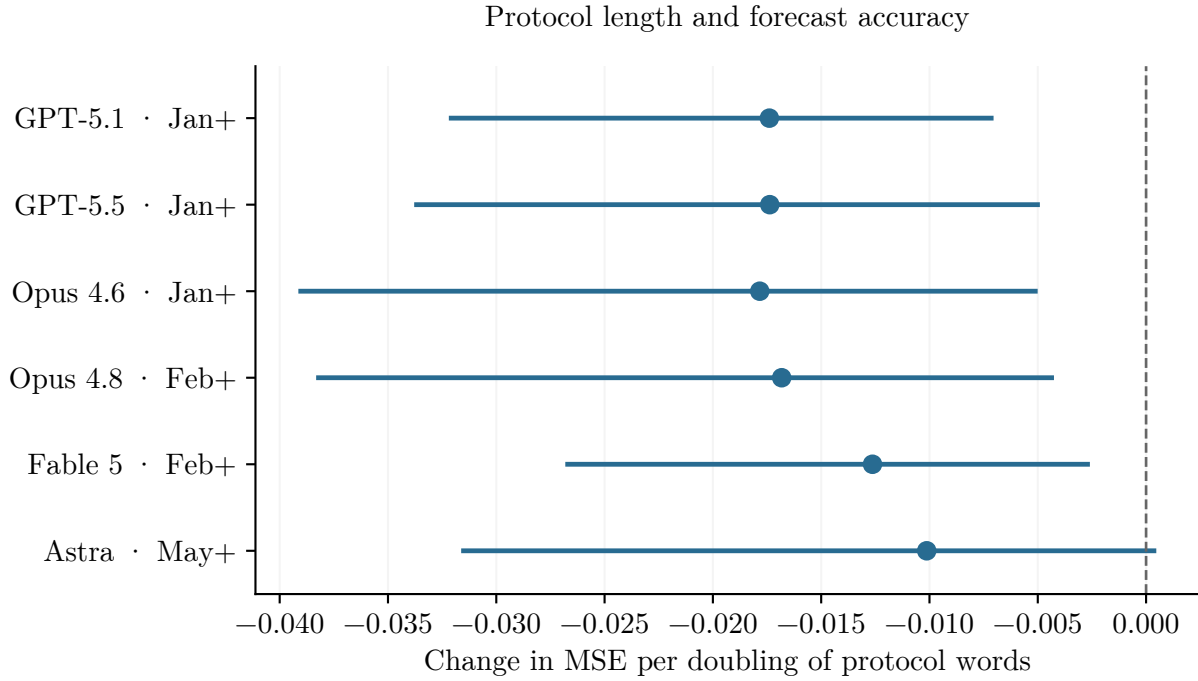


Figure A.7: Protocol length and forecast accuracy: regression coefficients

This figure shows the estimated change in paper-level MSE per doubling of protocol word count, controlling for source and giving papers equal weight. Lines show 95% percentile confidence intervals from 5,000 bootstrap resamples. Word counts use the protocol text supplied to the models, excluding instructions and forecasting questions.

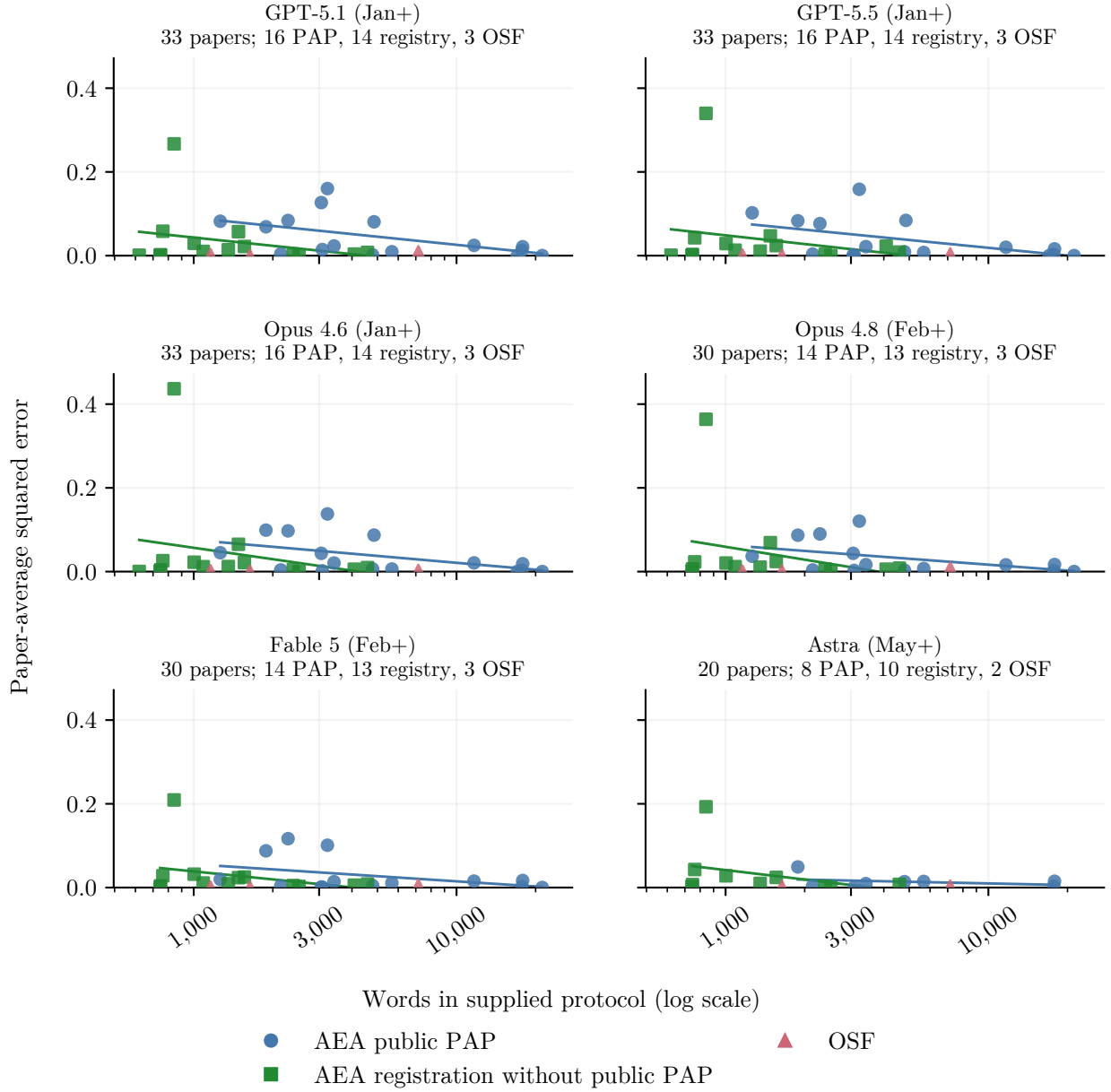


Figure A.8: Forecast accuracy and the length of supplied protocols

This figure shows detail in the relationship between protocol length and forecast accuracy. Each point is a paper. Solid lines fit paper-level MSE to log word count separately within each AEA source category, giving papers equal weight. No fitted line is shown for OSF because only two or three papers are available depending on the time period.

B Sensitivity analyses

B.1 Sensitivity to forecasting prompts

The experiments in this section use the same model-specific effort settings and output limits as the main analysis.

Figure B.1 compares five alternative prompts with the original prompt. These alternative prompts include three that make relatively small changes: prompting the LLM to act more as an independent expert, consider a reference class of results, and consider a balanced set of mechanisms. Two other alternative prompts introduce more structured reasoning: one again asking the LLM to consider relevant literature in the same reference class and one which asks the LLM to explicitly consider a mixture of scenarios. No prompt systematically improves accuracy across models.

B.2 Sensitivity to reasoning effort

Table B.1 compares the higher-effort forecasts used by default in the main analysis with the high-effort alternative. Moving to the alternative lowers MSE for GPT-5.5 (Jan+), GPT-5.5 (Feb+), GPT-5.5 (May+) and Fable 5 (May+), and raises it for the other model-sample comparisons. None of the paired differences is statistically significant at the 5% level.

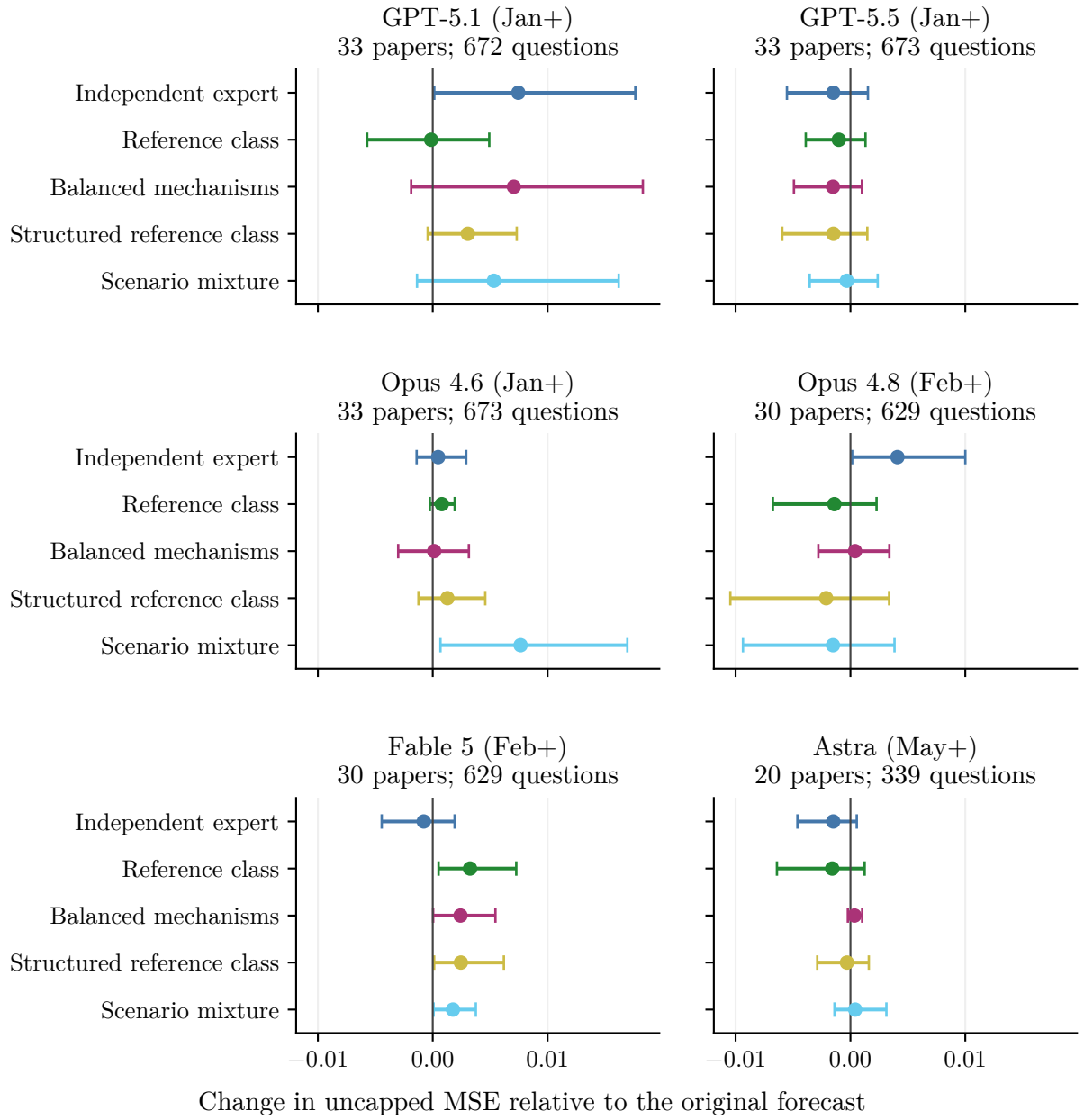


Figure B.1: Changes in forecast accuracy under alternative prompts.

Negative values indicate lower MSE than the original prompt. 95% percentile intervals from 10,000 paper bootstrap draws are shown.

Table B.1: Sensitivity to reasoning effort and output limits

Model	Sample	Main MSE	High-effort alternative MSE	Change
GPT-5.5	Jan+	0.03462	0.03285	-5.1%
Opus 4.6	Jan+	0.03692	0.03727	+0.9%
GPT-5.5	Feb+	0.03524	0.03313	-6.0%
Opus 4.6	Feb+	0.03768	0.03807	+1.0%
Opus 4.8	Feb+	0.03367	0.03644	+8.2%
Fable 5	Feb+	0.02560	0.02698	+5.4%
GPT-5.5	May+	0.03069	0.02745	-10.6%
Opus 4.6	May+	0.03505	0.03514	+0.3%
Opus 4.8	May+	0.03075	0.03312	+7.7%
Fable 5	May+	0.02363	0.02263	-4.2%
Astra	May+	0.02176	0.02238	+2.8%

The main specification uses xhigh for OpenAI and max for Anthropic, with a 64,000-token output cap, with the exception that GPT-5.1 is at high, its maximum available setting, and not shown here. The alternative uses high with a 32,000-token cap. Change is the percentage change in MSE relative to the main specification.

C Papers in the January-onward screened analysis sample

This list covers the 33 papers in the January-onward screened analysis sample. References identify the archived results papers; registry links identify their associated study records.

Acton, Riley K., Camila Morales, Julia A. Turner, Lois Miller, and Kalena Cortes. 2026. “The Labor Market Value of Community College Bachelor’s Degrees: Initial Evidence from a Resume Audit Study in Early Childhood Education.” NBER Working Paper 35404. [Paper](#). [AEARCTR-0016106](#).

Adena, Maja, Patrycja Janowska-Widomska, and Julian Harke. 2026. “Motivating Crowdworkers with Nonmonetary Incentives and Payment Framing—Evidence from a Large-Scale Experiment.” CRC TRR 190 Discussion Paper 576. [AEARCTR-0006744](#).

Agte, Patrick, Daniel Morales, Christopher Neilson, Sebastián Otero, and Gautam Rao. 2026. “Incentives, Evidence, and Reminders for Bureaucrats: Overcoming Barriers to Policy Scale Up.” Working paper. [AEARCTR-0003621](#).

Allcott, Hunt, E. Jason Baron, Thomas Dee, Angela L. Duckworth, Matthew Gentzkow, and Brian Jacob. 2026. “The Effects of School Phone Bans: National Evidence from Lockable Pouches.” NBER Working Paper 35132. [Paper](#). [OSF-8twh9](#).

Allcott, Hunt, Amy Finkelstein, Anna Grummon, and Matthew J. Notowidigdo. 2026. “The Effects of SNAP Sugary Drink Restrictions on Consumption and Welfare.” NBER Working Paper 35659. [Paper](#). [OSF-5nf3b](#).

Andersen, Andreas Makoto Fukuda, Gwen-Jirō Clochard, Guillaume Hollard, and Andreas Kot-sadam. 2026. “Contact and the Price of Prejudice.” University of Osaka, ISER Discussion Paper 1317. [AEARCTR-0018089](#).

Asirvatham, Hemanth, Chiara Betti, Rachel Brown, Arnaldo Camuffo, Aaron Chatterji, Chiara Fumagalli, Alfonso Gambardella, Myriam Mariani, Abhinav Pandey, Steve Ramos, Giovanni Salvucci, and Vladimir Simic. 2026. “Training Novices to Think, or Giving Them LLMs? Evidence from an RCT.” Working paper. [AEARCTR-0017124](#).

Baron, Jason, Arkadev Ghosh, and Richard Lombardo. 2026. “Predictive Risk Scores in the Public Sector: Experimental Evidence from Child-Protection Investigations.” Working paper. [AEARCTR-0012623](#).

Bernard, Louise, Robert Metcalfe, and Andrew R. Schein. 2026. “Critical Peak Rewards: Evidence from a Nationwide Demand Response Program.” Working paper. [AEARCTR-0013034](#).

Bhatt, Monica P., Jonathan Guryan, Fatemeh Momeni, Philip Oreopoulos, and Eleni Packis. 2026. “Supporting Student Engagement During Remote Learning: Three Randomized Controlled

- Trials in Chicago Public Schools.” NBER Working Paper 34996. [Paper](#). [AEARCTR-0006250](#).
- Braghieri, Luca, Leonardo Bursztyn, and Jan Fasnacht. 2026. “Threshold Disclosure in Collective Decisions.” Working paper. [AEARCTR-0016968](#).
- Burlig, Fiona, James Bushnell, and David Rapson. 2026. “If You Build It, They May Not Come: Willingness to Participate in Managed EV Charging.” Becker Friedman Institute Working Paper 2026-56. [AEARCTR-0014013](#).
- Carrell, Scott E., Michal Kurlaender, Paco Martorell, and Christina Sun. 2026. “School’s in for Summer? The Effect of Encouraging Summer Community College Enrollment.” AEA Papers and Proceedings 116: 362–366. [Paper](#). [AEARCTR-0011439](#).
- Castro, Caio and João F. Pugliese. 2026. “Good News and Bad News: The Effects of Consuming Professional News Outlets During the 2022 Brazilian Elections.” Working paper. [AEARCTR-0009940](#).
- Cockx, Bart, Johan Egebark, Greet Van Hove, Emilie Videnord, and Johan Vikström. 2026. “Motivating Job Seekers. A Field Experiment.” IFAU Working Paper 2026:8. [AEARCTR-0005502](#).
- Cruces, Guillermo, Diego Fernández Mejjide, Sebastian Galiani, Ramiro H. Gálvez, and María Lombardi. 2026. “Does Generative AI Narrow Education-Based Productivity Gaps? Evidence from a Randomized Experiment.” Universidad Torcuato Di Tella, Escuela de Gobierno, Documento de Trabajo 2026/03. [AEARCTR-0016607](#).
- de Barros, Andreas and Theresa Lubozha. 2026. “Targeting Foundational Skills at Scale: Skill Specificity and Transfer.” CESifo Working Paper 12542. [AEARCTR-0014922](#).
- Del Pozo Loayza, Cesar E., José D. López-Rivas, and Francisco Alpízar. 2026. “Trained to reduce food waste: Evidence from an RCT with micro-sized and small enterprises in Cusco’s hospitality industry.” Journal of Environmental Economics and Management 138: 103355. [AEARCTR-0012063](#).
- Fang, Ximeng, Stefania Innocenti, and Sonja Vogt. 2026. “Fun and Change: Video Game Entertainment Promotes Pro-Environmental Behaviour.” INET Oxford Working Paper 2026-13. [AEARCTR-0015162](#).
- Franco, Catalina, Natalie Irmert, and Siri Isaksson. 2026. “Does AI Help or Hurt Learning?” Working Paper 2026:2. [AEARCTR-0014930](#).
- Fréget, Louis, Ariell Reshef, and Claudia Senik. 2026. “Do Personal AI Conversations Reduce Loneliness? Evidence from a Pre-Registered Experiment.” CESifo Working Paper 12882. [AEARCTR-0016749](#).

- Gordon, Sarah H., Kevin H. Nguyen, Breanne Biondi-Bonetti, Jacob Wallace, Benjamin D. Sommers, and Anna Goldman. 2026. “Randomized communication nudges to increase primary care engagement among Medicaid enrollees in Maine.” *Health Affairs Scholar* 4(6): qxag100. [Paper](#). [AEARCTR-0012341](#).
- Holz, Justin E., Ricardo Perez-Truglia, Andrew Simon, and Alejandro Zentner. 2026. “Taxpayer Behavior in the Age of AI: A Field Experiment on Property Tax Appeals.” NBER Working Paper 35632. [Paper](#). [AEARCTR-0018430](#).
- Hyytinen, Ari, Jan Jääskeläinen, Antti Sieppi, Vesa-Heikki Soini, and Janne Tukiainen. 2026. “Entry barriers in public procurement: Evidence from conjoint survey experiment.” Aboe Centre for Economics Discussion Paper 176. [AEARCTR-0012628](#).
- Immordino, Giovanni, Mario Macis, Immacolata Marino, and Fabrizio Panebianco. 2026. “Human Enhancement Technologies: A Survey Experiment on Private Demand and Governance Preferences.” IZA Discussion Paper 18804. [AEARCTR-0016972](#).
- LearnLM Team, Google and Fab AI. 2026. “Teaching with Gemini: Measuring the impact of Guided Learning on student mathematics progress in Sierra Leone.” Research report, Google and Fab AI. [AEARCTR-0016651](#).
- Mont’Alverne, Camila, Patrícia Rossini, and Antonis Kalogeropoulos. 2026. “Do WhatsApp news channels help users navigate the information environment? A field experiment on information-related attitudes.” *New Media & Society*, 1–21. [Paper](#). [OSF-fv945](#).
- Oreopoulos, Philip and Nina Low. 2026. “One Click Away: AI Tutoring with Khanmigo in a Two-Year School Experiment.” EdWorkingPaper 26-1551. [AEARCTR-0013519](#).
- Oreopoulos, Philip, Oliver Keyes-Krysakowski, and Deepak Agarwal. 2026. “How In-School Supervised Ed-Tech Support Produces Massive Learning Gains: A Khan Academy Field Experiment in India.” NBER Working Paper 34683. [Paper](#). [AEARCTR-0013921](#).
- Rhodes, Elizabeth, David E. Broockman, Eva Vivalt, Patrick K. Krause, Sarah Miller, and Alexander W. Bartik. 2026. “Income Shocks and Intrahousehold Dynamics: Evidence from a Guaranteed Income Experiment.” NBER Working Paper 35589. [Paper](#). [AEARCTR-0006750](#).
- Thirumurthy, Harsha, Elizabeth F. Bair, Dean Yang, Teniola I. Egbe, Benard Ayieko, Julius Wesonga, Sue Napierala, Connie Celum, Mary E. Putt, Carol S. Camlin, and Mark Ayallo. 2026. “Crowding Out Temptation: Experimental Evidence on Savings and Health Behaviors in Kenya.” NBER Working Paper 35833. [Paper](#). [AEARCTR-0009484](#).
- Yu, Hang, Jared Stolove, Dean Yang, James Riddell IV, and Arlete Mahumane. 2026. “When Incentives Aren’t Enough: Evidence on Inattention and Imperfect Memory from HIV Medication Adherence.” NBER Working Paper 34917. [Paper](#). [AEARCTR-0003184](#).

Zhang, Lujia, Rajet Vatsa, Wei Chang, Sharon Akinyi, Sarah Little, Catherine Gakii, John Mungai, Cynthia Kahumbura, Sathyanath Rajasekharan, Anneka Wickramanayake, Margaret McConnell, and Jessica Cohen. 2026. “Impact of a Long-Term Mentorship Program for Child-birth Providers on Quality of Care: A Cluster Randomized Controlled Trial.” PLOS Global Public Health 6(9): e0006073. [Paper](#). [AEARCTR-0008449](#).

D Historical benchmark comparison

Results are shown in Table D.1. Because the treatment effects in these data differ from those in the 2026 set I consider in this paper, they are not immediately comparable (Appendix Figure D.1). To make them more comparable, I rescale them so that their root-mean-square magnitude matches that of the Jan+, Feb+, and May+ groups. Samples still differ, but the leave-one-out meta-analyses based on Vivalt (2020) do not seem to predict treatment effects as well as LLMs. A researcher could in principle do better than a meta-analysis if they built a model of treatment effects for a particular intervention-outcome combination. For example, one might assume:

$$\begin{aligned} Y_i &= \theta_i + u_i, & u_i &\sim \mathcal{N}(0, \sigma_i^2), \\ \theta_i &= \alpha + X_i\beta + e_i, & e_i &\sim \mathcal{N}(0, \tau^2), \end{aligned}$$

where Y_i is the estimated treatment effect in study i , θ_i is its true treatment effect, X_i contains study characteristics, α is an intercept, and β contains the corresponding coefficients. The sampling error u_i has variance σ_i^2 . e_i is the residual study effect, with variance τ^2 measuring residual heterogeneity remaining after accounting for X_i .

Table D.1 shows that closing the gap between the rescaled historical forecasts and the LLM forecasts would require reductions in τ^2 of at least roughly 30–60%, assuming prediction error consists of a component proportional to τ^2 plus sampling noise. For context, in Vivalt (2020), the single best-fitting covariate reduced estimated τ^2 by about 10% at the median and 20% on average in the intervention-outcome combinations examined.

The main sample in Table D.1 used intervention-outcome combinations with at least three studies, but these may lead to noisy estimates. Table D.2 restricts the historical benchmark to intervention–outcome combinations with at least five studies. These appear to do slightly better, but not by much.

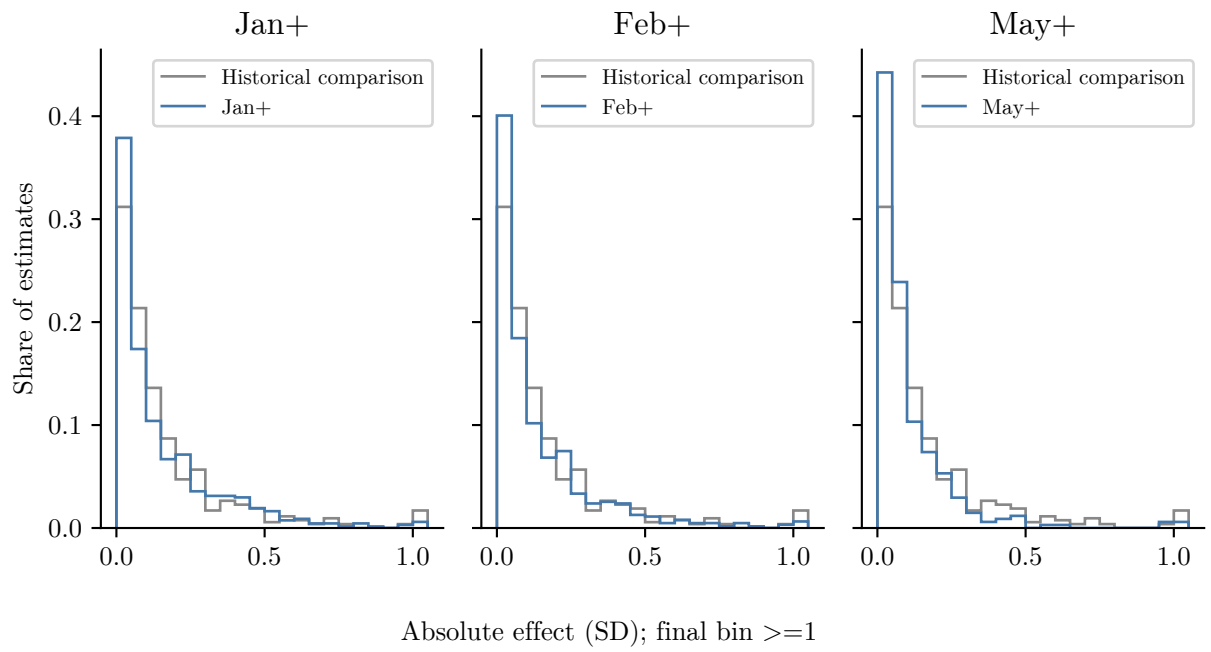


Figure D.1: Magnitudes of reported effects

This figure compares the set of papers available for AI forecasting in each time period with the historical results in [Vivalt \(2020\)](#).

Table D.1: AI MSE vs. MSE from historical benchmark

Model	AI MSE	Original MSE	p -value	Minimum τ^2 reduction required	Rescaled MSE	p -value	Minimum τ^2 reduction required
Panel A. Jan+ (33 papers, 673 questions)							
GPT-5.1	0.0365	0.1024	0.061	64%	0.0567	0.118	36%
GPT-5.5	0.0346	0.1024	0.058	66%	0.0567	0.032	39%
Opus 4.6	0.0369	0.1024	0.075	64%	0.0567	0.036	35%
Panel B. Feb+ (30 papers, 629 questions)							
GPT-5.1	0.0375	0.1024	0.066	63%	0.0573	0.160	35%
GPT-5.5	0.0352	0.1024	0.063	66%	0.0573	0.048	38%
Opus 4.6	0.0377	0.1024	0.082	63%	0.0573	0.054	34%
Opus 4.8	0.0337	0.1024	0.058	67%	0.0573	0.037	41%
Fable 5	0.0256	0.1024	0.030	75%	0.0573	0.034	55%
Panel C. May+ (20 papers, 339 questions)							
GPT-5.1	0.0286	0.1024	0.045	72%	0.0482	0.185	41%
GPT-5.5	0.0307	0.1024	0.061	70%	0.0482	0.100	36%
Astra	0.0218	0.1024	0.025	79%	0.0482	0.114	55%
Opus 4.6	0.0350	0.1024	0.098	66%	0.0482	0.059	27%
Opus 4.8	0.0308	0.1024	0.066	70%	0.0482	0.069	36%
Fable 5	0.0236	0.1024	0.030	77%	0.0482	0.090	51%

This table compares the MSE of the AI forecasts using the AEA RCT Registry and OSF with the MSE one would get by using leave-one-out meta-analysis to predict the left out study across the 47 intervention-outcome combinations with at least 3 papers in [Vivalt \(2020\)](#). These samples are different, but perhaps still informative. The “Original MSE” column builds off the original estimates while the “Rescaled MSE” column multiplies the historical effects and leave-one-out predictions by the ratio of the AI sample’s root-mean-square reported effect to its historical counterpart, separately for each panel, to partially adjust for the different samples. The p -values test equality of the AI MSE compared to the MSE implied by the data in [Vivalt \(2020\)](#). Bootstrap tests use 50,000 draws independently resampling AI papers and historical groups. Rescaling is recomputed in every draw. The minimum required reduction in τ^2 is $\max\{0, 1 - \text{MSE}_{\text{AI}}/\text{MSE}_{\text{H}}\}$, using the original or rescaled historical MSE. Comparisons remain descriptive.

Table D.2: AI MSE vs. MSE from historical benchmark: intervention-outcomes with at least 5 papers

Model	AI MSE	Original MSE	p -value	Minimum τ^2 reduction required	Rescaled MSE	p -value	Minimum τ^2 reduction required
Panel A. Jan+ (33 papers, 673 questions)							
GPT-5.1	0.0365	0.0811	0.117	55%	0.0530	0.168	31%
GPT-5.5	0.0346	0.0811	0.108	57%	0.0530	0.052	35%
Opus 4.6	0.0369	0.0811	0.142	54%	0.0530	0.068	30%
Panel B. Feb+ (30 papers, 629 questions)							
GPT-5.1	0.0375	0.0811	0.130	54%	0.0536	0.220	30%
GPT-5.5	0.0352	0.0811	0.120	57%	0.0536	0.074	34%
Opus 4.6	0.0377	0.0811	0.157	54%	0.0536	0.097	30%
Opus 4.8	0.0337	0.0811	0.109	58%	0.0536	0.057	37%
Fable 5	0.0256	0.0811	0.047	68%	0.0536	0.044	52%
Panel C. May+ (20 papers, 339 questions)							
GPT-5.1	0.0286	0.0811	0.078	65%	0.0450	0.231	36%
GPT-5.5	0.0307	0.0811	0.110	62%	0.0450	0.130	32%
Astra	0.0218	0.0811	0.038	73%	0.0450	0.135	52%
Opus 4.6	0.0350	0.0811	0.174	57%	0.0450	0.081	22%
Opus 4.8	0.0308	0.0811	0.117	62%	0.0450	0.079	32%
Fable 5	0.0236	0.0811	0.047	71%	0.0450	0.108	48%

This table compares the MSE of the AI forecasts using the AEA RCT Registry and OSF with the MSE one would get by using leave-one-out meta-analysis to predict the left out study across the 34 intervention-outcome combinations with at least 5 papers in [Vivalt \(2020\)](#). These samples are different, but perhaps still informative. The “Original MSE” column builds off the original estimates while the “Rescaled MSE” column multiplies the historical effects and leave-one-out predictions by the ratio of the AI sample’s root-mean-square reported effect to its historical counterpart, separately for each panel, to partially adjust for the different samples. The p -values test equality of the AI MSE compared to the MSE implied by the data in [Vivalt \(2020\)](#). Bootstrap tests use 50,000 draws independently resampling AI papers and historical groups. Rescaling is recomputed in every draw. The minimum required reduction in τ^2 is $\max\{0, 1 - \text{MSE}_{\text{AI}}/\text{MSE}_{\text{H}}\}$, using the original or rescaled historical MSE. Comparisons remain descriptive.