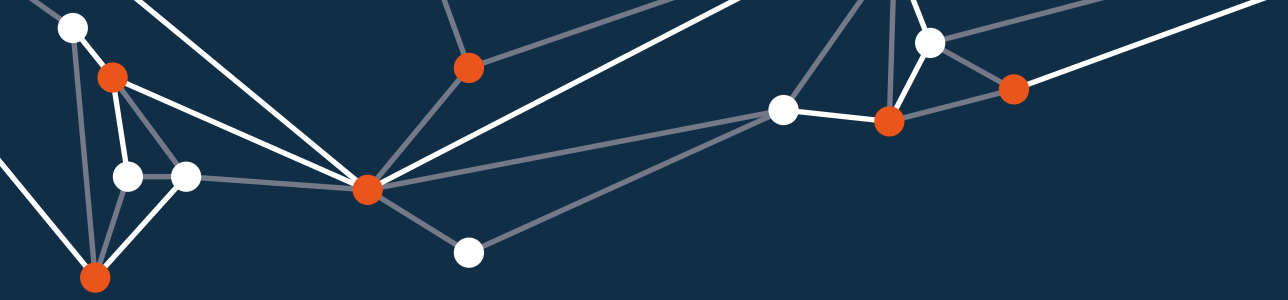


Forecasting Research Institute

Assessing Near-Term Accuracy in the Existential Risk Persuasion Tournament

Authors: Simas Kučinskas, Josh Rosenberg, Rebecca Ceppas de Castro, Zach Jacobs, Jordan Canedy, Philip E. Tetlock, Ezra Karger

First released on: September 2, 2025

Assessing Near-Term Accuracy in the Existential Risk Persuasion Tournament

Authors: Simas Kučinskas[1*], Josh Rosenberg[1], Rebecca Ceppas de Castro[1], Zach Jacobs[1], Jordan Canedy[1], Philip E. Tetlock[1,2], Ezra Karger[1,3]

Abstract

In June–October 2022, we convened 169 people to participate in the "Existential Risk Persuasion Tournament" (XPT). The XPT participants included both superforecasters with proven forecasting track records and domain experts with subject-matter expertise. The tournament incentivized accurate forecasting and persuasive argumentation about long-term risks humanity may face, including risks from artificial intelligence (AI), climate change, nuclear war, and pandemics. This report analyzes respondents' forecasting accuracy on 38 near-term questions that resolved by mid-2025. Key findings include: (a) there was overall performance parity between superforecasters and domain experts, with both groups underestimating AI progress and overestimating improvements in climate technology; (b) both superforecasters and domain experts substantially outperformed a baseline of educated members of the general public; (c) at the individual level, the median superforecaster and median domain expert performed statistically indistinguishably from simple extrapolation algorithms; (d) at the aggregate level, superforecasters and domain experts showed improved accuracy and some evidence of outperforming simple extrapolation algorithms; (e) there was no statistically significant correlation between near-term accuracy and long-term existential risk forecasts.

This research would not have been possible without the support of the Musk Foundation, Open Philanthropy, and the Long-Term Future Fund. We greatly appreciate the assistance and input of Sam Glover, Rory Svarc, and Bridget Williams throughout the project. The views expressed in this paper do not necessarily reflect those of the Federal Reserve Bank of Chicago or the Federal Reserve System.

1 = Forecasting Research Institute, 2 = Wharton School of the University of Pennsylvania, 3 = Federal Reserve Bank of Chicago.
*Corresponding author: simas@forecastingresearch.org

Executive Summary

This report evaluates the **accuracy of near-term forecasts** made by domain experts and superforecasters in the Existential Risk Persuasion Tournament (**XPT**; Karger et al., 2023).

Background

The XPT tournament took place in June–October 2022. The tournament convened 169 participants to generate probabilistic forecasts about humanity's long-term future and potential global risks such as climate change, nuclear war, pandemics, and artificial intelligence (AI). Of these participants, 89 were superforecasters with track records of high accuracy on near-term questions, while 80 were domain experts. In addition, we sampled hundreds of public participants for comparison. The XPT represents the largest existential risk forecasting tournament to date, uniquely combining superforecasters and domain experts to predict humanity's long-term risks.

The tournament included 59 forecasting questions set to resolve at dates ranging from mid-2024 to as late as 2100. These questions broke down into 172 subquestions over multiple forecasting horizons and, in some cases, across different countries. Out of these 172 subquestions, 38 have known outcomes (i.e., are "resolved") as of mid-2025. We note that the XPT tournament concluded prior to the public release of ChatGPT in November 2022.

Key Findings on Accuracy

Performance parity between superforecasters and domain experts. The near-term questions revealed no meaningful accuracy differences between superforecasters and experts forecasting on questions within their domain of expertise. Both groups achieved nearly identical accuracy scores. The performance gap between the most- and least-accurate XPT participant groups spanned just 0.18 standard deviations, comparable to the difference between median and slightly above-median performance. These small differences were not statistically significant, indicating that neither a proven forecasting track record nor domain expertise provided a consistent edge for these near-term predictions.

Individual forecasters outperformed public participants but not simple algorithms. Both superforecasters and domain experts strongly outperformed a sample of educated public participants, who scored 1.82 standard deviations below the median XPT participant. However, individual forecasters' performance was not statistically distinguishable from two simple algorithms: a "no-change" forecast and trend extrapolation. These simple algorithms performed well partly because many questions involved low-probability events (which did not occur) or slow-moving variables (where trends persisted).

Aggregate forecasts demonstrated the wisdom of crowds. Median aggregation of XPT participants' forecasts achieved a substantial improvement over individual performance, increasing accuracy by roughly 1 standard deviation. These aggregated predictions showed

weak but positive evidence of outperforming the “no-change” forecast, though not trend extrapolation. This finding reinforces the well-established principle that combining multiple forecasts improves accuracy.

Main Insights across Subject Areas

Despite the strong overall performance of aggregate forecasts, XPT participants systematically misjudged progress in specific domains.

Respondents underestimated AI progress, especially superforecasters. XPT participants significantly underestimated the pace of AI advancement across multiple benchmarks. For three standard AI benchmarks—MATH, MMLU, and QuALITY—domain experts assigned probabilities of 21.4%, 25.0%, and 43.5% respectively to the outcomes achieved by the end of 2024. Superforecasters were even more pessimistic, assigning only 9.3%, 7.2%, and 20.1% respectively. The International Mathematical Olympiad results proved particularly surprising: AI systems achieved gold-level performance in July 2025, an outcome to which domain experts assigned only an 8.6% probability and superforecasters a mere 2.3% probability. Overall, superforecasters assigned an average probability of just 9.7% to the observed outcomes across these four AI benchmarks, compared to 24.6% from domain experts.

Climate technology progress was overestimated. In contrast to AI, forecasters were overly optimistic about the development of green technology. In 2024, the cost of hydrogen produced using renewable electricity remained higher than anticipated at $7.50 USD/kg (median forecasts of $4.50 by superforecasters and $3.50 USD/kg by domain experts), and direct air CO_2 capture technology captured only 0.01 $MtCO_2$/year (median forecasts of 0.32 by superforecasters and 0.60 $MtCO_2$/year by domain experts).

Implications for Long-Term Risks

No correlation between near-term accuracy and long-term existential risk forecasts. There was no statistically significant correlation between forecasters’ near-term accuracy and their forecasts of long-term risks. Ideally, we would use near-term forecasting ability to assess the reliability of forecasts about humanity’s long-term future. Unfortunately, in our XPT data, near-term forecasting accuracy did not consistently align with any particular position on long-term risks. Overall, near-term forecasting accuracy provides limited evidence at present for identifying who makes the most credible long-term risk forecasts.

Next Steps

Given the faster-than-expected progress on AI capabilities, it is more important than ever to understand the likely future trajectory and impact of AI. In our current and future work, we aim to shed more light on these questions. Our current projects on this front include a longitudinal panel of AI experts and a survey of economists on the expected economic impacts of AI. Through these systematic efforts to gather expert perspectives, we will provide empirically grounded insights that can inform policy and decision-making.

1. Introduction

The Existential Risk Persuasion Tournament (XPT; Karger et al., 2023) convened 169 participants from June to October 2022 to forecast questions about humanity's long-term future and the impact of global risks such as climate change, nuclear war, pandemics, and artificial intelligence (AI). Of these 169 participants, 89 were experienced forecasters with a track record of high accuracy on near-term questions ("superforecasters"), and the other 80 were specialists working in domains related to global risks and humanity's future ("experts"). Additionally, hundreds of public participants provided their answers to the same forecasting questions in 2023 and 2024.

We recruited superforecasters with assistance from Good Judgment, Inc. To find experts, we contacted organizations, academic departments, and research labs working on existential-risk-related issues; we also made several posts via social media and websites such as the Effective Altruism Forum. We received hundreds of expressions of interest and offered slots to the most qualified among the interested applicants. The final expert sample included 32 AI experts, 12 biorisk experts, 12 nuclear experts, 9 climate experts, and 15 "general" experts who study existential risks more broadly (referred to as "x-risk generalists"). Many in the expert pool were affiliated with the Effective Altruism (EA) community; 42% of experts participating in the XPT reported having attended an EA meetup in the past.

The median expert in the XPT forecasted a 20% probability of global catastrophe—defined as a loss of at least 10% of the global population—and a 6% probability of human extinction by 2100. Superforecasters viewed the world as less risky, forecasting a 9% and 1% probability of global catastrophe and human extinction by 2100, respectively. This held across domains, though not uniformly: superforecasters and experts were much further apart on risk related to AI than on the risk of nuclear war.

Participants in the tournament forecasted on questions set to resolve at various dates ranging from as early as mid-2024 to as late as 2100. The 59 forecasting questions in the XPT broke down into 172 subquestions. Of these, 32 questions (38 subquestions) have resolved as of the writing of this report. The resolved questions provide us with a unique opportunity to evaluate forecasting accuracy across different expertise groups, identify key surprises, and explore the relationship between near-term forecasting accuracy and predictions of long-term existential risks.

While we analyze all resolved questions in our dataset, our confidence in resolutions varies across questions (see Table A1.2 in the Appendix). Of the 32 resolved forecasting questions, 47% (15/32) have been definitively resolved based on authoritative data sources, while 53% (17/32) have been provisionally resolved based on available evidence or expert consultation. These provisional resolutions reflect two constraints. First, some questions require expert panels for adjudication (particularly biorisk questions lacking clear ground truth). Second, others await authoritative data publications like International Energy Agency (IEA) reports or labor statistics from the Organisation for Economic Co-operation and Development (OECD).

2. Forecasting Performance

2.1 Accuracy Metrics

We measure forecasting performance using two main accuracy metrics.

Our primary accuracy metric is the **Accuracy Score**. *Accuracy Score* evaluates forecasting performance using the original XPT scoring rules (log score for binary questions; S score for continuous questions). *Accuracy Score* is standardized to measure performance relative to the median XPT participant (i.e., the median across experts and superforecasters). For example, an *Accuracy Score* of 0.25 means a forecaster was 0.25 standard deviations more accurate than the median XPT participant. An *Accuracy Score* of 0.25 would place them roughly in the top 40% of all forecasters. Higher *Accuracy Score* values indicate better accuracy.

Our secondary accuracy metric is **Standardized Absolute Forecast Error** (*SAFE*). *SAFE* measures how "surprising" the actual outcome was relative to forecasters' expectations. For example, a *SAFE* of 1.0 means the outcome was 1 standard deviation from the forecast—corresponding to a moderate but not extreme surprise. Lower *SAFE* values indicate better accuracy.

Table 2.1 provides a summary and interpretation of our main accuracy metrics. **Technical details are provided in the Appendix**.

Metric	Description	Interpretation	Use case
Accuracy Score (*primary*)	Average standardized score across all questions. *Higher* values are better.	*Accuracy Score* = 0.25 means a forecaster is 0.25 standard deviations more accurate than the median XPT participant, placing them roughly in the top 40% of all forecasters.	Used to measure relative performance.
Standardized Absolute Forecast Error, ***SAFE*** (*secondary*)	Average absolute forecast error in units of predictive standard deviations. *Lower* values are better.	*SAFE* = 1.0 means outcomes are on average one standard deviation from forecasters' expectations (corresponding to 16th/84th percentile realizations).	Used to measure absolute performance; how "surprising" questions were to forecasters.

Table 2.1: *Primary and secondary accuracy metrics used to evaluate forecasting accuracy.*

To provide an apples-to-apples comparison between the different groups of XPT participants, **we calculate accuracy metrics at the individual forecaster level**.

Box 1: Individual versus aggregate forecasts

When analyzing forecasting performance, it is important to distinguish between individual and aggregate forecasts. *Individual forecasts* represent each forecaster's predictions, while *aggregate forecasts* combine predictions from multiple forecasters within a group.

For individual-level accuracy, we calculate metrics for each forecaster separately, and then take the median across all individuals in a group. For aggregate-level accuracy, we first combine forecasts via median aggregation, and then calculate the accuracy of that combined forecast.

To compare accuracy between different groups of XPT participants, we use individual-level metrics. This is important to ensure fair comparisons. The median subquestion in the XPT has 32 superforecaster predictions versus only 4 domain-expert predictions. Since aggregating more forecasters improves accuracy via a wisdom-of-the-crowd effect, comparing group aggregates would unfairly advantage superforecasters due to their greater sample size.

Outside of group comparisons, however, we primarily analyze aggregate forecasts. In particular, we use aggregate forecasts when examining substantive questions—such as whether forecasters correctly anticipated AI progress or developments in climate technology (see Section 3). The reason is that aggregation yields more accurate predictions. As a result, the aggregate forecasts produce the most reliable measure of the XPT participants' collective judgment.

2.2 Relative Accuracy

2.2.1 Main Results

Figure 2.1 summarizes the overall forecasting performance.

The graph provides the *Accuracy Score* of the median XPT participant (i.e., individual-level accuracy) by subgroup:

- *Superforecaster*: Forecasters with a proven track record of high accuracy on near-term forecasting questions;
- *Domain Expert*: Subject-matter experts answering questions within their specific area of expertise;
- *Non-domain Expert*: Subject-matter experts answering questions outside their primary area of expertise;
- *X-risk Generalist*: Experts specializing in existential risks.

Note that the same expert may be classified differently across questions. For example, an AI expert is classified as a domain expert when forecasting progress on AI benchmarks but as a non-domain expert when predicting green hydrogen costs. The sample size and composition of the final dataset is provided in the Appendix (Table A3.1). Accuracy results at the group level are given in the Appendix (Table A3.3).

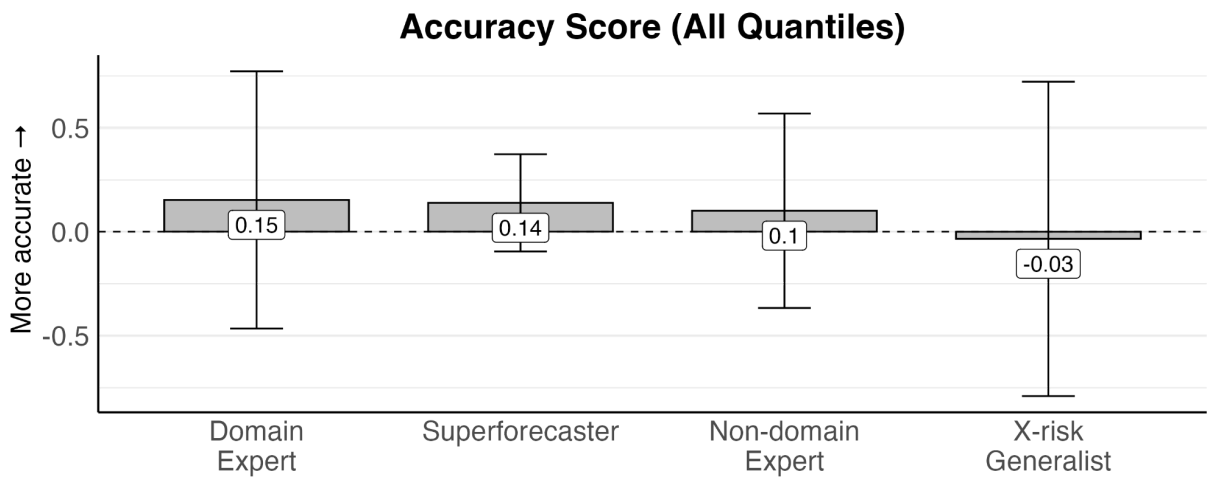


Figure 2.1: *For each group (non-domain expert, domain expert, superforecaster, x-risk generalist), the error bars indicate the Accuracy Score of the median individual in that group. The whiskers provide 95% bootstrap confidence intervals.*

Overall, **performance differences between groups were small**, with only a 0.18 standard-deviation gap between the top and bottom groups. For context, a difference of 0.18 in the *Accuracy Score* corresponds to a difference of approximately 8 percentiles—comparable to the difference between someone performing at the median (50th percentile) versus someone performing slightly above average (around the 58th percentile). Superforecasters and domain experts achieved an almost identical *Accuracy Score*. Intuitively, these results indicate that there was no consistent pattern in accuracy: for some questions, domain experts were more accurate; for others, superforecasters were closer to the truth. In the Appendix (Table A1.1), we provide a question-by-question table with superforecaster and domain expert predictions (group-level aggregates) and their forecast errors, highlighting the same pattern.

Consistent with the above finding, the **performance differences between groups were not statistically significant**, as we document below in Figure 2.3. Therefore, we cannot confidently conclude that superforecasters, domain experts, or other groups demonstrated meaningfully higher forecasting accuracy. This finding is consistent with previous research showing that superforecasters do not have a consistent edge over domain experts (or vice versa).

Domain experts were slightly more accurate when predicting within their area of expertise. However, this difference was small in absolute value (a difference of 0.05 in the *Accuracy Score*) and not statistically significant. This finding suggests limited gains from specialized knowledge in this specific forecasting context.

2.2.2 Performance against Benchmarks

Next, we compare the quality of predictions made by XPT participants (experts and superforecasters) to two benchmarks:

- Sample of public participants;

- Simple prediction algorithms (see “Methods” in the Appendix for details):
    - Naive “no-change” forecast (predict no change);
    - Naive “extrapolation” forecast (extrapolate the current trend).

Since we did not elicit the full set of quantile predictions for the public-participant sample, only the 50th-percentile predictions are used for this benchmarking exercise.

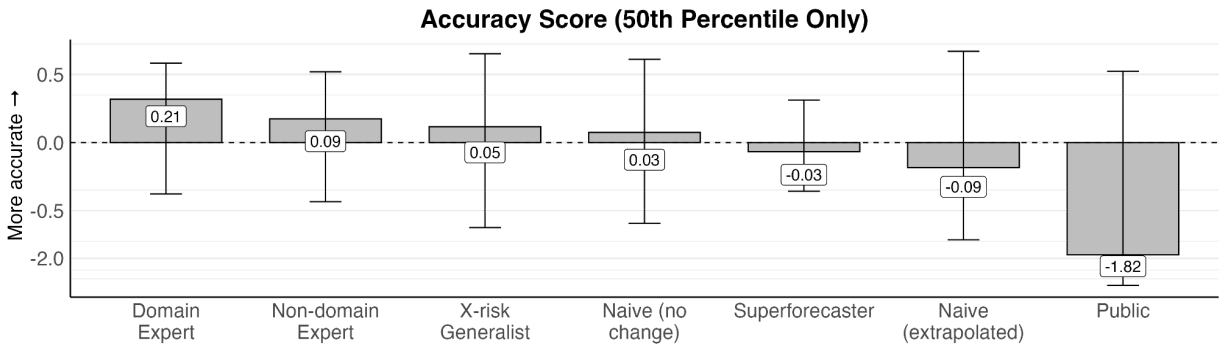


Figure 2.2: *For each group (domain expert, non-domain expert, x-risk generalist, superforecaster, public), the bars indicate the Accuracy Score of the median individual in that group. For the two prediction algorithms (no change and extrapolated), the bars directly indicate their performance. The whiskers indicate 95% bootstrap confidence intervals. Only 50th-percentile predictions are used in the construction of this graph. The y-axis is log-transformed, so visual distances may understate true differences.*

Figure 2.2 provides the comparison with our benchmarks. We observe the following takeaways:

- **XPT participants outperformed public participants.** The median public participant performed substantially worse than XPT forecasters, with an *Accuracy Score* of –1.82. This underperformance is large: a 1.82-point gap in the *Accuracy Score* corresponds to the difference between participants at the 50th and 3rd percentiles of the forecasting accuracy distribution. As shown below in Figure 2.3, this difference is weakly statistically significant ($p < 0.10$) when comparing public participants to the full XPT sample. Domain experts and non-domain experts showed stronger outperformance ($p < 0.05$), while superforecasters exhibited weaker outperformance ($p < 0.10$).
- **The median XPT participant did not outperform statistical benchmarks.** The accuracy differences between individual XPT participants and statistical benchmarks were small and not statistically significant. In fact, the simple “no-change” benchmark (*Accuracy Score* of 0.03) slightly outperformed both the median XPT participant and the median superforecaster, highlighting the difficulty of beating naive statistical rules.

We note that certain features of the XPT tournament may have favored simple prediction algorithms. First, a substantial portion of subquestions (8/38) concerned low-probability events that did not occur during the resolution period. These included questions about biological and nuclear weapon use (for example, Q15–18 and Q31). For all these subquestions, the no-change prediction of zero matched the actual outcome perfectly. Second, several questions tracked slowly-evolving variables for which historical trends provide strong predictive power, such as labor force participation rates (Q38) and nuclear warhead counts (Q33). By contrast, in dynamic domains like AI, these simple algorithms performed substantially worse. As we document in the

Appendix (Table A3.4), the no-change and extrapolation algorithms achieved *SAFE* scores of 1.89 and 1.35 respectively on AI questions—substantially worse than their full-sample values of 1.04 and 0.94.

Finally, we statistically test the relative performance of different forecasts, including aggregated group-level predictions. The results are provided in Figure 2.3.

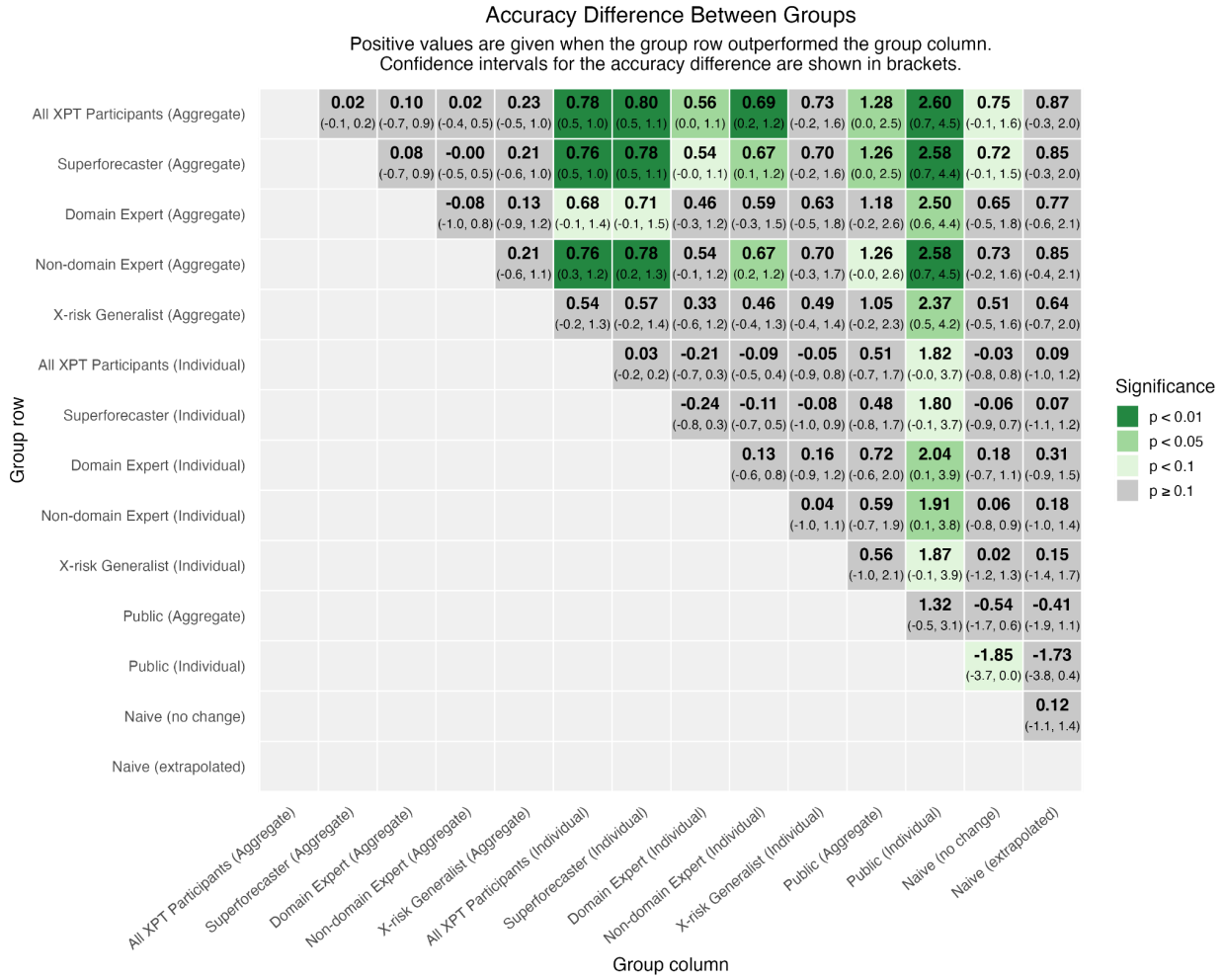


Figure 2.3: *Comparison of Accuracy Score differences across different forecasts; only 50th-percentile predictions are used to calculate the Accuracy Score. Bootstrap 95% confidence intervals appear in parentheses.*

A key insight that emerges is that **aggregated XPT forecasts were substantially more accurate and showed evidence of outperforming statistical benchmarks**. Consistent with the forecasting literature, aggregated forecasts substantially outperformed individual forecasts. The aggregate of all XPT participants achieved an *Accuracy Score* of 0.97 when using all quantile predictions and 0.78 when only median forecasts were used—a large improvement over the median individual participant (see Table A3.3 in the Appendix). While the aggregated

forecast outperformed both naive benchmarks by a large margin in absolute terms, statistical significance varied. The aggregated forecast showed weak statistical evidence of outperforming the "no-change" benchmark ($p < 0.10$) but did not statistically significantly outperform the "extrapolation" benchmark.

Due to the limited number of resolved questions, our statistical power to detect small accuracy differences between groups is constrained. However, Figure 2.3 shows that, at the individual level, the 95% confidence interval for the accuracy difference between superforecasters and domain experts is (-0.8, 0.3). Here, negative numbers indicate greater accuracy by domain experts. This finding allows us to rule out large performance differences: with 95% confidence, the true accuracy gap between these groups is less than 0.8 standard deviations in either direction.

2.2.3 Robustness Tests and Other Analyses

A natural concern when evaluating forecasting performance is whether the results depend on the chosen accuracy metric. To address this concern, we examined forecasting performance using six different accuracy measures, including our primary *Accuracy Score* and alternative metrics like standardized absolute forecast error (*SAFE*), percentile accuracy, and mean standardized squared error; see Table A3.2 in the Appendix for the full results. Our **core findings remain robust across all metrics** (i.e., the differences between XPT participant groups remain small; XPT participants outperform public participants; individual XPT participants have similar accuracy to the two naive statistical benchmarks).

In the Appendix (Appendix 4: Forecaster Calibration), we also analyze forecaster calibration. Overall, we find that **forecasters are overconfident at the individual level but well-calibrated when aggregated at the group level**. At the individual level, forecasters are overconfident when predicting less likely tail events (i.e., they underestimate the probability of tail events). The fact that group-level forecasts are well-calibrated provides additional confidence when using predictive standard deviations to calculate the SAFE metric, as the group-level predictive standard deviations appear to accurately reflect the uncertainty present in the real world.

Finally, we examined whether near-term forecasting accuracy correlates with *intersubjective accuracy*, i.e., participants' ability to predict other forecasters' predictions. While previous research has found that intersubjective accuracy often correlates with real-world forecasting performance, **intersubjective accuracy was not correlated with near-term accuracy** in our data (see Figure A3.1 in the Appendix). This null result may suggest that intersubjective accuracy is less informative in our specific empirical context. Alternatively, our sample of 38 resolved subquestions may be too small to reliably detect a meaningful relationship.

3. Key Surprises and Insights

We next examine areas in which aggregate forecasts—which demonstrated strong overall accuracy through a wisdom-of-the-crowd effect—most notably diverged from reality. We identify the "most surprising" questions based on standardized absolute forecast errors (*SAFE*) at the group level, revealing systematic patterns in what forecasters found difficult to predict (Tables 3.1 and 3.2). We first present the top-10 most surprising questions for each group, and then dive deeper into three key domains where forecasters' expectations most diverged from actual outcomes: biological weapons (Section 3.1), climate technology (Section 3.2), and artificial intelligence (Section 3.3).

3.1 Biological Weapons

Both domain experts and superforecasters overestimated the number of countries with biological weapons programs by the end of 2024. Experts predicted an average of 6.5 countries, while superforecasters predicted 5 countries, an overestimation by a factor of 2.5–3.3 relative to our projected resolution of 2 countries. For several specific countries (i.e., China, Iran, Syria, and Israel), both groups also overestimated the fraction of a panel of 100 biosecurity experts who would agree that the country has an active biological weapons program. Here, multiple countries had forecast errors with *SAFE* values exceeding 1, indicating moderate surprises. However, as discussed in more detail in "Ambiguous Resolutions" in the Appendix, it is difficult to unambiguously resolve this question, which could explain part of the apparent surprise.

Question	Median Forecast	Resolution	SAFE	N
45. Maximum Compute Used in an AI Experiment	100,000	578,703.7	1.92	33
49. Largest Number of Parameters in a Machine Learning Model	100 trillion	10 trillion	1.71	31
30. Cost of Hydrogen	4.5 USD/kg	7.5 USD/kg	1.70	32
40. "Massive Multitask Language Understanding" Benchmark	77.75%	88.7%	1.59	32
20. Individual Countries with Biological Weapons Programs (China)	70%	30%	1.51	26
21. Number of Countries with Biological Weapons Programs	5	2	1.45	32
39. MATH Dataset Benchmark	71%	87.92%	1.38	30
20. Individual Countries with Biological Weapons Programs (Iran)	60%	30%	1.18	28
35. GPT Revenue (Hanson Wins Bet that GPT Revenue < $1B)	53.5%	0%	1.07	32
20. Individual Countries with Biological Weapons Programs (Israel)	40%	10%	1.01	27

Table 3.1: *Most surprising questions, superforecasters (group-level forecast). The table provides the top-10 questions with the largest standardized absolute forecast errors (SAFE) for the group. N denotes the number of forecasters in the group.*

Question	Median Forecast	Resolution	SAFE	N
32. Total Nuclear Warheads	9,949	12,331	2.93	1
49. Largest Number of Parameters in a Machine Learning Model	150 trillion	10 trillion	2.74	7
30. Cost of Hydrogen	3.5 USD/kg	7.5 USD/kg	2.27	2
21. Number of Countries with Biological Weapons Programs	6.5	2	2.17	4
29. Annual Direct Air CO2 Capture	0.6 Mt/year	0.01 Mt/year	1.52	7
20. Individual Countries with Biological Weapons Programs (Iran)	61.5%	30%	1.24	4
38. Labor Force Participation Rate in OECD	77.2%	79.86%	1.22	4
20. Individual Countries with Biological Weapons Programs (Syria)	52.5%	25%	1.10	4
35. GPT Revenue (Hanson Wins Bet that GPT Revenue < $1B)	45%	0%	0.90	6
20. Individual Countries with Biological Weapons Programs (China)	51%	30%	0.79	3

Table 3.2: *Most surprising questions, domain experts (group-level forecast). The table provides the top-10 questions with the largest standardized absolute forecast errors (SAFE) for the group. N denotes the number of forecasters in the group.*

3.2 Climate Technology

Forecasters were overly optimistic about progress in climate technology. Both groups expected a more substantial decrease in the cost of hydrogen produced using renewable electricity: superforecasters expected the cost of hydrogen production to decrease to 4.5 USD/kg in 2024, while domain experts predicted an even greater decline to 3.5 USD/kg. By contrast, we currently project a resolution of 7.5 USD/kg for the question. The *SAFE* values for this question are in the range of 1.70–2.27, suggesting large surprises. Similarly, XPT participants anticipated greater advances in carbon removal. For total direct air capture and storage, domain experts and superforecasters predicted 0.6 and 0.32 $MtCO_2$/year in 2024, respectively, while we currently project just 0.01 $MtCO_2$/year.

3.3 Artificial Intelligence

Both domain experts and superforecasters misjudged the pace and direction of AI progress. Both groups predicted lower values for the maximum compute used in an AI experiment by the end of 2024, with superforecasters underestimating the actual maximum by a factor of five. At the same time, both domain experts and superforecasters overestimated the size of the largest machine learning models by the end of 2024 (1.00E+14 parameters and 4.00E+14 parameters respectively), projecting parameter counts ten times higher than provisionally resolved (1.00E+13 parameters). However, as we note in the Appendix (Section A1.2), this overestimation likely has to do with incorrect base rate information provided to participants during the XPT tournament.

XPT participants systematically underestimated AI progress on multiple benchmarks, with superforecasters exhibiting larger underestimation. Figure 3.1 shows the probability XPT participants assigned to observed outcomes on various AI benchmarks, calculated using an estimated density function (see "Methods" in the Appendix). GPT-4 Turbo achieved 87.82% on the MATH Dataset Benchmark in April 2024; domain experts and superforecasters had assigned a 21.4% and a 9.3% probability, respectively, to reaching this level by June 30, 2024. Both GPT-4o and Claude 3.5 Sonnet achieved 88.7% on MMLU by mid-2024, an outcome that had been assigned a 25.0% and a 7.2% probability for the June 30, 2024 resolution date. RAPTOR + GPT-4 scored 69.3 on QuALITY's hard subset in June 2023—a full year before the resolution date—yet domain experts and superforecasters had assigned only a 43.5% and a 20.1% probability to this achievement by June 30, 2024. Across these three benchmarks, superforecasters assigned probabilities 12–23 percentage points below those of domain experts.

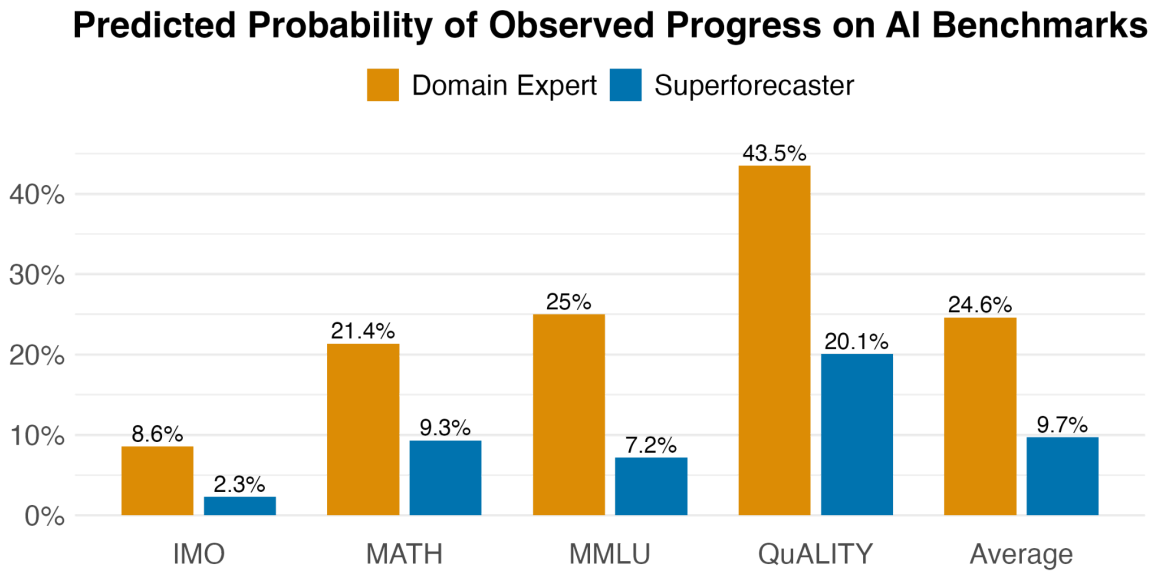


Figure 3.1: *Superforecasters' and domain experts' predicted probabilities of observed progress on AI benchmarks. Probabilities were calculated based on the estimated probability density functions (see Appendix 5) and the observed resolution values. Appendix 2 provides the methodological details on the density function estimation.*

Among the most surprising developments was the performance of AI systems on the International Mathematical Olympiad (IMO). While not officially an AI benchmark, the IMO in recent years has "become an aspirational challenge for AI systems as a test of their advanced mathematical problem-solving and reasoning capabilities." Domain experts and superforecasters did not anticipate an AI system to win a gold medal in the International Mathematical Olympiad (IMO) until after 2030. In July 2025, both Google DeepMind and OpenAI reported that their models achieved gold-level performance in the IMO 2025 competition—5 years earlier than the median expert prediction and 10 years earlier than the median superforecaster prediction. Domain experts and superforecasters only expected an 8.6% and a 2.3% probability of this achievement on or before 2025.

We note that the XPT tournament concluded prior to the public release of ChatGPT at the end of 2022, which marked the beginning of an intense phase of AI investment and capability acceleration. While domain experts were more calibrated to trends in AI progress than superforecasters, at times even their judgment failed to anticipate the speed of advancement. These results align closely with previous reports about how experts were surprised by progress in language models in 2022 and 2023, particularly as it related to the MMLU, MATH, and the International Mathematical Olympiad.

4. Long-Term Risk Implications

A key goal of the original XPT tournament was to obtain forecasts for long-term risks facing humanity. XPT participants forecasted two types of risks: *catastrophic risks* (the probability of more than 10% of the global population dying within a 5-year period) and *extinction risks* (the probability of human extinction or a reduction of the global population below 5,000). The tournament assessed these risks across multiple domains: genetically-engineered and naturally-occurring pathogens, artificial intelligence, nuclear weapons, non-anthropogenic causes (such as asteroids or supervolcanoes), and overall risk from all causes combined.

A natural question is whether more accurate near-term forecasters made systematically different long-term risk predictions. Figure 4.1 suggests that there is **no meaningful relationship between near-term accuracy and long-term risk forecasts**. Across accuracy quartiles (from least accurate in quartile 1 to most accurate in quartile 4), median risk estimates remain fairly flat for all risk categories, and there is no statistically significant correlation between accuracy and long-term risk forecasts. The correlation coefficients all cluster around zero, ranging from -0.08 to 0.14, and they are not statistically significant.

In the Appendix (Figure A3.2), we examine how long-term risk forecasts relate to near-term accuracy in our sample of public participants. An advantage of using this sample is that most public participants provided a forecast on every question, eliminating issues surrounding self-selection into questions. In particular, the median public participant answered 36 out of the 38 resolved subquestions. For the public participants, unlike the main XPT sample, we observe a statistically significant *negative* correlation (i.e., the most accurate public forecasters predicted lower risks).

Overall, **our findings challenge the hope that near-term accuracy can reliably identify forecasters with more credible long-term risk predictions**. These results are consistent with the analysis from the original XPT report. The original XPT report found that, for “AI-concerned” (the third of participants with the highest forecast of AI extinction risk by 2100) and “AI-skeptic” (the third of participants with the lowest forecast of AI extinction risk by 2100) groups, their near-term forecasts were in strong agreement (see Table 26 in Appendix 4). The same was also true for superforecasters and domain experts (see Table 28 in Appendix 4).

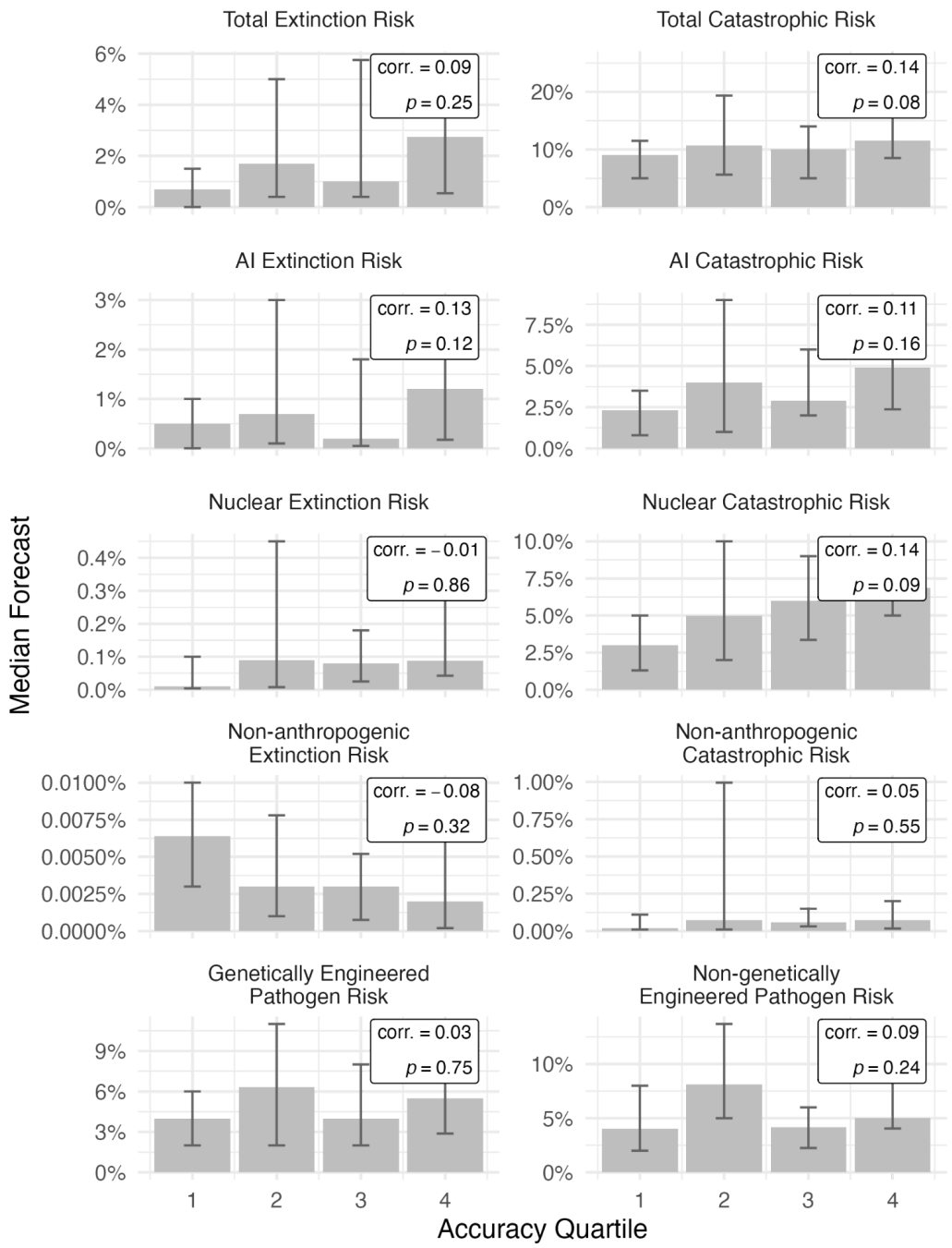


Figure 4.1: *XPT participants' forecasts on catastrophic and extinction risks by 2100. "Catastrophic risk" is defined as the probability of 10% or more of humans dying within a 5-year period (except for pathogen risks, which use a 1% threshold). "Extinction risk" is defined as the probability of human extinction or a reduction of the global population below 5,000. Participants are divided into quartiles based on their near-term accuracy, from least (1) to most (4) accurate. Error bars represent 95% bootstrap confidence intervals for the median risk forecast within each quartile. Only forecasters with at least 10 resolved near-term forecasts are included. Labels show the Spearman rank correlation between individual-level accuracy and long-term risk forecasts as well as the corresponding p-value.*

5. Conclusions

This report provides the first empirical assessment of forecasting accuracy in the Existential Risk Persuasion Tournament (XPT). We conclude by discussing the limitations of this work and highlighting next steps.

5.1 Limitations

Some methodological limitations should be considered when interpreting our results:

- **Limited statistical power.** With only 38 resolved subquestions—further subdivided across different domains—our ability to detect statistically significant differences between forecaster groups is constrained. Most observed accuracy differences between groups did not reach statistical significance, limiting any conclusions about relative expertise.
- **Limited implications for long-term risks.** This analysis covers only questions resolved by mid-2025. Despite observing, for example, faster-than-expected AI progress, this short timeframe provides limited basis for updating beliefs on long-term existential risks.
- **Non-representative expert sample.** The XPT relied on a nonrepresentative expert sample with a 34% attrition rate by the end of the tournament. (See Appendix 1 in the original XPT report.) The experts who participated may not accurately represent the broader expert communities in their respective fields.
- **Post-hoc benchmark definition.** Simple algorithmic benchmarks (no change, extrapolation) were developed after data collection rather than defined a priori. This post-hoc approach may introduce hindsight bias and make tournament participants appear less accurate than they actually were.
- **Ambiguous resolutions.** While 38 subquestions have resolved, our confidence in each resolution varies from question to question. While many questions have been definitively resolved (i.e., according to the criteria specified in the original XPT report), others have provisional resolutions that may change in the future. For more details on potentially ambiguous resolutions, see Appendix 1.

5.2 Looking Forward to 2030

While the questions resolved by mid-2025 have provided valuable initial insights, we are looking forward to the next wave of questions set to resolve in 2030. These questions will offer deeper insights into potential existential risks:

- **AI development and impact.** Given the faster-than-expected progress on AI benchmarks, we are interested to track how this acceleration continues in the coming years. Question #51 asks whether Nick Bostrom affirms the existence of AGI by 2030, where superforecasters estimated just a 1% probability compared to domain experts' 9%. Another key milestone is Question #44 ("Date of first publicly known advanced AI"). For this question, superforecasters predicted 2060 while domain experts predicted 2046. Beyond technical advancements, we will assess broader economic impacts through forecasts on US computer R&D spending (Question #37), labor force participation in

OECD countries (Question #38), and the percentage of US GDP from software and information services (Question #36).

- **Climate trajectory and technology.** Critical climate questions with 2030 resolution dates include global surface temperature change (Question #25), where superforecasters predicted 1.47°C warming versus domain experts' 1.4°C estimate. We will also assess progress on climate technologies through questions about green hydrogen production costs (Question #30), direct air carbon capture (Question #29), and electricity share from solar and wind energy (Question #28). These resolutions will be particularly telling given the current overestimation of climate technology development.
- **Global risk forecasts.** While most existential risk forecasts for 2030 were very low, we will track several important risk predictions that resolve by this date. For public health emergencies, both superforecasters and domain experts predicted approximately 2 declarations of a public health emergency of international concern (PHEIC) with at least 10,000 deaths by 2030 (Question #22). We will also monitor forecasts about nuclear weapon use causing significant casualties (Question #31). As these and other 2030 questions resolve, they will also enable us to answer crucial meta-questions: What is the relationship between near- and medium-term (5–8 years) forecasting accuracy? Do forecasters with high medium-term accuracy make systematically different predictions on long-term existential risks?

5.3 Next Steps

Building on the insights from this initial analysis, we plan to take the following next steps:

- **Develop specialized AI insights.** Given the faster-than-expected progress on AI benchmarks, researchers at the Forecasting Research Institute are in the process of launching multiple dedicated projects to better understand the likely future trajectories and impacts of AI. These projects include establishing a longitudinal panel of AI experts and conducting a survey of economists on AI's potential economic and labor market effects.
- **Track future resolutions.** We will continue tracking the resolution of questions posed in the XPT. We may also re-engage the original XPT participants and gather data on how their forecasts have changed in light of recent AI advances and other developments.

6. References

1. Armstrong, J.S. (2001). Combining forecasts. In Armstrong, J.S. (ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners* (pp. 417–439). Springer.
2. Brockwell, P.J. and Davis, R.A. (2002). *Introduction to Time Series and Forecasting*. Springer.
3. Durbin, J. and Koopman, S.J. (2012). *Time Series Analysis by State Space Methods*. Oxford University Press.
4. Hyndman, R.J. and Koehler, A.B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688.
5. Jose, V.R.R. and Winkler, R.L. (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting*, 24(1), 163–169.
6. Karger, E., Rosenberg, J., Jacobs, Z., Hickman, M., Hadshar, R., Gamin, K., Smith, T., Williams, B., McCaslin, T. and Tetlock, P.E. (2023). Forecasting existential risks: Evidence from a long-run forecasting tournament. *FRI Working Paper #1*. Online access at https://forecastingresearch.org/xpt.
7. Wand, M.P. (1997). Data-based choice of histogram bin width. *The American Statistician*, 51(1), 59–64.

7. Appendices

Appendix 1: Questions and Resolutions

A1.1 Question-by-Question Resolution Table

Question	Domain	Resolution	Median forecast (SAFE)		Resolution Status
			Superforecaster	Domain Expert	
13. Non-Coronavirus mRNA Vaccine	Biorisk	98,056 doses	75,000 doses (0.007)	60,000 doses (0.011)	Projected
14. Novel Infectious Disease Surveillance System	Biorisk	0% (No such system was established)	20% (0.481)	27.5% (0.662)	Projected
15. Non-State Actor Bioweapon 1k Deaths	Biorisk	0 events	0 events (0)	0 events (0)	Official
16. State Actor Bioweapon 1k Deaths	Biorisk	0 events	0 events (0)	0 events (0)	Official
17. Non-State Actor Bioweapon 100k Deaths	Biorisk	0 events	0 events (0)	0 events (0)	Official
18. State Actor Bioweapon 100k Deaths	Biorisk	0 events	0 events (0)	0 events (0)	Official
19. Lab Leaks	Biorisk	0.002 expected events	0.01 expected events (0.087)	0.03 expected events (0.194)	Projected
20. Individual Countries with Biological Weapons Programs (U.S.)	Biorisk	2%	20% (0.601)	2% (0)	Projected (ambiguous)
20. Individual Countries with Biological Weapons Programs (Russia)	Biorisk	80%	80% (0)	68% (0.569)	Projected (ambiguous)
20. Individual Countries with Biological Weapons Programs (China)	Biorisk	30%	70% (1.507)	51% (0.791)	Projected (ambiguous)
20. Individual Countries with Biological Weapons Programs (North Korea)	Biorisk	70%	77.5% (0.34)	70% (0)	Projected (ambiguous)

20. Individual Countries with Biological Weapons Programs (Israel)	Biorisk	10%	40% (1.013)	30.5% (0.693)	Projected (ambiguous)
20. Individual Countries with Biological Weapons Programs (Iran)	Biorisk	30%	60% (1.183)	61.5% (1.242)	Projected (ambiguous)
20. Individual Countries with Biological Weapons Programs (Syria)	Biorisk	25%	35% (0.4)	52.5% (1.101)	Projected (ambiguous)
21. Number of Countries with Biological Weapons Programs	Biorisk	2 countries	5 countries (1.447)	6.5 countries (2.17)	Projected (ambiguous)
22. PHEIC Declarations with 10k Deaths	Biorisk	0 declarations	0.5 declarations (0.805)	0 declarations (0)	Projected
23. Assassinations with Biological Weapons	Biorisk	0 assassinations	0 assassinations (0)	0 assassinations (0)	Projected
24. Malaria Deaths	Biorisk	585,521.7 deaths	612,500 deaths (0.253)	597,500 deaths (0.112)	Projected
26. Cost of Utility-Scale Solar Energy	Climate	$0.0378	$0.038 (0.018)	$0.033 (0.441)	Official
28. Solar and Wind Energy	Climate	15%	14% (0.439)	14.25% (0.329)	Official
29. Annual Direct Air CO2 Capture	Climate	0.01 MT CO2/yr	0.32 MT CO2/yr (0.785)	0.6 MT CO2/yr (1.519)	Projected
30. Cost of Hydrogen	Climate	$7.50	$4.50 (1.705)	$3.50 (2.273)	Projected
31. Nuclear Weapon Use	Nuclear	0% (No such event occurred)	1.5% (0.106)	2% (0.141)	Official
32. Total Nuclear Warheads	Nuclear	12,331 warheads	12,700 warheads (0.455)	9,949 warheads (2.935)	Official
33. Countries with Nuclear Warheads	Nuclear	9 countries	9 countries (0)	9 countries (0)	Official
35. GPT Revenue	AI	0% (Hanson lost the bet)	53.5% (1.075)	45% (0.901)	Projected
36. US GDP From Software	AI	3.25%	3.45% (0.352)	3.6% (0.616)	Official
37. US Computer R&D Spending	AI	$217,022,161,400	$192,500,000,000 (0.633)	$240,000,000,000 (0.593)	Projected
38. Labor Force Participation Rate in OECD	AI	79.86%	78.25% (0.738)	77.2% (1.216)	Projected

39. MATH Dataset Benchmark	AI	87.92	71 (1.377)	80 (0.644)	Official
40. Massive Multitask Language Understanding Benchmark	AI	88.7	77.75 (1.59)	84.8 (0.566)	Official
41. QuALITY Dataset Benchmark	AI	69.3	58 (0.851)	69 (0.023)	Official
42. AI Wins International Mathematical Olympiad	AI	2025	2035 (0.748)	2030 (0.374)	Projected
45. Maximum Compute Used in an AI Experiment	AI	578703.7 petaFLOPS-days	100000 petaFLOPS-days (1.923)	420680 petaFLOPS-days (0.635)	Official
46. Largest AI Experiment Cost of Compute	AI	$45,762,654.80	$35,000,000 (0.198)	$65,000,000 (0.354)	Projected (ambiguous)
47. Lowest Price of GFLOPS	AI	$0.01 (0)	$0.011 (0.179)	$0.011 (0.201)	Projected
49. Largest Number of Parameters in a Machine Learning Model	AI	1.00E+13 parameters	1.00E+14 parameters (1.712)	4.00E+14 parameters (2.743)	Projected (ambiguous)
50. Negative Public Opinion of AI	AI	32.9%	33% (0.012)	33% (0.012)	Projected

Table A1.1: *An overview of XPT questions which have been resolved as of the publication date of this report. Certain questions listed here have multiple resolution dates; each resolution associated with those questions refers to the question's 2024-resolving component. For the full question details, including resolution criteria, please refer to Appendix 5 in the original XPT report. Standardized absolute forecast errors (SAFE) provided next to the median forecasts in parentheses.*

A1.2 Resolution Statuses

Question	Domain	Resolution Status	Notes
13. Non-Coronavirus mRNA Vaccine	Biorisk	Projected	mRESVIA, an mRNA vaccine targeting RSV from Moderna, was approved by the FDA in 2024 and was available commercially in the United States by the end of that year. As of the end of 2024, it was the only known vaccine matching the resolution criteria of this question. FRI projected the number of doses administered to be 98,056 [86207, 110837] based on a model from these estimates and advice from a biorisk expert. This question will be resolved officially by a panel of experts.
14. Novel Infectious Disease Surveillance	Biorisk	Projected	A novel infection disease surveillance system meeting our resolution criteria was not announced by the end of 2024.

System			This question will be resolved officially by a panel of experts.
15. Non-State Actor Bioweapon 1k Deaths	Biorisk	Official	An event in which a non-state actor's bioweapon was responsible for the deaths of 1,000 people did not occur.
16. State Actor Bioweapon 1k Deaths	Biorisk	Official	An event in which a state actor's bioweapon was responsible for the deaths of 1,000 people did not occur.
17. Non-State Actor Bioweapon 100k Deaths	Biorisk	Official	An event in which a non-state actor's bioweapon was responsible for the deaths of 100,000 people did not occur.
18. State Actor Bioweapon 100k Deaths	Biorisk	Official	An event in which a state actor's bioweapon was responsible for the deaths of 100,000 people did not occur.
19. Lab Leaks	Biorisk	Projected	Outbreaks of cholera, dengue, and malaria were responsible for the deaths of more than 1,000 people from 2022 through 2024. Each of these diseases has regularly established vectors of transmission, and outbreaks of these diseases are extremely unlikely to have been caused by lab leaks. A plausible expected number of events was estimated by a panel of FRI research assistants and an external biorisk expert to be 0.002. This question will be resolved officially by a panel of experts.
20. Individual Countries with Biological Weapons Programs	Biorisk	Projected (ambiguous)	FRI researchers came up with the below estimates based on biorisk experts' current views on the probability that each country has biological weapons programs. We had two independent biorisk experts vet the credibility of and give feedback on the following estimates: ● US: 2% ● Russia: 80% ● China: 30% ● North Korea: 70% ● Israel: 10% ● Iran: 30% ● Syria: 25% This question will be resolved officially by a panel of experts. For more details on the resolution for this question, see the "Ambiguous Resolutions" subsection of this appendix.
21. Number of Countries with Biological Weapons Programs	Biorisk	Projected (ambiguous)	Based on the views of consulting biorisk experts, the determination was made that a panel of biorisk experts would most likely conclude that a median of two countries currently maintain biological weapons programs: Russia and North Korea. This question will be resolved officially by a panel of experts. For more details on the resolution for this question, see the "Ambiguous Resolutions" subsection of this appendix.
22. PHEIC Declarations with 10k Deaths	Biorisk	Projected	The World Health Organization (WHO) declared two public health emergencies of international concern (PHEICs) in this timeframe involving Clade II mpox and Clade I mpox, neither of which caused 10,000 deaths within the established timeframe. The former outbreak, the PHEIC of which ended in 2023, killed 207 people; the latter outbreak, which is ongoing, has killed several hundred people and is unlikely to cross the 10,000 death threshold required for this question to resolve positively. This question will be resolved officially once the PHEIC on Clade

			I mpox is lifted.
23. Assassinations with Biological Weapons	Biorisk	Projected	From the start of the XPT through 2024, three heads of state died while in office. It is extremely unlikely that any were killed by biological weapons. This question will be resolved officially by a panel of experts.
24. Malaria Deaths	Biorisk	Projected	To project an answer for this question, we extrapolated WHO data from 2023 to 2024 and estimated 585,521.7 deaths [557348.1, 613695.3] from malaria in 2024. This question will be resolved officially using results from Institute for Health Metrics and Evaluation's annual Global Burden of Disease reports.
26. Cost of Utility-Scale Solar Energy	Climate	Projected	According to the Department of Energy, the levelized cost of energy (LCOE) for a utility-scale photovoltaic system was $47/MWh in 2024 Q1, or $0.0378/kWh in 2017 USD. This question will be resolved officially by the Department of Energy's next SunShot report, or by a panel of experts if such a report is not published.
28. Solar and Wind Energy	Climate	Official	The International Energy Agency (IEA) reported in its 2025 Global Energy Review that wind and solar photovoltaics made up 8% and 7% of global electricity generation in 2024, summing to a combined 15% (p. 27).
29. Annual Direct Air CO2 Capture	Climate	Projected	According to a February 2024 report from the IEA, “the installed capture capacity of DACS today is less than 0.01 Mt CO_2/year,” (p. 28). Sufficient progress has not been made beyond this point to suggest that the amount of CO_2 being captured is significantly larger. The data referred to in the above report comes from 2023. The question will be resolved officially when the IEA publishes relevant data from 2024, or by a panel of experts if such data is not published.
30. Cost of Hydrogen	Climate	Projected	The *Financial Times*, citing Argus Media, reported in June 2024 that "the cost of producing green hydrogen is still about $5 per kilogramme higher than for grey hydrogen, making it three times more expensive to produce," implying a cost of green hydrogen of $7.50. This roughly aligns with the IEA's Global Hydrogen Review 2024 report, which suggested a price of $7.80 for 2023 (p.82). This question will be resolved officially based on data from the IEA’s Global Hydrogen Review 2025 report, or a panel of experts if such data is not published.
31. Nuclear Weapon Use	Nuclear	Official	Nuclear weapons were not used to kill anyone from the beginning of the XPT through the end of 2024.
32. Total Nuclear Warheads	Nuclear	Official	The Federation of American Scientists updated its “Status of World Nuclear Forces” article in March 2025, estimating that nuclear stockpiles worldwide contained 12,331 warheads as of the beginning of that year.
33. Countries with Nuclear Warheads	Nuclear	Official	The Federation of American Scientists updated its “Status of World Nuclear Forces” article in March 2025. While the number of nuclear warheads in stockpiles worldwide increased, the number of countries with nuclear weapons remained nine.
35. GPT Revenue	AI	Projected	Robin Hanson publicly conceded the bet in a social media post

			on March 16, 2025. It is virtually certain, even outside of the context of the bet, that GPT has generated more than $1 billion in revenue; *The Information* reported in September 2024 that, based on public comments from OpenAI's COO, "the artificial chatbot is conservatively generating more than $225 million in revenue per month, or $2.7 billion on an annual basis, based on publicly available prices of its subscriptions." The resolution criteria leave open the possibility that Hanson could take back the concession of the bet between now and the end of 2025, at which point the question would be resolved officially by a panel of experts.
36. US GDP From Software	AI	Official	The Bureau of Economic Analysis (BEA)'s industry data suggests that 3.25% of total 2024 US GDP can be attributed to the "Publishing industries, except internet (includes software)" and "Data processing, internet publishing, and other information services" industries.
37. US Computer R&D Spending	AI	Projected	The National Center for Science and Engineering Statistics reported that US businesses spent a combined $204.87 billion in research and development in the "Information" and "Computer systems design and related services" industries in 2022. Extrapolating from this point, based on Aswath Damodaran's January 2025 estimates of year-by-year R&D spending in related industries, FRI projects that $217,022,161,400 was spent on R&D in these industries in 2024. This question will be resolved officially when the NCSES releases relevant data for 2024.
38. Labor Force Participation Rate in OECD	AI	Official	The OECD has not yet published labor force participation rate (LFPR) statistics for 2024. Based on this model, projecting based on partial data from member countries of the OECD, we estimate an LFPR of 79.86%. This question will be resolved officially when the OECD releases relevant OECD-wide data for 2024.
39. MATH Dataset Benchmark	AI	Official	GPT-4 Turbo achieved an overall score on the MATH Dataset Benchmark of 87.82% in April 2024. After the official resolution date, Gemini 2.0 Flash achieved an overall score on the MATH Dataset Benchmark of 89.7% in December 2024.
40. Massive Multitask Language Understanding Benchmark	AI	Official	Both GPT-4o and Claude 3.5 Sonnet achieved scores on the MMLU Benchmark of 88.7% in mid-2024. Some models, including Claude 3.5 Sonnet and Gemini Ultra, have touted MMLU scores at or above 90% using Chain-of-Thought (CoT) prompting. We have decided to exclude CoT-derived scores given that some CoT prompts include information specifically tailored for each question by prompt writers, as opposed to other methods which offer, at most, more generalizable information useful across the MMLU question set.
41. QuALITY Dataset Benchmark	AI	Official	According to the official QuALITY leaderboard, RAPTOR (collapsed tree) + GPT-4 achieved an SAT-style score of 69.3 on the hard subset of the QuALITY Dataset in June 2023. In September 2024, powerdrill.ai's "Baseline model: RAPTOR + gpt-4o w/ query intent & entity understanding" achieved an SAT-style score of 69.7 on the hard subset of the QuALITY dataset. Mike Wang's "Clustering and Decomposition using Qwen2.5-7b

			and chat using DeepSeek" model surpassed this shortly after the resolution period ended in January 2025, achieving a score of 76.8.
42. AI Wins International Mathematical Olympiad	AI	Projected	Both OpenAI and Google Deepmind claim to have achieved gold medal-level at IMO 2025. While these performances do not formally meet the standards of the IMO Grand Challenge, which require proofs to be produced in Lean and for models to have been released to the public and open-source prior to the competition, our resolution criteria also specify that a panel of experts could resolve this if any AI model, open or closed, is determined to have the technical capability of winning the challenge; we believe a panel of experts would likely agree with this statement. This question will be resolved officially when a model officially wins the IMO Grand Challenge or by a panel of experts determining that at least one AI model has the technical capability of winning the Challenge.
45. Maximum Compute Used in an AI Experiment	AI	Official	According to Epoch AI, Gemini 1.0 Ultra was trained with just over 5E+25 FLOPS, making it the largest model in these terms in their database with a publication date before 2025. This training compute estimate converts to 578703.7 petaFLOPS-days. The current top model according to Epoch is Grok-3, which was trained with ~4.64E+26 FLOPS. Had the model specifications been published in 2024, this question would have therefore resolved to 5,370,370.37 petaFLOPS-days.
46. Largest AI Experiment Cost of Compute	AI	Projected (ambiguous)	Epoch AI estimates that Gemini 1.0 Ultra was the largest AI experiment trained in the established time period. It additionally estimates that Gemini 1.0 Ultra's final training run cost $29,827,341.92 ($26,525,309.45 in 2021 USD). However, some in the AI community have suggested that using just the cost of a model's final training run to represent training costs is misleading and tends to undersell other costs related to overhead, hardware maintenance, and other costs.. The resolution criteria specified in the original report for this question do not suggest, one way or the other, whether these costs should be included or excluded, and state that a panel of experts will determine cost methodology. Because it is uncertain which method is more similar to an eventual panel of experts' cost estimate, researchers at FRI estimated costs of compute taking into account assumptions around additional costs which would resolve the question to $62.5 million. Then, taking the average of these assumptions and Epoch's published estimate, we projected a resolution to this question of $**45,762,654.80** (2021 USD). For more details on the resolution for this question, see the "Ambiguous Resolutions" subsection of this appendix.
47. Lowest Price of GFLOPS	AI	Projected	Researchers at FRI compiled a list of GPUs to determine the lowest price of GFLOPS. The lowest price was approximately $0.01 for 1 GFLOPS, from the Radeon RX 7600 platform.
48. ImageNet Classification Training Efficiency	AI	Unresolved	Due to a lack of published data round this specific measure of training efficiency, we were unable to establish a resolution for this question.
49. Largest Number of Parameters in a Machine Learning Model	AI	Projected (ambiguous)	After consulting with Epoch AI, researchers at FRI found that M6-10T was trained with 10 trillion parameters, making it the largest number of parameters in a machine learning model.

			For more details on the resolution for this question, see the "Ambiguous Resolutions" subsection of this appendix.
50. Negative Public Opinion of AI	AI	Official	FRI ran a representative survey of 385 American adults on the survey platform Prolific and found that 32.9% felt that the development of AI systems had mostly been a bad thing for society.

Table A1.2: *An overview of resolved XPT questions and context around how they were resolved. An "official" resolution indicates that the question resolved in a way that fits with the question's original resolution criteria; "projected" indicates that the question has either passed its resolution date or FRI researchers have strong reason to believe that resolution has been met, but we have chosen to resolve the question in a way that differs in some way from the question's original resolution criteria; and "unresolved" indicates a question which has resolved but for which we are unable to determine a sufficient resolution due to lack of data. For the full question details, including resolution criteria, please refer to Appendix 5 in the original XPT report.*

A1.3 Ambiguous Resolutions

A1.3.1 Biological Weapons

Our projected resolutions for "20. Individual Countries with Biological Weapons Programs" and "21. Number of Countries with Biological Weapons Programs" differ significantly from superforecaster and domain expert forecasts produced during the tournament. In particular, both groups overestimated by 20 percentage points or more the probabilities our researchers and expert consultants assigned to China and Israel having active biological weapons programs (and had similarly large overestimates for Syria and Iran); and both groups forecasted that the total number of countries with biological weapons programs would be more than double what we project.

Several issues may have contributed to these discrepancies:

- **Lack of clear base rates.** We did not establish a clear base rate for these questions, though we did link to several background reading sources. Several forecasters expressed uncertainty because of this and forecasted near the ignorance prior, adjusting for their own understanding of each country.
- **Ambiguity around "dual use" stipulation in the resolution criteria.** We define a bioweapons program in our resolution criteria as "a government-run program that aims to weaponize or stockpile biological agents for hostile use." We follow up by saying "[d]ual-use weapons that also have a defensive purpose but could be used for hostile purposes count for the resolution of this question." Many forecasters cited this stipulation in their rationales as a reason that they had relatively high forecasts or as a reason that others should raise their forecasts, with some arguing that nationally funded gain-of-function research involving potentially harmful pathogens might resolve the question positively. We believe these types of examples would not necessarily resolve

the question positively as written, but the ambiguity around this point may have led to increased forecasts.

- **Expectations around expert panel biases and beliefs.** Questions 20 and 21 explicitly rely on a panel of experts for resolution, and this point was discussed by XPT forecasters. Some forecasters stated that they felt the panel would likely be composed largely of American experts with most of the rest of the panel being composed of experts from other countries allied with the United States; these forecasters concluded that these experts would be biased toward lower probabilities that America has a bioweapons program and higher probabilities that American adversaries (in particular, China) have bioweapons programs. In attempting to project a resolution for this question, we explicitly instructed our researchers and external biorisk consultants to estimate an answer to the question that was representative of the biorisk community at large, rather than from a specific viewpoint.
- **Small number of bioweapons-knowledgeable biorisk experts.** 12 biorisk experts participated in the XPT. While some may have had expertise in pandemics, they were not necessarily experts in biological weapons—or vice versa. This uneven distribution of expertise, along with the absence of clear base rates, may have left forecasters on Questions 20 and 21 uncertain about how bioweapons experts would quantitatively express their belief. Only three of the 12 biorisk experts gave forecasts for Question 20, and only four gave forecasts for Question 21. Of the four experts who took part on the latter question, one forecasted roughly in line with what our biorisk panel estimated, but did not seem to sway other members of their team toward that forecast; another explicitly stated that they did not have sufficient background knowledge on bioweapons to give a reasonable forecast.

A1.3.2 Cost of Compute

Several aspects of the resolution criteria for “46. Largest AI Experiment Cost of Compute” and publicly available information related to the question make it difficult to resolve unambiguously. While the resolution criteria do specify definitions for the “largest” experiment, multiple interpretations exist. While most of the resolution criteria suggest that “largest” refers to maximum compute used in training, one bullet point suggests that “[f]or this question, we are interested in the most expensive experiment.” We have determined this latter suggestion to be a typo given that it is the only instance where we emphasize cost rather than compute in determining how large an experiment is, but it may have led to confusion among forecasters.

Determining which model used the maximum amount of compute in training is also difficult, since this information is not publicly available for most models. We write that this portion of the question can be resolved by credible estimates from an authoritative source, and have thus relied on Epoch AI’s Notable AI Models dataset to determine which model satisfies the resolution criteria. While we consider Epoch’s data to be authoritative, these estimates are not necessarily stable given how recent many of the relevant models are. For instance, on December 30, 2024, Epoch’s database listed Grok-2 as having been trained with 5.30E+25 FLOP, making it the largest model and beating Gemini 1.0 Ultra’s training compute of ~5E+25;

as of May 12, 2025, Epoch has since revised its estimate of Grok-2's training compute down to 2.96E+25 FLOP, making Gemini 1.0 Ultra the largest model to have been published before the end of 2024. While the current projected resolution using Gemini 1.0 Ultra's cost estimates makes it seem as if superforecasters were more accurate, our initial estimates (made with feedback from staff at Epoch AI) of Grok-2's cost of compute were around $74 million, which would have made domain experts more accurate.

Finally, as stated in the Resolution Statuses table, we do not specify how an eventual panel of experts should resolve the cost of a chosen model. Epoch AI gives estimates for training costs ranging from around $30 million to $130 million depending on various assumptions and definitions around training cost.

When we use Epoch AI's data to produce a result based on our resolution criteria, we get to an estimate of around $65m, or around **$62.5m** in 2021 USD, of which:

- AI Accelerator Chip Cost: ~$29.1 million.
- Other Server Components Cost: ~$18.6 million.
- Cluster-Level Interconnect Cost: ~$11.2 million.
- Energy Cost: ~$6.4 million.

Epoch itself reports an estimate of **$29.8m** ($26.5m in 2021 USD) in its Notable AI Models dataset comes from one of their hardware cost estimation approaches, which uses specific fixed input values in its calculation. In short, this approach calculates the cost per chip-hour for TPU v4 hardware and multiplies it by the total estimated training time.

The difference between this lower estimate and the higher estimates we present above are a result of two factors: how uncertainty is handled and whether experimentation overhead is factored into the calculations. The $30 million is a point estimate based on fixed input assumptions for the final training run only, while the $65 million figure is the median result from a probabilistic analysis that explicitly models the uncertainties in hardware costs and includes a multiplier (typically 1.2x to 4x) to account for the substantial compute used during model development before the final training run.

Both estimates include the same set of components: AI accelerator chips, other server components, cluster-level interconnect costs, and energy costs. Both use hardware depreciation calculations to derive a cost per chip-hour. However, the probabilistic approach uses a wide distribution for hardware cost inputs (with a 90% confidence interval spanning from $5.9M to $110M), and the experimentation overhead factor significantly contributes to the higher median estimate.

Taking the mean between our approach and Epoch's Notable AI Models approach, we project a resolution to this question of **$45.8m** (2021 USD), but acknowledge a much wider range of reasonable resolutions.

A1.3.3 Largest Number of Parameters in a Machine Learning Model

Superforecasters and domain experts both had median forecasts for “49. Largest Number of Parameters in a Machine Learning Model” that overestimated the parameter count by an order of magnitude. This likely has to do with incorrect base rate information provided to participants during the tournament.

In our background document for this question, we included a link to the Parameter, Compute and Data Trends in Machine Learning dataset, a precursor to Epoch AI’s Notable AI Models dataset. That dataset lists BaGuaLu as a model like any other in the dataset. This seems to be a reference to MoDa174-T, a proof-of-concept model described in the BaGuaLu paper used to demonstrate that the training system described by the authors has the capability to train models up to 173.9 trillion parameters, close to our participants’ forecasts. However, our resolution criteria state that a model “cannot be merely a description of a possible system, or a demonstration of scaling a system without application to a task” in order to count toward the resolution of this question, making MoDa174-T ineligible. References to BaGuaLu have since been removed from Epoch AI’s Models datasets.

The model we currently use to resolve the question, M6-10T, was forwarded to us by Epoch AI. We have not found a model trained with a larger number of parameters.

Appendix 2: Methods

A2.1 Data Processing

A2.1.1 Data Sources

This report uses forecast data and participant group labels (for superforecasters and experts) available in the original project's GitHub repository. We also include responses from the public survey.

A2.1.2 Forecast Selection and Processing

As part of the XPT's persuasion element, participants submitted forecasts at multiple stages aligned with the tournament's design. For this analysis, we use only the final forecast each participant submitted for each question. In most cases, this corresponds to forecasts made at Stage 4 of the tournament. However, some participants exited the study earlier; for them, we use the most recent forecast available, which may come from an earlier stage. Forecast timing is tracked using the timestampId variable.

In addition to eliciting forecasts on their own beliefs, participants were asked to estimate others' beliefs for a subset of questions. In this report, we focus mostly on participants' own forecasts. In Appendix 3, we show some results that use participants' intersubjective forecasts.

Public participants did not engage in the persuasion component and submitted a single forecast representing their personal belief. Public forecasts were sometimes extreme or poorly calibrated, resulting in large spreads in their accuracy that distorted group comparisons. To deal with this, we winsorized public responses using the interquartile range (IQR) of XPT participants' responses for each questions (specifically, using Q25 - 3 * IQR and Q75 + 3 * IQR for the lower and upper bounds, respectively). Winsorization improves the apparent performance of the public, making it a more difficult benchmark for evaluating experts and superforecasters. Since our primary goal is to compare performance across those expert groups, this conservative correction is appropriate.

A2.1.3 Data Filtering

Participants were given default forecasts reflecting ignorance priors for each question. Forecasts that exactly matched these defaults were excluded from the analysis. We also exclude null (i.e., missing) forecasts from all summary statistics.

A2.1.4 Analysis Code

Forecasts on resolved questions were evaluated for accuracy using code in this GitHub repository. Other new analysis that is part of this report but not featured in the original XPT report is also stored in this repository.

A2.2 Accuracy Confidence Intervals

Because differences in accuracy scores across participants and groups were often small, we computed 95% bootstrap confidence intervals to better assess the robustness and significance of these differences. We used a two-stage bootstrapping approach to account for two key sources of variability: question sampling and participant sampling.

In the first stage, we resampled questions with replacement. This was critical because not all participants answered every question, and we wanted to avoid penalizing participants who happened to answer particularly difficult questions. By resampling questions, we generated distributions that better reflected the variability across the full question set.

In the second stage, we also resampled participants within each group. This allowed us to estimate uncertainty around our aggregation approach. This step was especially important for groups with smaller sample sizes, such as domain experts. In these cases, resampling participants helped assess whether observed scores were representative of typical group performance or potentially driven by outliers (e.g., a single highly accurate or inaccurate forecaster).

For each resampled dataset (defined by a specific participant and question bootstrap iteration), we computed the relevant accuracy metric. We then aggregated these across all iterations to construct bootstrap distributions for each group and used the 2.5th and 97.5th percentiles of these distributions to define the 95% confidence intervals. The pseudocode below describes the algorithm in more detail.

A2.2.1 Generating Bootstrap Dataset

- **data**: dataset with individual participant forecasts and resolved outcomes for each question.

```
None
function: get_boot_data(input: data)

       # Stage 1: Question resampling

       1. Sample questions from data with replacement -> sampled_questions

       # Stage 2: Stratified user resampling

       2. For each group, sample users with replacement -> sampled_users

       3. For each sampled user, append all their forecasts (within
       sampled_questions) -> boot_data
```

```
        4. Add bootstrap_id column to tag the dataset

(output: boot_data)
```

A2.2.2. Computing Confidence Intervals

- **n_boot:** number of bootstrap iterations.

```
None
initialize empty list boot_results

for i in (1 to n_boot):

        1. Generate one bootstrapped dataset -> boot_data = get_boot_data()

        2. Process bootstrapped dataset to get aggregate forecasts ->
processed_boot_data

        3. Compute individual- and group-level accuracy metrics

        4. Store accuracy metrics in boot_results[i]

After all iterations:
Compute 2.5% and 97.5% quantiles of boot_results for each metric -> confidence
interval bounds
```

A3.3 Naive Forecasts

In addition to the public responses, we included two **naive benchmarks** to compare to participants' accuracy across questions. These benchmarks consist of:

1. **No-change forecasts**, which simply report the most recent historical base rate available at the time of the tournament; and
2. **Extrapolated forecasts**, which project forward from the last available data using a constant growth rate derived from the most recent year-over-year change. Specifically, the forecast is calculated by applying the same proportional change observed between the two most recent data points to future years.

For some questions for which base rates were unavailable, we used a large language model (Claude) to obtain the no-change forecasts.

A3.4 Accuracy Metrics: Definitions

A3.4.1 Accuracy Score

This metric combines performance on binary and continuous questions into a single standardized measure. In accordance with the rules communicated to the tournament participants, binary questions are scored using the log score, and continuous questions are scored using the S-score rule.

1. For each participant *i* and question *q*, calculate the raw score, $S_{i,q}$, for that question:
    a. *Binary questions*: The raw score is $S_{i,q} = ln(f_{i,q})$ for questions resolving to 1, and $S_{i,q} = ln(1 - f_{i,q})$ for questions resolving to zero. All forecasts are winsorized at 0.001 and 0.999.
    b. *Continuous questions*: The raw score is given by the S-score and calculated as
    $$S_{i,q} = (-1) \times \frac{1}{5} \sum_{i=1}^{5} \alpha_i \times max(0, y_q - f_{i,q}^{(\alpha_i)}) + (1 - \alpha_i) \times max(0, f_{i,q}^{(\alpha_i)} - y_q)$$
    where $f_{i,q}^{(\alpha)}$ is the α-quantile forecast made by forecaster *i* on question *q*, and $\alpha_i \in \{0.05, 0.25, 0.5, 0.75, 0.95\}$. We multiply the score by (-1) so that higher scores indicate better accuracy.
2. Standardize the raw scores question-by-question using the median and standard deviation of the middle 80% of XPT participants:
    $$\widehat{S}_{i,q} = \frac{S_{i,q} - med_i(S_{i,q})}{std.\,dev._i(S_{i,q})}$$
3. Average the scores across questions to get a single score per participant: $\widehat{S}_i = \frac{1}{Q} \sum_{q=1}^{Q} \widehat{S}_{i,q}$.
4. Subtract the score of the median XPT participant from step (3) to have the metric expressed relative to the median XPT participant.

A3.4.2 Accuracy Score (50th Percentile Only)

Calculated identically to *Accuracy Score*. The only difference is that in step (1b), only $\alpha = 0.50$ forecasts are used (i.e., 50th percentile predictions).

A3.4.3 Standardized Absolute Forecasting Error (SAFE)

1. For each participant i and question q, calculate the absolute forecast error $\varepsilon_{i,q} = |f_{i,q} - y_q|$ where $f_{i,q}$ is the prediction, and y_q is the realized (i.e., resolved) value. For continuous questions, use the 50th percentile prediction as $f_{i,q}$, i.e., $f_{i,q}^{(0.50)}$ (for simplicity, we drop the (α) superscript).

2. Calculate the subjective standard deviation for question q as $\sigma_q = (f_q^{(0.75)} - f_q^{(0.25)})/1.349$ where $f_q^{(p)}$ is the (100 * p)-th percentile aggregate forecast of all superforecasters and experts (aggregated using trimmed-mean aggregation, trimming 5% from the top and bottom to avoid the impact of any outliers).
    a. For binary outcomes, estimate the subjective standard deviation via the standard Bernoulli formula of $\sigma_q = \sqrt{f_q(1 - f_q)}$.
3. Calculate the standardized absolute forecasting error as $SAFE_{i,q} = |f_{i,q} - y_q|/\sigma_q$.

While the SAFE metric is, to our knowledge, new to the judgemental forecasting literature, it is based on several well-established ideas in statistics and forecasting:

- **Scaled absolute forecast errors.** SAFE metric follows a key idea from the Mean Absolute Scaled Error (MASE) proposed by Hyndman and Koehler (2006), who motivate scaling to make errors comparable across questions and data sets. In particular, MASE scales the absolute forecast errors by the mean absolute error from a naive (no-change) forecast. The only difference between SAFE and MASE is that we scale absolute forecast errors by the predictive standard deviation.
- **Robust standard deviation estimate.** We estimate the predictive standard deviation as IQR / 1.349. As is well-known, when the underlying data are normally distributed, this estimator is consistent for the population standard deviation (see, for example, Wand, 1997, p. 61).
- **Standardizing by predictive standard deviation.** Dividing each forecast error by its predictive standard deviation is identical to the standardised prediction error used in textbook time series diagnostics; see Brockwell and Davis (2002, p. 165) for ARIMA and Durbin and Koopman (2012, Chapter 2.12) for state-space models.

The use of trimmed-mean aggregation (used for aggregating 25th and 75th percentile forecasts) is recommended by Armstrong (2001) and is found to improve upon simple averaging by Jose and Winkler (2008).

A3.4.4 Percentile Accuracy

1. Calculate raw accuracy scores as in step (1) of Overall Accuracy to get $S_{i,q}$ for each forecaster i and question q.
2. Calculate the percent of public participants with lower raw scores than each participant on each question:
$$100 \times \frac{count_j(S_{j,q}^{public} < S_{i,q})}{\#\ public\ respondents}.$$
3. Average this quantity over all questions for each participant to get their percentile accuracy:
$$percentile\ accuracy_i = \frac{100}{Q} \sum_{q=1}^{Q} 100 \times \frac{count(S_{i,q}^{public} < S_{i,q})}{\#\ public\ responses}.$$

A3.4.5 Cumulative Percentile Rank

This is a second-order metric based on the percentile accuracy estimates:

1. Compute the percentile accuracy for each participant (as described above), including for individual public respondents.
2. For each participant, estimate the proportion of individual public respondents with a lower percentile accuracy:

$$cumulative\ percentile\ rank_i = \frac{count_j(percentile\ accuracy_i > percentile\ accuracy_j^{public})}{\#\ public\ respondents}.$$

A3.4.5 Mean Standardized Squared Error (MSSE)

1. For each participant *i* and question *q*, compute the squared error for their 50th percentile prediction:

$$SE_{i,q} = (f_{i,q} - y_q)^2.$$

2. For the standardization, estimate the median $\overline{SE_q}$ and standard deviation $\hat{\sigma}_{SEq}$ of the middle 80% of squared errors for each question, as with *Accuracy Score*. For this step, we use only the squared errors for XPT participants (excluding the public and other benchmark forecasts).

3. Calculate the standardized squared error for each participant and question q,

$$SE_{i,q}^{standard} = \frac{SE_{i,q} - \overline{SE_q}}{\hat{\sigma}_{SEq}}.$$

4. Take the mean over Q questions for each participant to get the mean standardized squared error.

$$MSSE_i = \frac{1}{Q}\sum_{q=1}^{Q} SE_{i,q}^{standard}.$$

A3.5 Distribution Estimation

Superforecasters, domain experts, and non-domain experts provided forecasts for five quantiles (5th, 25th, 50th, 75th, and 95th percentiles) for each question. For each group, we calculated the median forecast at each quantile and used these five values as points to construct an empirical cumulative distribution function with linear interpolation between the median quantiles values. We then generated pseudo-random samples from each empirical CDF using inverse transform sampling and fitted these samples to the Johnson's S_U-distribution to obtain smooth probability density functions. This Johnson's S_U distribution family was chosen for its flexibility modeling both skewed and symmetric distributions.

Appendix 3: Additional Empirical Results

A3.1 Sample Size and Composition

Group		Subject Area				
		All	AI	Bio	Climate	Nuclear
Number of resolved subquestions		38	13	18	4	3
Number of total forecasts	Superforecasters	1,267	414	619	130	104
	Domain experts	183	72	85	18	8
	Non-domain experts	354	78	203	38	35
	X-risk generalists	133	38	73	8	14
	XPT participants (all above)	*1,937*	*602*	*980*	*194*	*161*
	Public	13,785	5,119	6,128	1,219	1,319
	Total all groups	**15,722**	**5,721**	**7,108**	**1,413**	**1,480**
Number of individual forecasters	Superforecasters		87	87	87	87
	Domain experts		29	13	10	14
	Non-domain experts		30	46	49	45
	X-risk generalists		14	14	14	14
	Public		516	516	516	516

Table A3.1: *Summary of sample sizes. We count quantile predictions for the same subquestion as a single prediction for the purposes of this table. Note that these numbers differ slightly from the headline sample size numbers cited in the introduction since not all participants predicted the resolved subquestions.*

A3.2 Alternative Accuracy Metrics (Individual-Level)

	Accuracy Score	Accuracy Score (50th Percentile Only)	SAFE	Mean Standardized Squared Error	Percentile Accuracy (%)	Cumulative Percentile Rank (%)
Superforecasters	0.139 (-0.095, 0.373)	-0.027 (-0.26, 0.206)	1.02 (0.767, 1.272)	0.785 (0.481, 1.089)	55.321 (48.592, 62.051)	96.117 (86.499, 100)
Domain Experts	0.153 (-0.466, 0.772)	0.214 (-0.286, 0.713)	0.957 (0.485, 1.429)	0.539 (-0.007, 1.086)	58.034 (44.456, 71.611)	98.058 (74.66, 100)
Non-domain Experts	0.101 (-0.367, 0.569)	0.087 (-0.374, 0.547)	0.732 (0.363, 1.101)	0.659 (0.074, 1.243)	52.631 (42.845, 62.417)	91.65 (71.65, 100)
X-risk Generalists	-0.034 (-0.791, 0.722)	0.052 (-0.837, 0.94)	0.795 (0.333, 1.256)	0.784 (-0.548, 2.117)	54.51 (42.201, 66.82)	93.592 (74.027, 100)
XPT Participants (all above)	0	0	0.904 (0.675, 1.133)	0.75 (0.52, 0.98)	54.624 (48.48, 60.767)	94.854 (84.949, 100)
Public		-1.823 (-4.202, 0.555)	2.117 (1.567, 2.668)	6.884 (-4.124, 17.891)	40.387 (37.833, 42.94)	49.903 (49.612, 50.194)
Naive (no change)		0.031 (-0.737, 0.798)	1.04 (0.589, 1.491)	1.568 (-0.164, 3.299)	54.863 (46.391, 63.335)	95.34 (81.262, 100)
Naive (extrapolated)		-0.093 (-1.196, 1.009)	0.935 (0.556, 1.314)	2.947 (-1.627, 7.522)	58.955 (51.145, 66.764)	98.447 (91.066, 100)

Table A3.2: *Comparison of median individual participant's performance on the XPT near-term questions as scored by six metrics. 95% bootstrap confidence intervals are shown in the parentheses.*

A3.3 Group-Level Accuracy

	Accuracy Score	Accuracy Score (50th Percentile Only)	SAFE	Mean Standardized Squared Error	Percentile Accuracy (%)	Cumulative Percentile Rank (%)
Superforecasters	0.992 (0.709, 1.276)	0.756 (0.474, 1.037)	0.606 (0.386, 0.826)	-0.091 (-0.228, 0.046)	62.391 (54.344, 70.438)	99.223 (94.364, 100)
Domain Experts	0.485 (-0.37, 1.34)	0.678 (-0.143, 1.499)	0.633 (0.074, 1.193)	0.198 (-2.526, 2.922)	62.635 (52.965, 72.305)	99.417 (87.087, 100)
Non-domain Experts	0.968 (0.425, 1.51)	0.756 (0.271, 1.242)	0.607 (0.216, 0.999)	-0.094 (-0.612, 0.424)	62.064 (52.468, 71.659)	99.223 (90.869, 100)
X-risk Generalists (Aggregate)	0.546 (-0.187, 1.278)	0.545 (-0.159, 1.248)	0.695 (0.301, 1.088)	0.146 (-1.333, 1.626)	62.581 (53.629, 71.534)	99.417 (92.813, 100)
XPT Participants (all above)	0.971 (0.723, 1.219)	0.777 (0.533, 1.021)	0.595 (0.382, 0.808)	-0.118 (-0.2, -0.036)	62.112 (54.404, 69.821)	99.223 (95.049, 100)
Public		-0.505 (-1.726, 0.716)	1.435 (0.729, 2.141)	3.889 (-1.5, 9.277)	50.952 (44.701, 57.202)	87.961 (73.204, 100)
Naive (no change)		0.031 (-0.737, 0.798)	1.04 (0.589, 1.491)	1.568 (-0.164, 3.299)	54.863 (46.391, 63.335)	95.34 (81.262, 100)
Naive (extrapolated)		-0.093 (-1.196, 1.009)	0.935 (0.556, 1.314)	2.947 (-1.627, 7.522)	58.955 (51.145, 66.764)	98.447 (91.066, 100)

Table A3.3: *Comparison of group-level performance on the XPT near-term questions as scored by six metrics. 95% bootstrap confidence intervals are shown in the parentheses.*

In line with previous research, aggregated group-level forecasts outperformed those made by individual participants. For example, the SAFE scores for aggregate forecasts from XPT participants range from 0.60-0.70, compared to 0.80-1.02 for the median individual. Group-level performance also consistently beat benchmarks, including both naive forecasts and the public. Notably, the aggregate public forecasts show significant improvement from individual public participants, though they still represent the weakest performing group.

A3.4 AI Questions: Individual- and Group-Level Accuracy

When focusing only on AI-related questions (Table A3.4), Superforecasters outperform domain and non-domain experts across all metrics, though the differences are not statistically significant. Interestingly, x-risk generalists perform on par with—or even slightly better than—Superforecasters on some metrics. Public respondents show worse performance on AI questions, and they continue to underperform compared to all other groups and even naive benchmarks.

At the group-forecast level (Table A3.5), aggregation significantly improves the Public's performance. Notably, domain experts outperform Superforecasters when forecasts are aggregated, reversing the individual-level ranking—though again, the difference is not statistically significant.

	Accuracy Score	Accuracy Score (50th Percentile Only)	SAFE	Mean Standardized Squared Error	Percentile Accuracy (%)	Cumulative Percentile Rank (%)
Superforecasters	0.063 (-0.195, 0.32)	0.128 (-0.116, 0.371)	1.011 (0.589, 1.433)	0.257 (-0.014, 0.528)	68.693 (56.572, 80.813)	96.893 (85.126, 100)
Domain Expert	-0.069 (-0.898, 0.76)	-0.06 (-0.875, 0.754)	1.064 (0.326, 1.801)	0.64 (-0.308, 1.588)	63.243 (44.273, 82.214)	91.068 (62.129, 100)
Non-domain Expert	-0.13 (-1.358, 1.099)	-0.067 (-1.262, 1.128)	1.352 (0.638, 2.067)	0.524 (-0.866, 1.913)	61.772 (42.112, 81.431)	89.32 (59.318, 100)
X-risk Generalist	0.174 (-0.741, 1.089)	0.122 (-0.909, 1.153)	1.037 (0.299, 1.775)	0.204 (-0.94, 1.347)	69.995 (51.319, 88.671)	96.796 (75.922, 100)
XPT Participants (all above)	0	0	1.064 (0.711, 1.418)	0.43 (0.21, 0.65)	66.146 (54.755, 77.537)	95.243 (83.734, 100)
Public		-2.688 (-4.096, -1.28)	2.699 (1.74, 3.658)	8.313 (2.886, 13.74)	45.67 (43.554, 47.786)	49.903 (49.709, 50.097)
Naive (no change)		-1.075 (-2.86, 0.71)	1.886 (0.82, 2.951)	3.776 (-0.804, 8.356)	62.591 (44.359, 80.824)	90.485 (65.619, 100)
Naive (extrapolated)		-0.296 (-1.713, 1.121)	1.354 (0.585, 2.123)	1.754 (-1.088, 4.597)	75.891 (64.642, 87.14)	99.029 (92.522, 100)

Table A3.4: *Comparison of median individual participant's performance on the AI-specific near-term questions as scored by six metrics. 95% bootstrap confidence intervals are shown in the parentheses.*

	Accuracy Score	Accuracy Score (50th Percentile Only)	SAFE	Mean Standardized Squared Error	Percentile Accuracy (%)	Cumulative Percentile Rank (%)
Superforecasters	0.564 (0.219, 0.908)	0.414 (0.055, 0.772)	0.876 (0.46, 1.291)	0.035 (-0.187, 0.258)	71.961 (57.602, 86.32)	97.864 (85.714, 100)
Domain Experts	0.74 (-0.269, 1.749)	0.885 (-0.34, 2.111)	0.683 (-0.414, 1.78)	-0.302 (-3.721, 3.117)	75.475 (60.549, 90.402)	99.029 (82.91, 100)
Non-domain Experts	0.725 (-0.483, 1.933)	0.611 (-0.378, 1.601)	0.798 (0.026, 1.569)	-0.131 (-1.75, 1.488)	73.786 (55.688, 91.884)	98.447 (78.723, 100)
X-risk Generalists	0.33 (-0.589, 1.248)	0.403 (-0.845, 1.651)	0.878 (0.202, 1.554)	0.097 (-3.407, 3.601)	70.747 (54.796, 86.699)	97.67 (81.839, 100)
XPT Participants (all above)	0.677 (0.405, 0.95)	0.604 (0.292, 0.916)	0.781 (0.387, 1.174)	-0.122 (-0.264, 0.02)	72.54 (59.043, 86.037)	98.252 (87.862, 100)
Public		-1.181 (-2.528, 0.166)	2.162 (1.121, 3.204)	2.477 (-0.423, 5.377)	53.702 (45.831, 61.574)	74.563 (59.389, 89.737)
Naive (no change)		-1.075 (-2.86, 0.71)	1.886 (0.82, 2.951)	3.776 (-0.804, 8.356)	62.591 (44.359, 80.824)	90.485 (65.619, 100)
Naive (extrapolated)		-0.296 (-1.713, 1.121)	1.354 (0.585, 2.123)	1.754 (-1.088, 4.597)	75.891 (64.642, 87.14)	99.029 (92.522, 100)

Table A3.5: *Comparison of group-level performance on the AI-specific near-term questions as scored by six metrics. 95% bootstrap confidence intervals are shown in the parentheses.*

A3.5 Near-Term Accuracy versus Intersubjective Accuracy

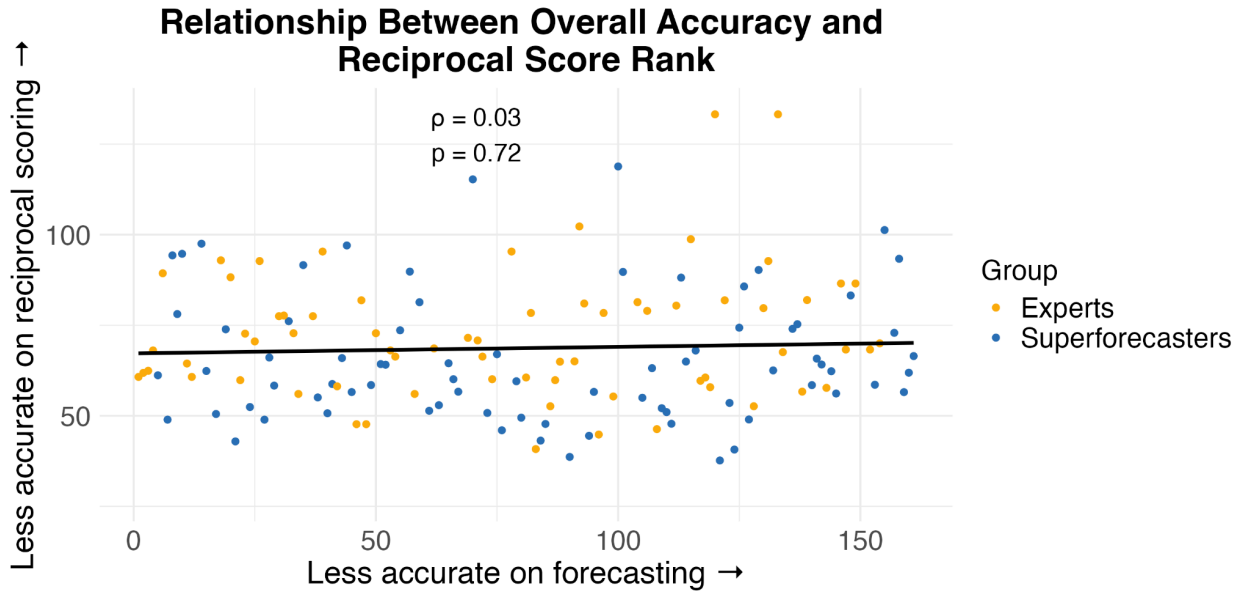


Figure A3.1: *Comparison of overall accuracy on the XPT near-term questions versus intersubjective accuracy (measured by the reciprocal score).*

As discussed in Section 2, we found no statistically significant correlation between overall accuracy on near-term questions and intersubjective accuracy on existential risk questions (as measured by the reciprocal score). Figure A3.1 shows the relationship between the average rank of XPT forecasters in terms of near-term overall accuracy scoring on the *x*-axis and the average rank of XPT forecasters in terms of reciprocal scoring on the ten catastrophic and extinction risk questions asked during the XPT. To be included in the figure above, forecasters were required to have submitted answers to at least three resolved forecasting questions.

A3.6 Public Long-Term Risk Forecasts

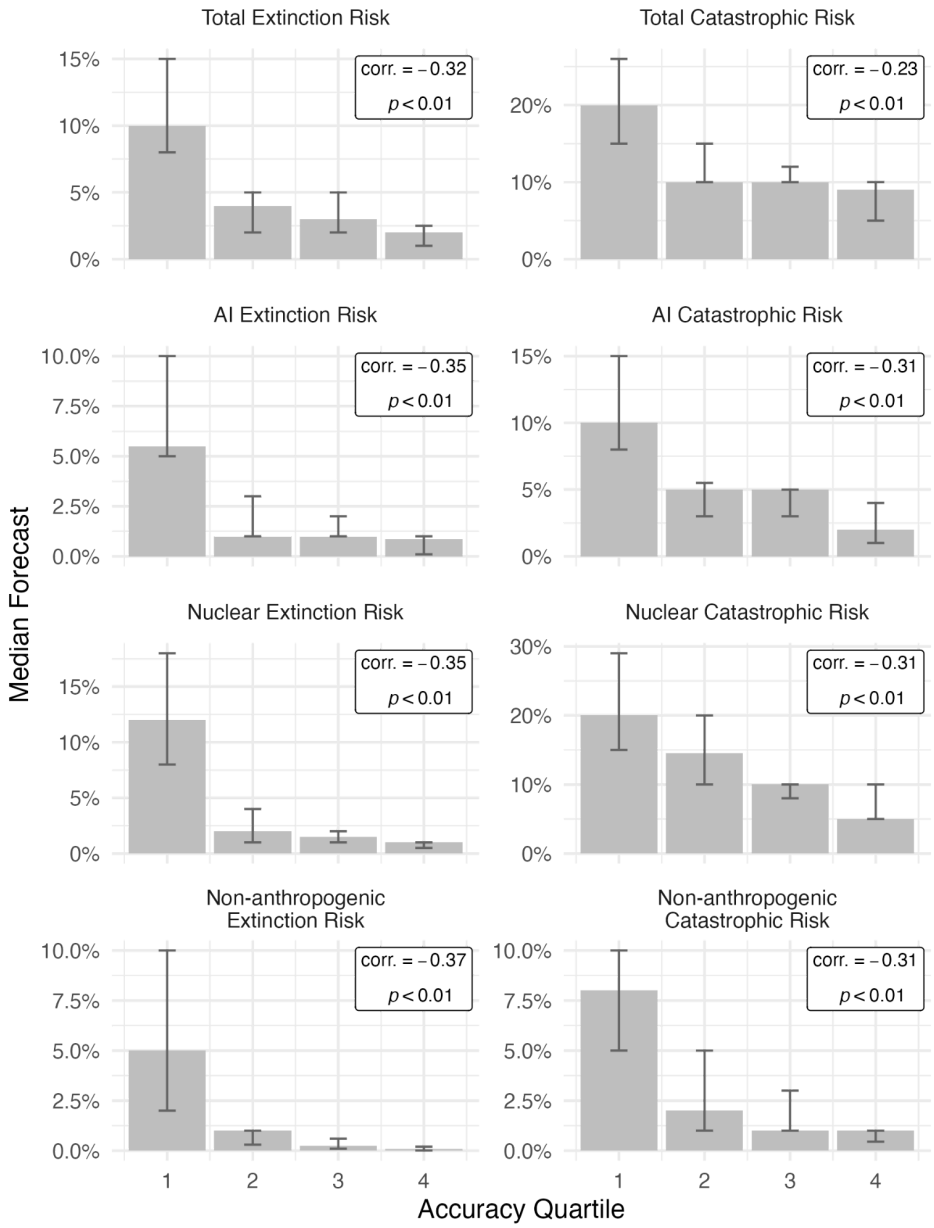


Figure A3.2: *As in Figure 4.1, but for the public participants. The public's views on existential risks by 2100. "Catastrophic risk" is defined as the probability of 10% or more of humans dying within a 5-year period (except for pathogen risks, which use a 1% threshold). "Extinction risk" is defined as the probability of human extinction or a reduction of the global population below 5,000. Participants are divided into quartiles based on their near-term accuracy, from least (1) to most (4) accurate. Error bars represent 95% bootstrap confidence intervals for the median risk forecast within each quartile. Labels show the Spearman rank correlation between individual-level accuracy and long-term risk forecasts as well as the corresponding p-value.*

Appendix 4: Forecaster Calibration

In this section, we analyze XPT participant calibration. Calibration is a key property of high-quality forecasts. For well-calibrated forecasters, their expressed uncertainty correctly captures the true uncertainty in the world.

For this analysis, we evaluate calibration by measuring *prediction interval coverage*. This approach is appropriate for our dataset for several reasons. First, many of our questions involve continuous outcomes rather than binary events. Second, our dataset contains many low-probability questions for which multiple percentile predictions may be identical (e.g., P5 = P25 = 0). Finally, prediction interval coverage provides an intuitive metric that directly addresses whether forecasters correctly estimate the uncertainty in their predictions.

We examine two key interval coverages:

- **50% prediction interval:** The range between the 25th and 75th percentile predictions.
- **90% prediction interval:** The range between the 5th and 95th percentile predictions.

For perfectly calibrated forecasters, we would expect 50% of actual outcomes to fall within their 50% prediction intervals, and 90% of outcomes to fall within their 90% prediction intervals. If, for example, the 50% prediction interval is too narrow (i.e., forecasters are *overconfident*), then we expect less than 50% of the realized outcomes to fall within the interval, and vice versa when the prediction interval is too wide (i.e., forecasters are *underconfident*).

Group	50% Prediction Interval Coverage	90% Prediction Interval Coverage
Domain Expert	52.82 (41.77–63.86)	77.46 (65.12–89.81)
Non-domain Expert	48.55 (37.39–59.72)	75.72 (69.15–82.30)
Superforecaster	43.02 (34.77–51.27)	68.02 (60.55–75.49)
X-risk Generalist	59.57 (47.98–71.17)	84.04 (74.74–93.34)

Table A4.1: *Forecaster prediction interval coverage (individual-level forecasts). 95% confidence intervals shown in the parentheses; standard errors clustered at the question level.*

Table A4.1 provides evidence on prediction interval coverage for individual forecasts. We observe the following key insights:

- **Good mid-range calibration.** For 50% prediction intervals, we cannot reject the null hypothesis of calibrated forecasts for any group, indicating generally good calibration in the middle of forecasters' distributions;
- **Systematic overconfidence at extremes.** For 90% intervals, most groups showed statistically significant overconfidence: Their intervals contained the actual outcome less often than they should.

This pattern of overconfidence at the extremes is consistent with the moderately high SAFE metrics discussed in the main report.

Group	50% Prediction Interval Coverage	90% Prediction Interval Coverage
Domain Expert	64.29 (46.21–82.36)	78.57 (63.09–94.05)
Non-domain Expert	60.71 (42.29–79.14)	89.29 (77.62–100.95)
Superforecaster	57.14 (38.48–75.81)	85.71 (72.51–98.91)
X-risk Generalist	60.71 (42.29–79.14)	92.86 (83.14–102.57)

Table A4.2: *Forecaster prediction interval coverage (group-level forecasts). 95% confidence intervals shown in the parentheses; heteroskedasticity-robust standard errors*

In Table A4.2, we can see the same exercise performed for group-level forecasts (median aggregation). The following key results emerge:

- **Improved overall calibration.** Group forecasts showed better calibration than individual forecasts, particularly at the 90% level. We cannot reject the null hypothesis of calibrated forecasts for any group and prediction interval;
- **Mild underconfidence in mid-range.** For 50% intervals, all groups were somewhat underconfident: Their intervals contained more outcomes than expected. However, this difference is not statistically significant;
- **Wider confidence intervals.** The confidence intervals for group-level analysis are broader due to the smaller sample size, reducing our ability to make definitive statistical claims.

Overall, we conclude that XPT forecasters are well-calibrated at the group level. At the individual level, there was notable overconfidence in tail predictions (90% prediction intervals).

Appendix 5: Distribution of Forecasts by Question

In this section we provide estimated probability density functions (PDFs) of the aggregate forecasts for each group and question. The *implicit percentile* refers to the percentile assigned to the resolution value under the fitted PDF.

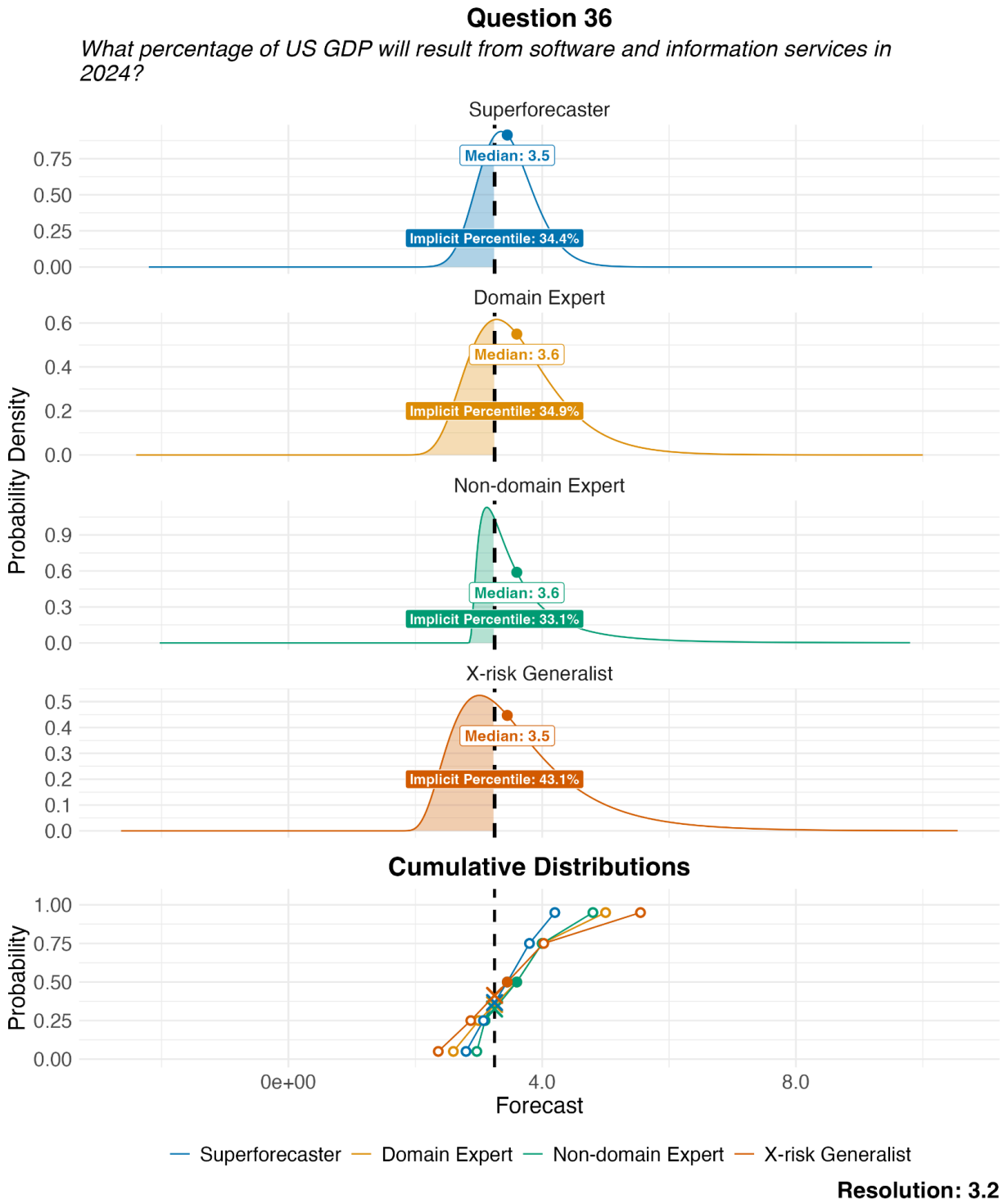
We do not show distributions for the following biorisk-related questions:

- **Q15:** How many times will a non-state actor use biological weapons that involve a contagious agent be the cause death for at least 1,000 people by the end of 2024?
- **Q16:** How many times will a state actor use biological weapons that involve a contagious agent be the cause death for at least 1,000 people by the end of 2024?
- **Q17:** How many times will a non-state actor use biological weapons that involve a contagious agent to kill at least 100,000 people by the end of 2024?
- **Q18:** How many times will a state actor use biological weapons that involve a contagious agent to kill at least 100,000 people by the end of 2024?
- **Q19:** What will be the expected number of events in which contagious biological agents that have escaped from labs be the cause of death for at least 1,000 people by the end of 2024?
- **Q22:** How many times will the WHO declare a new Public Health Emergency of International Concern (PHEIC) for a disease that will be the cause of death of at least 10,000 people by the end of 2024?
- **Q23:** What will be the expected number of events in which country leaders are assassinated by a biological weapon involving a contagious agent by the end of 2024?

For all of these questions, group medians across percentiles (5th, 25th, 50th, 75th, and 95th), were often uniformly zero. In some cases, only the 95th percentile was non-zero. These extremely flat forecast patterns made it difficult for the algorithm to fit meaningful distributions, so we do not show the PDFs for these questions.

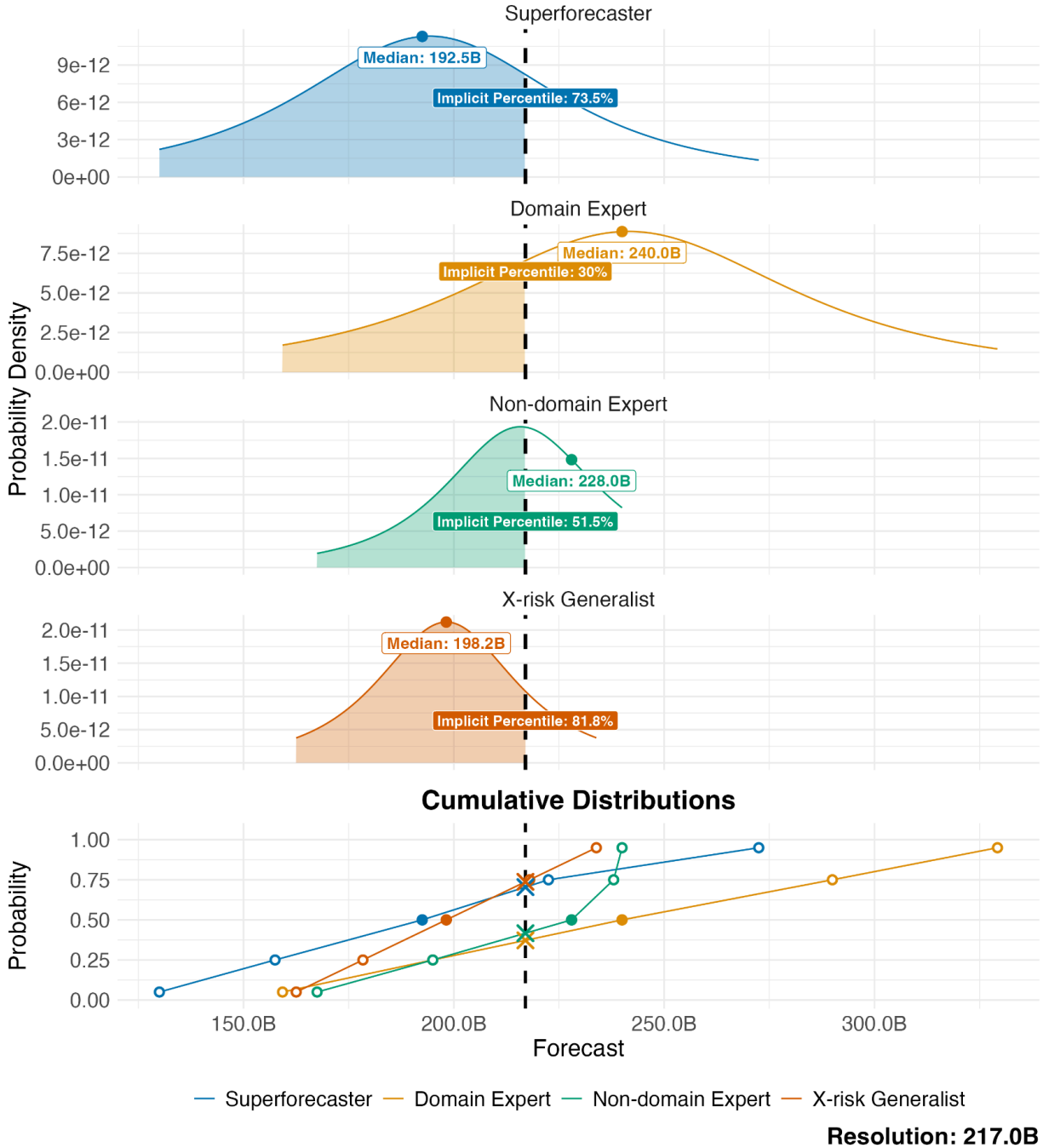
The same issue arose for **Q33:** *How many countries will be estimated to have at least one nuclear warhead by the end of 2024?* Several groups gave identical values (9 countries) across all quantiles, resulting in not enough variation to fit a distribution.

A5.1 Distributions for AI-Related Questions



Question 37

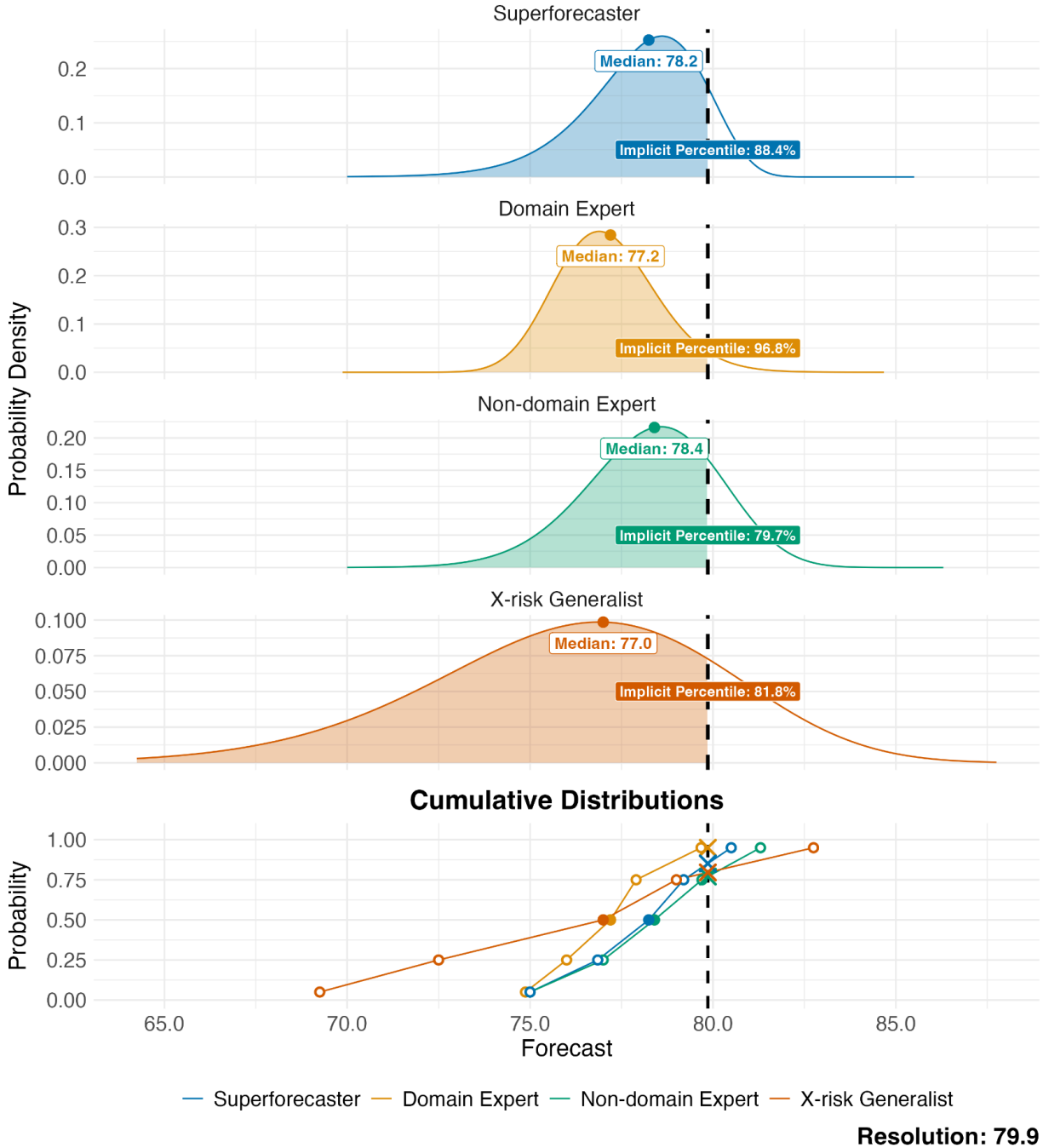
How much money will be spent on research and development by US companies in the 'Information' and 'Computer systems design' industries in 2024?



Resolution: 217.0B

Question 38

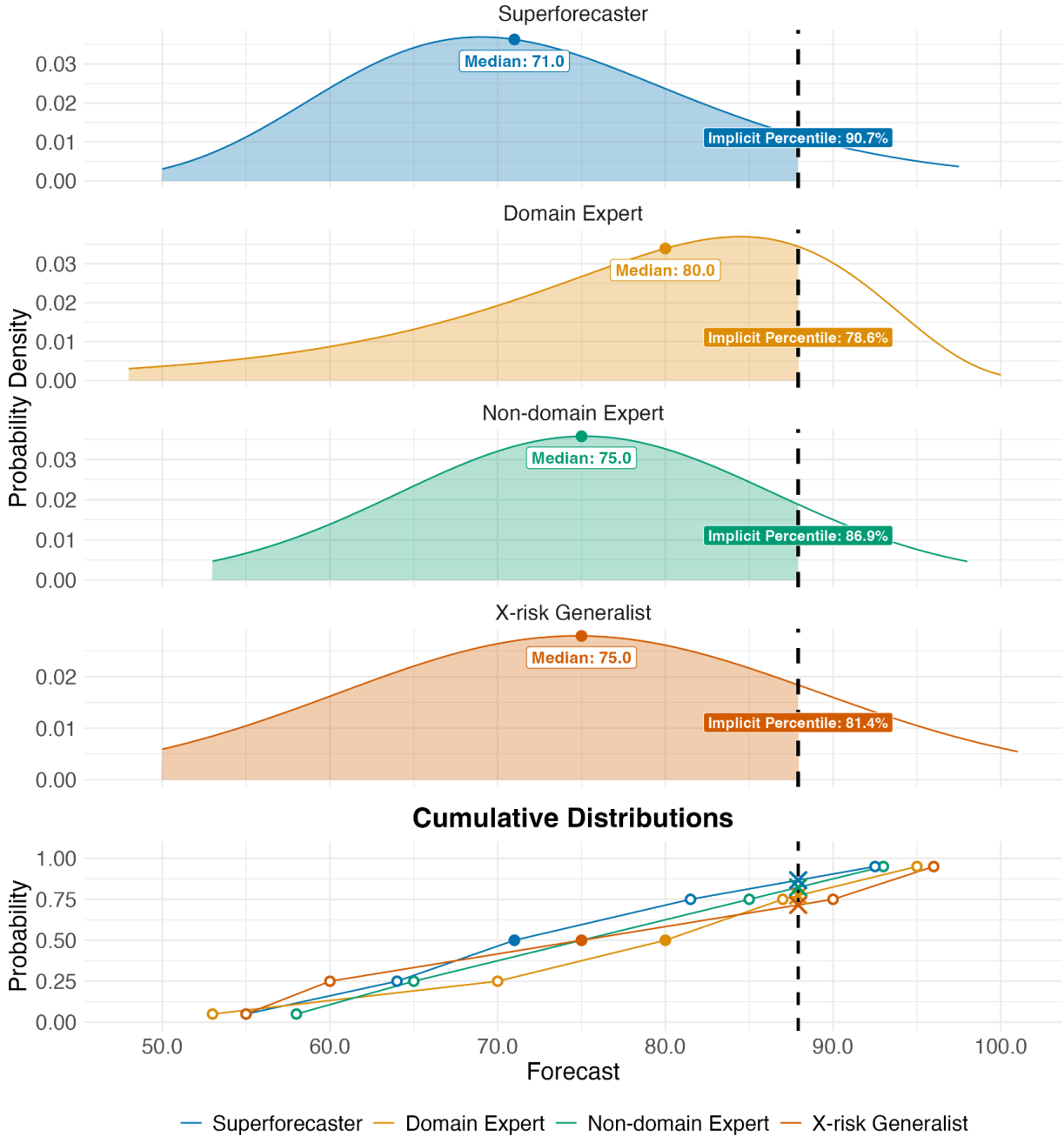
What will be the labor force participation rate in OECD countries in the year 2024?



Resolution: 79.9

Question 39

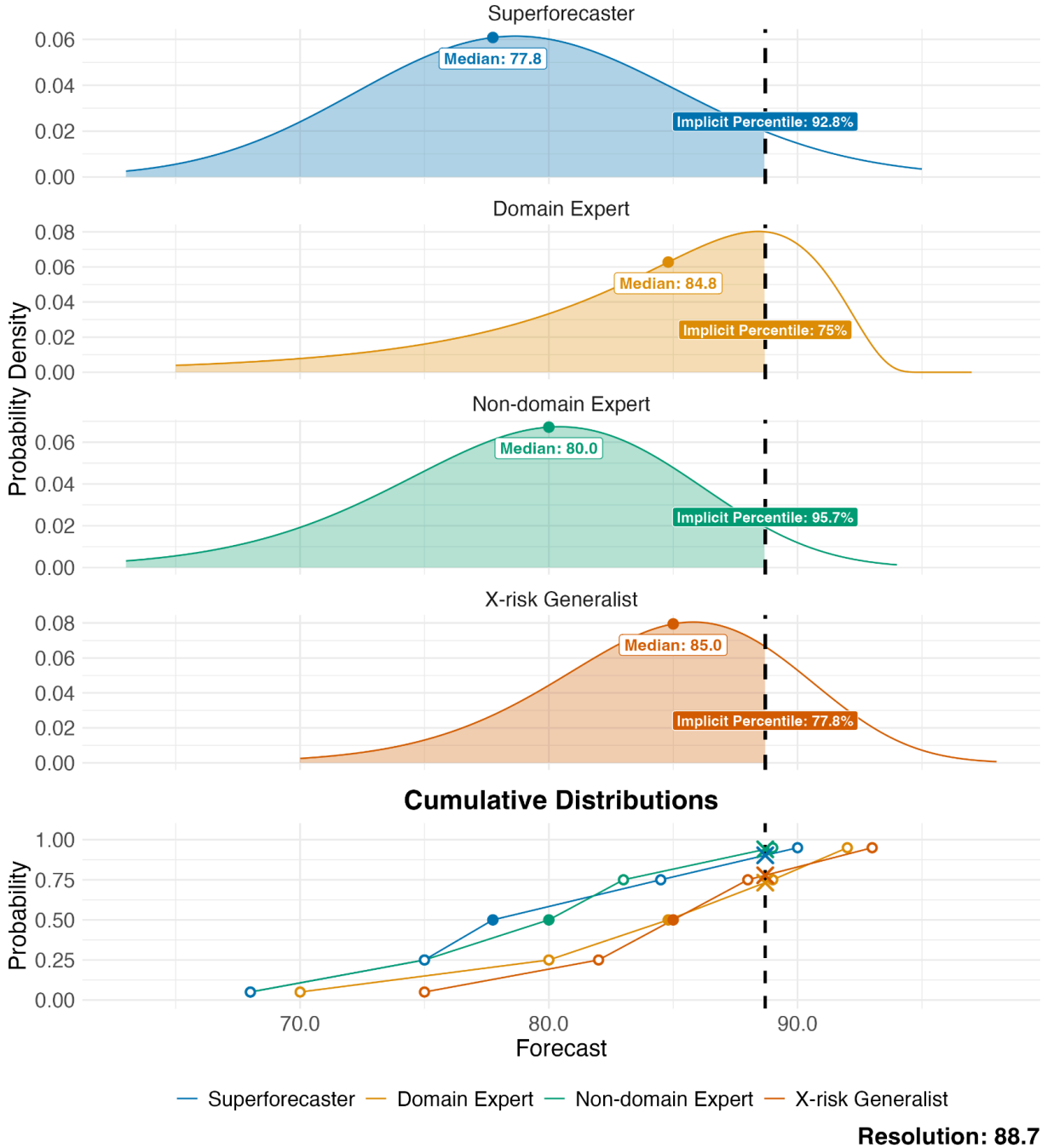
What will be the state-of-the-art accuracy of a machine-learning model on the MATH Dataset by June 30, 2024?



Resolution: 87.9

Question 40

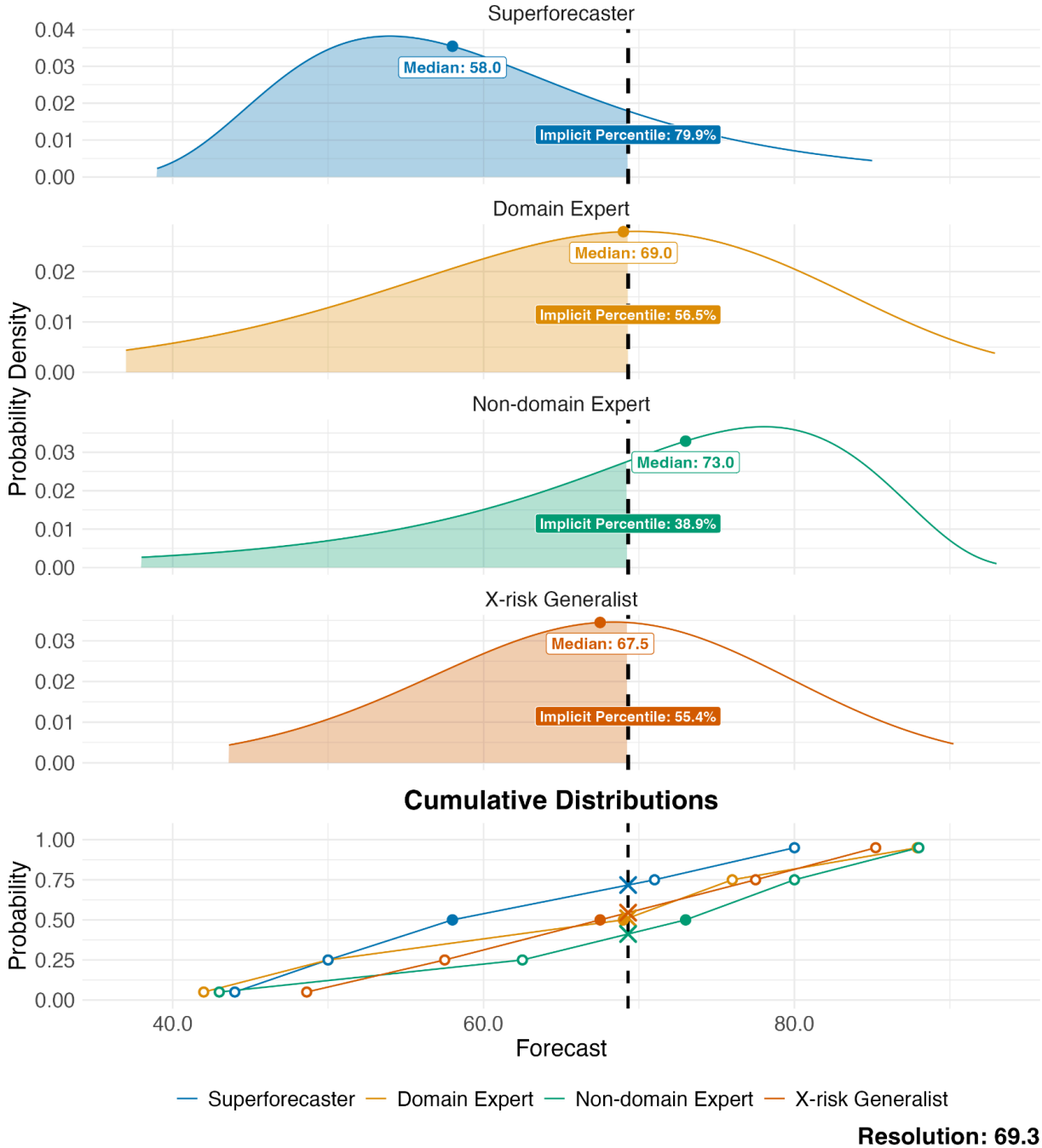
What will be the state-of-the-art few-shot or transfer accuracy on the Massive Multitask Language Understanding dataset by June 30, 2024?



Resolution: 88.7

Question 41

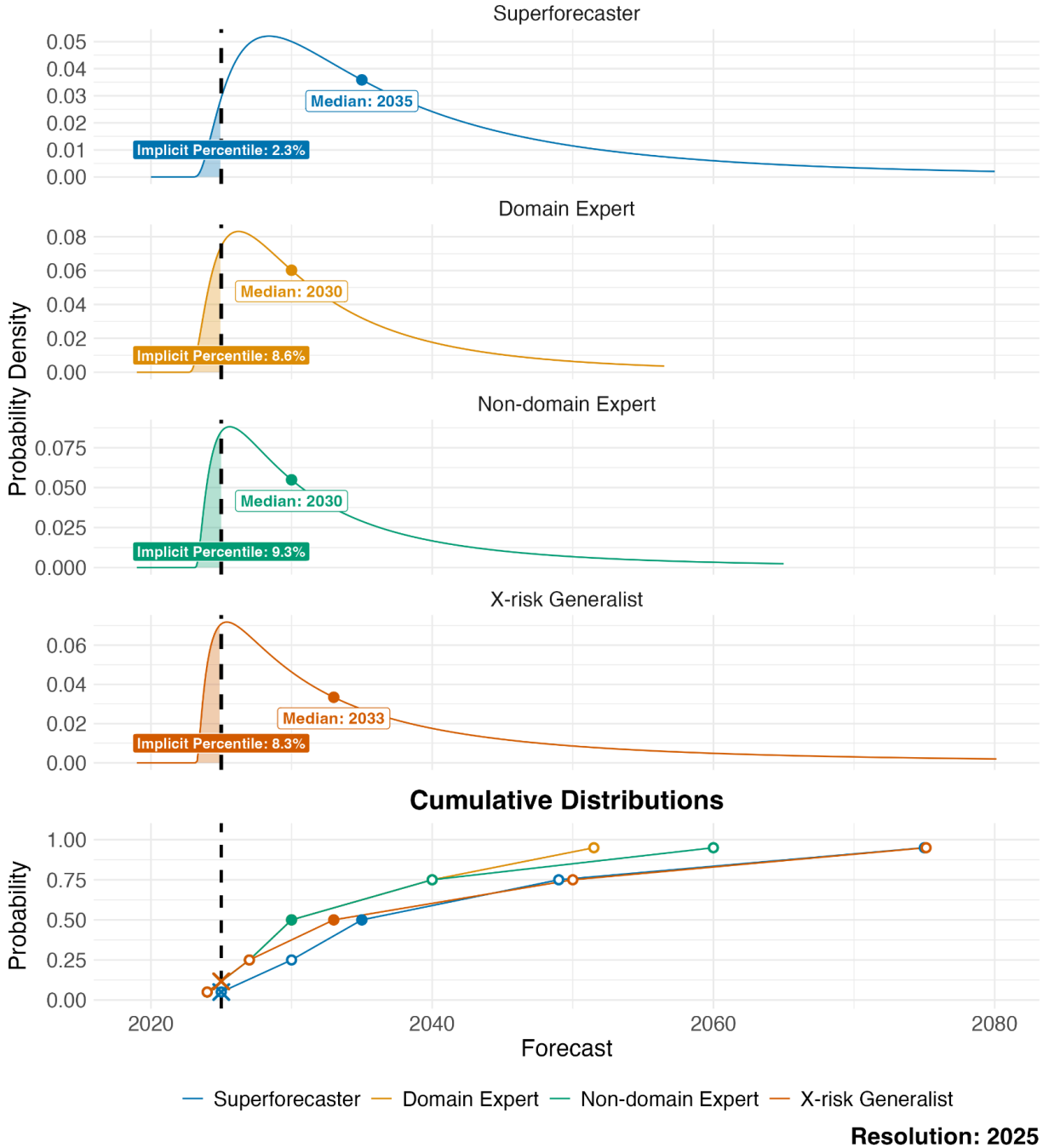
What will be the best SAT-style score with a machine learning model on the hard subset of the QuALITY dataset by June 30, 2024?



Resolution: 69.3

Question 42

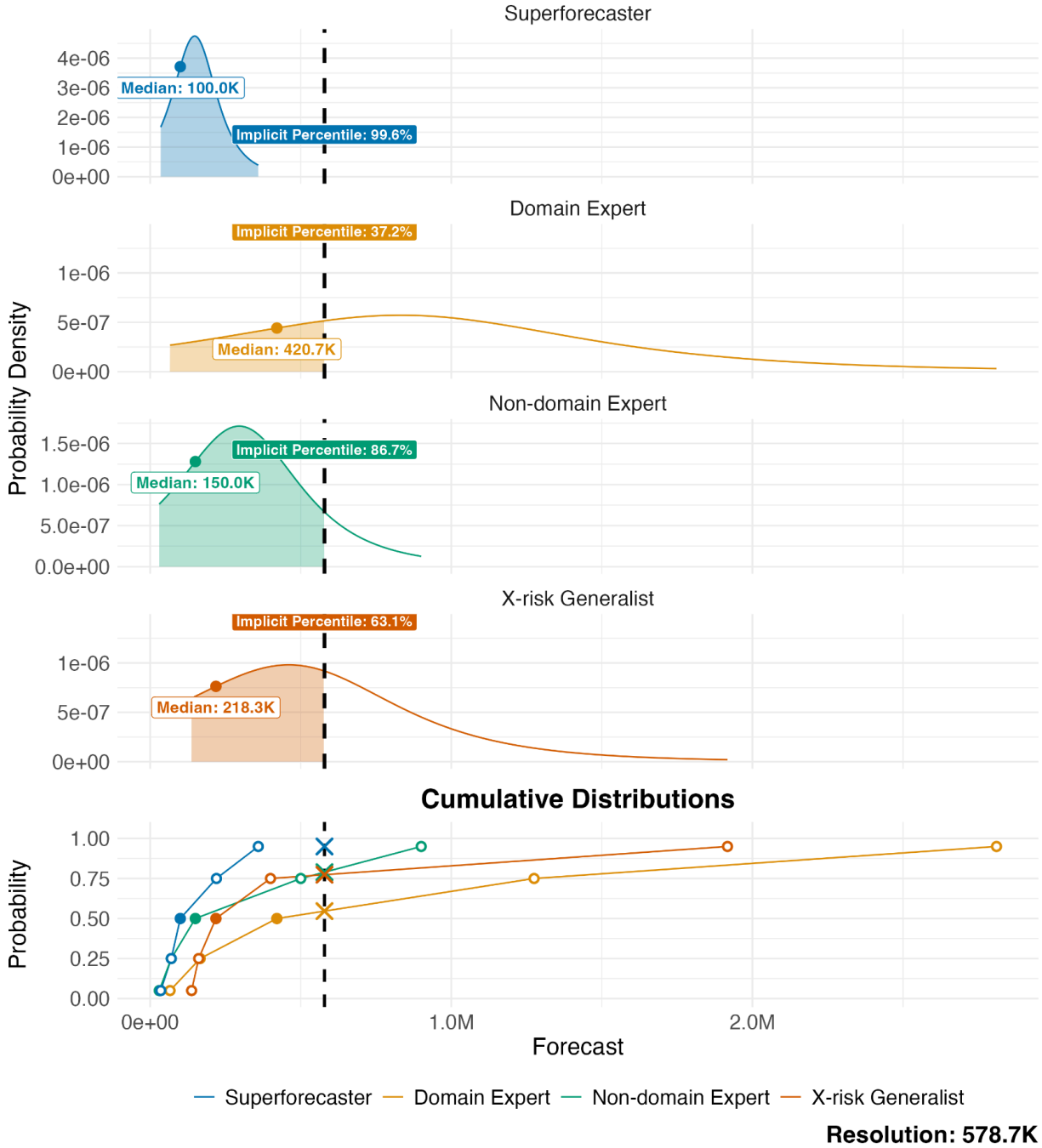
By what year will an AI win a Gold Medal in the International Mathematical Olympiad (IMO)?



Resolution: 2025

Question 45

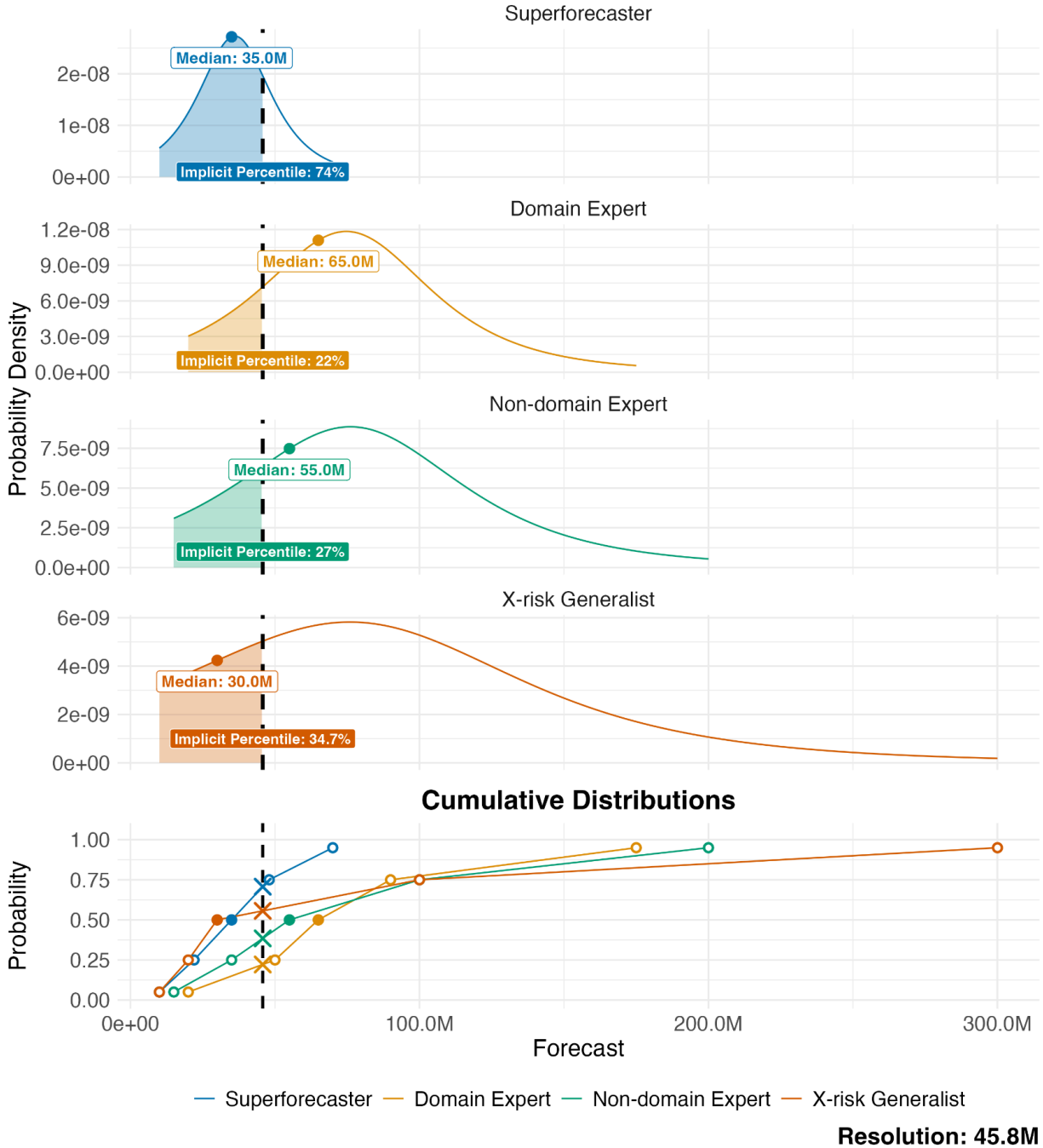
What will be the maximum compute (measured in petaFLOPS-days) used for training in an AI experiment by the end of 2024?



Resolution: 578.7K

Question 46

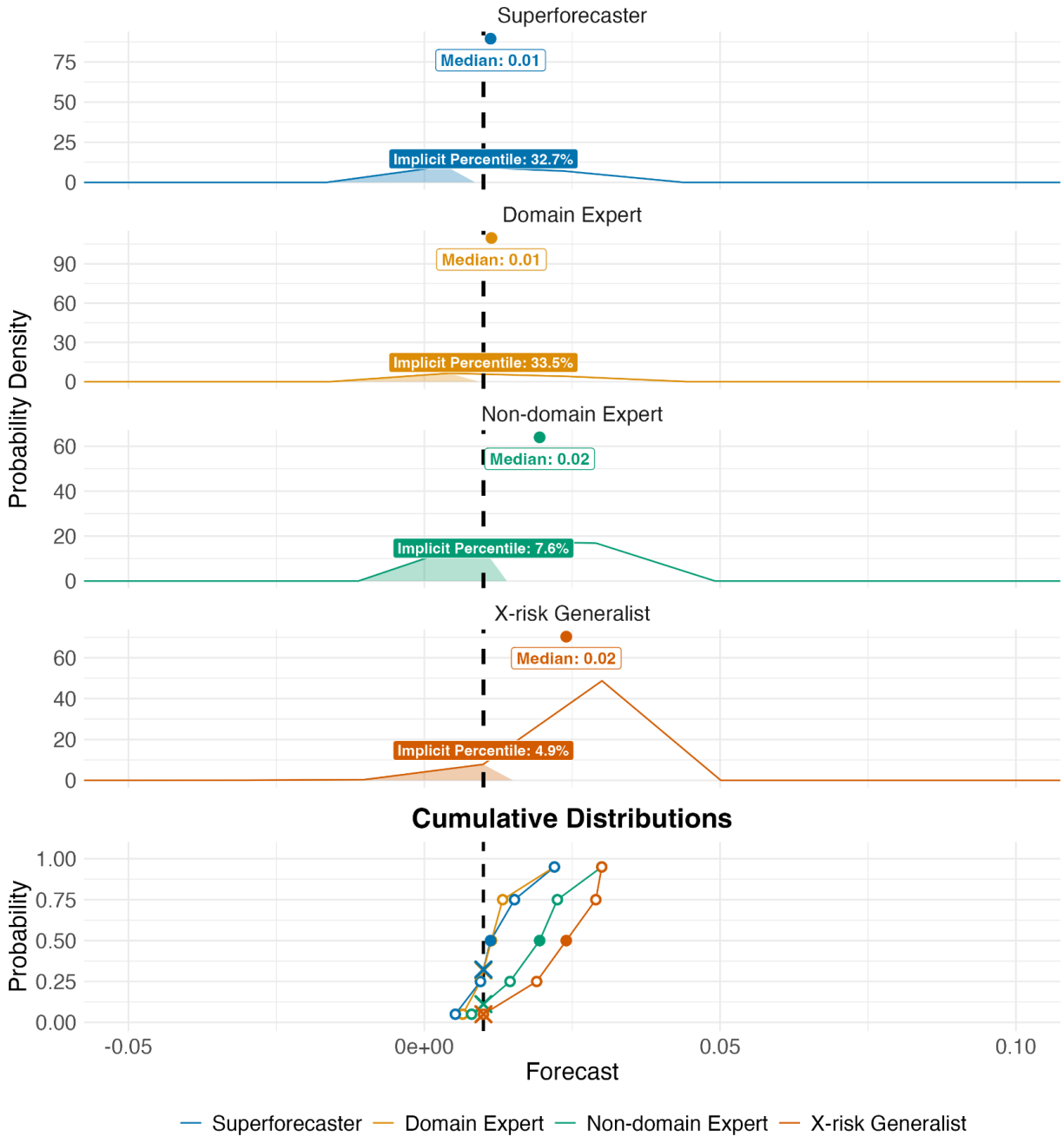
How much will be spent on compute in the largest AI experiment by the end of 2024?



Resolution: 45.8M

Question 47

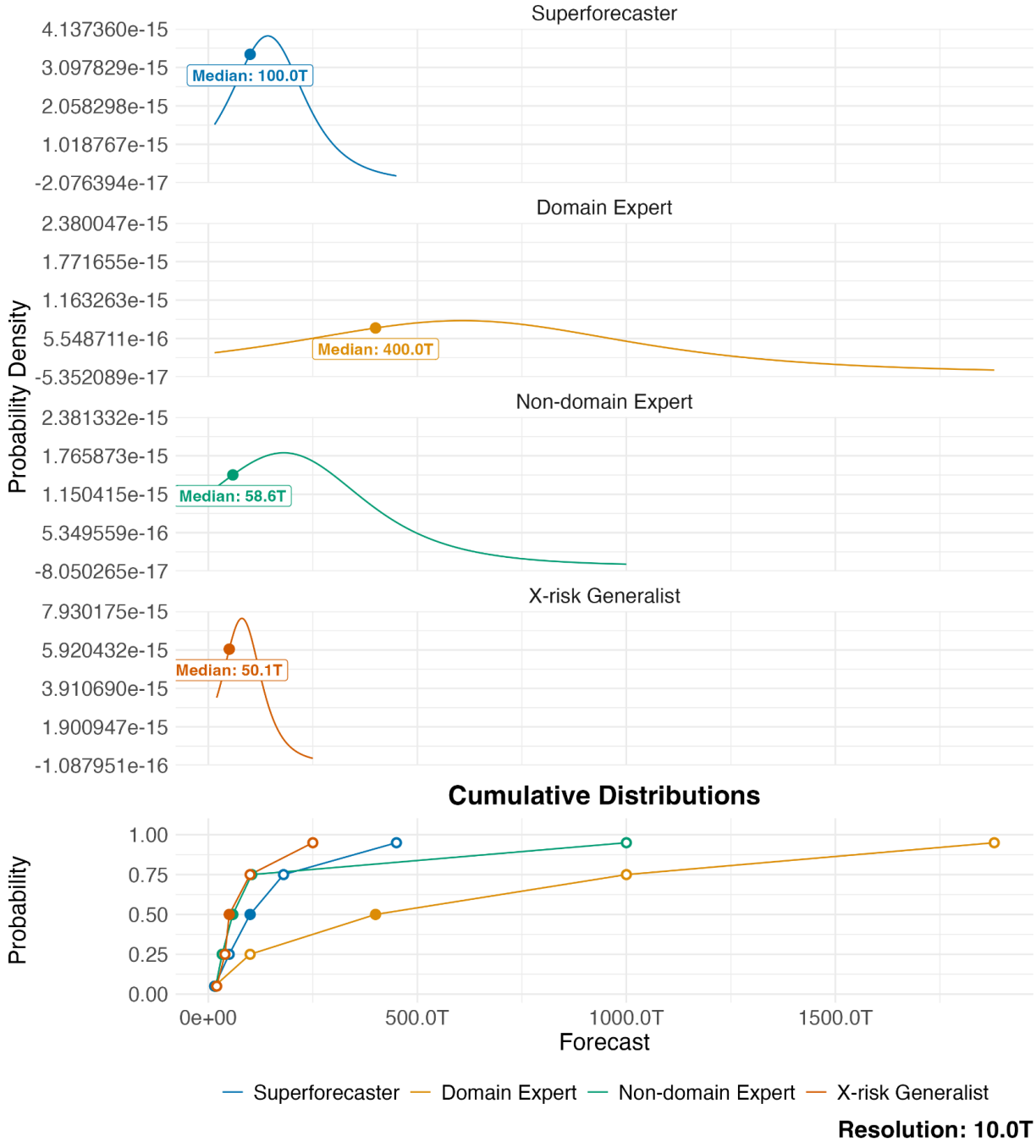
What will be the lowest price, in 2021 US dollars, of 1 GFLOPS with a widely-used processor by the end of 2024?



Resolution: 0.01

Question 49

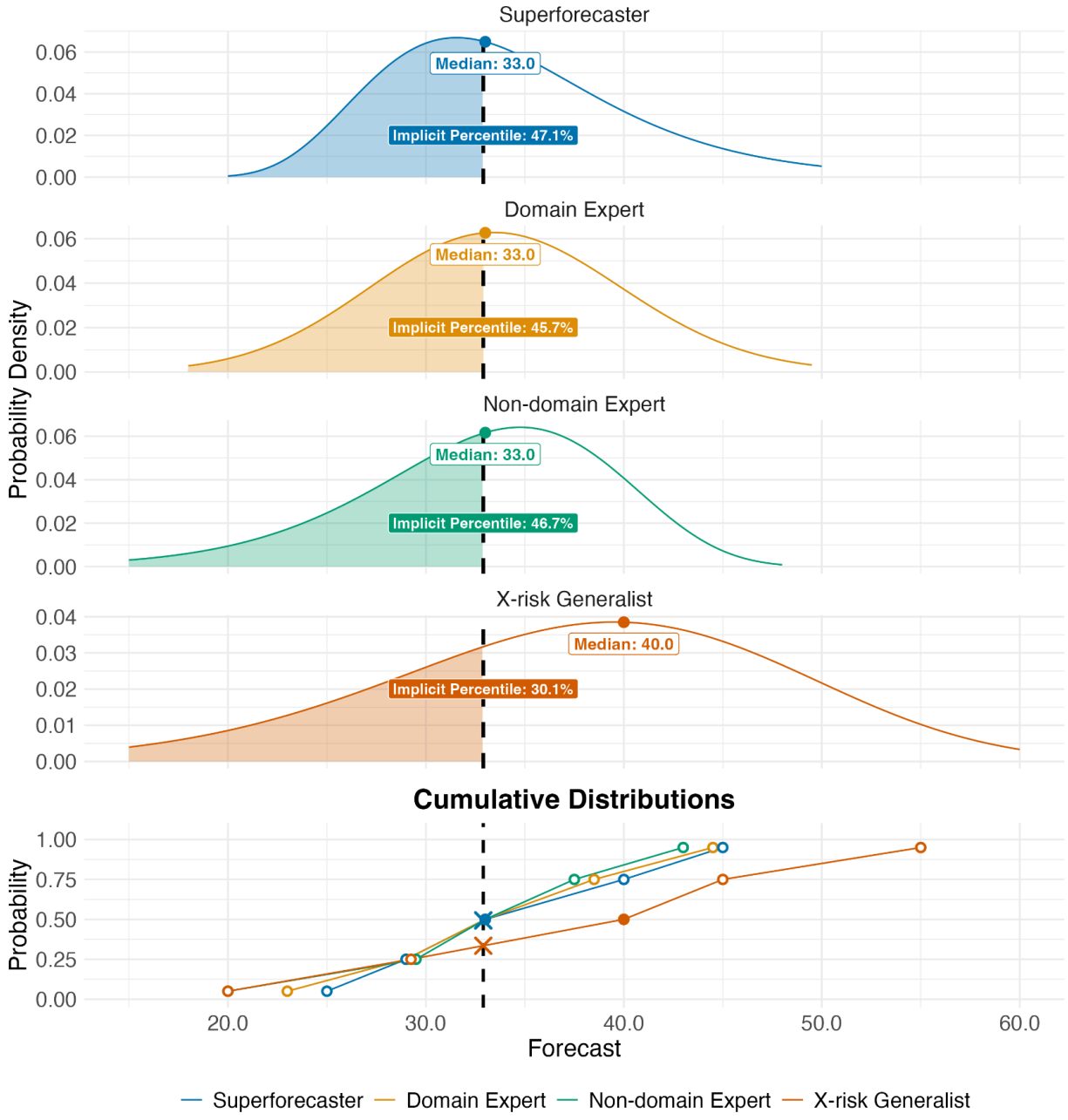
What will be the largest number of parameters of a machine learning model trained by the end of 2024



Resolution: 10.0T

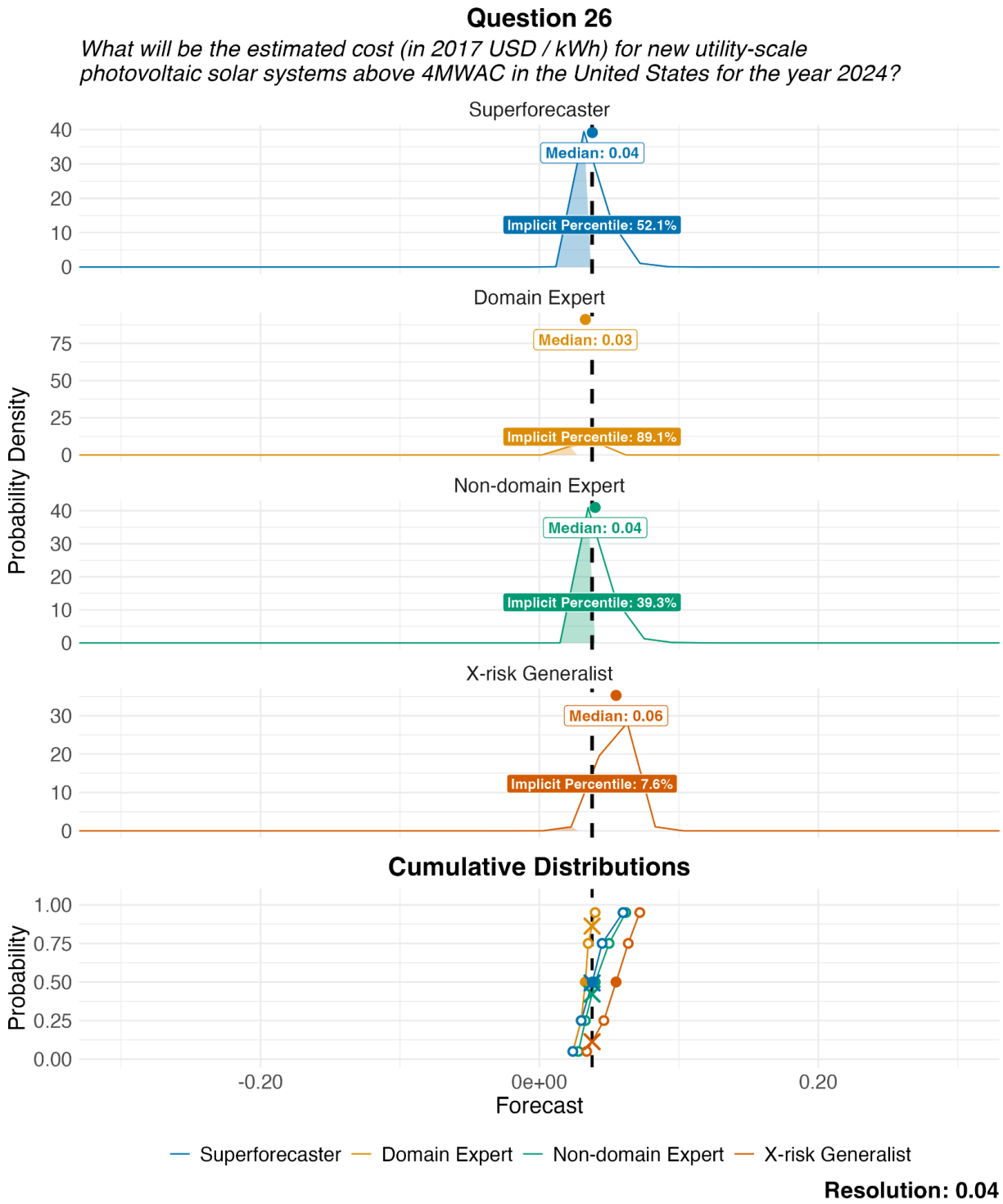
Question 50

Assume that Pew Research re-runs the survey linked here. What % of people in the median country in the survey will say that the development of artificial intelligence has mostly been a bad thing for society in 2024?



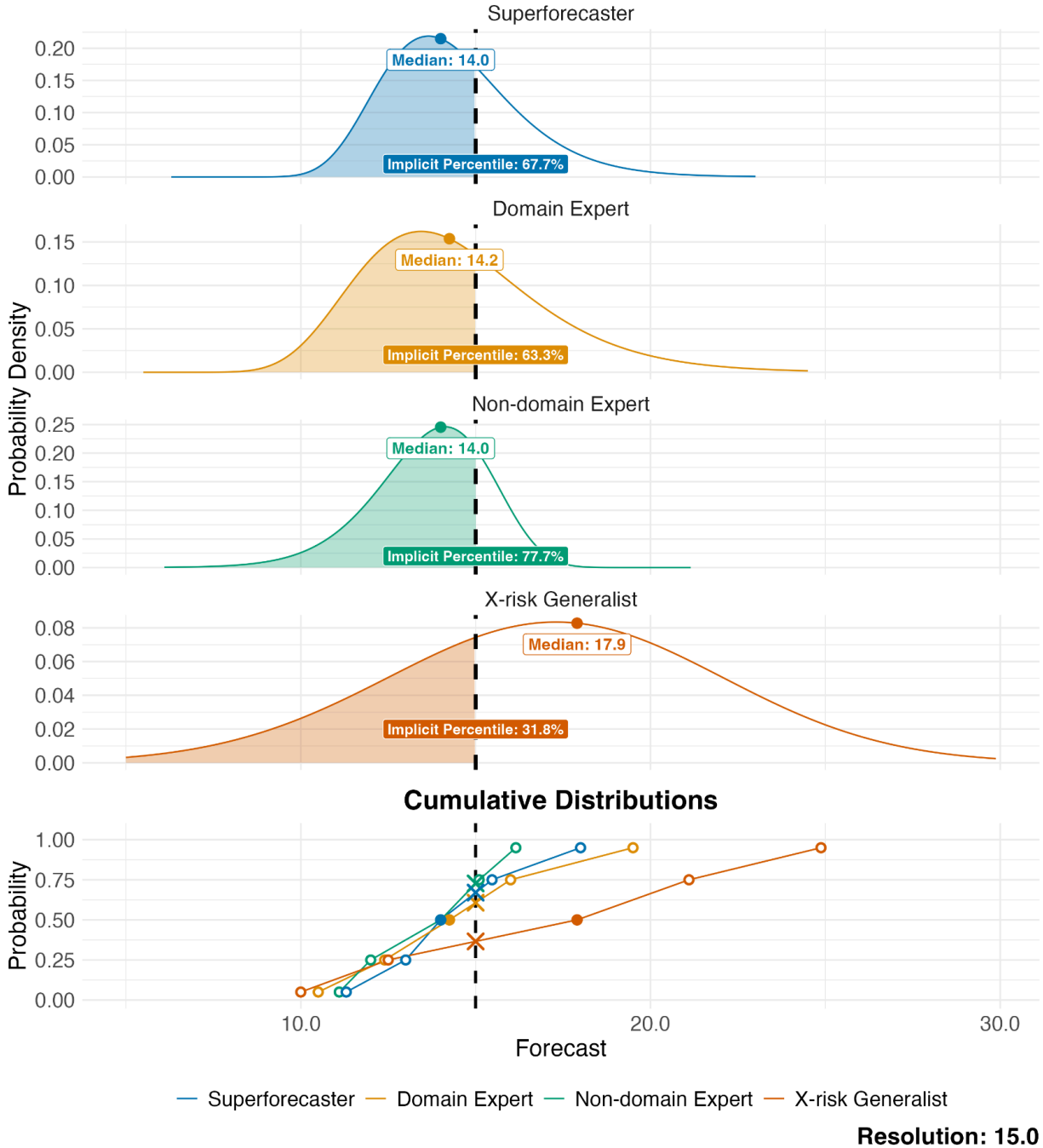
Resolution: 32.9

A5.2 Distributions for Climate-Related Questions



Question 28

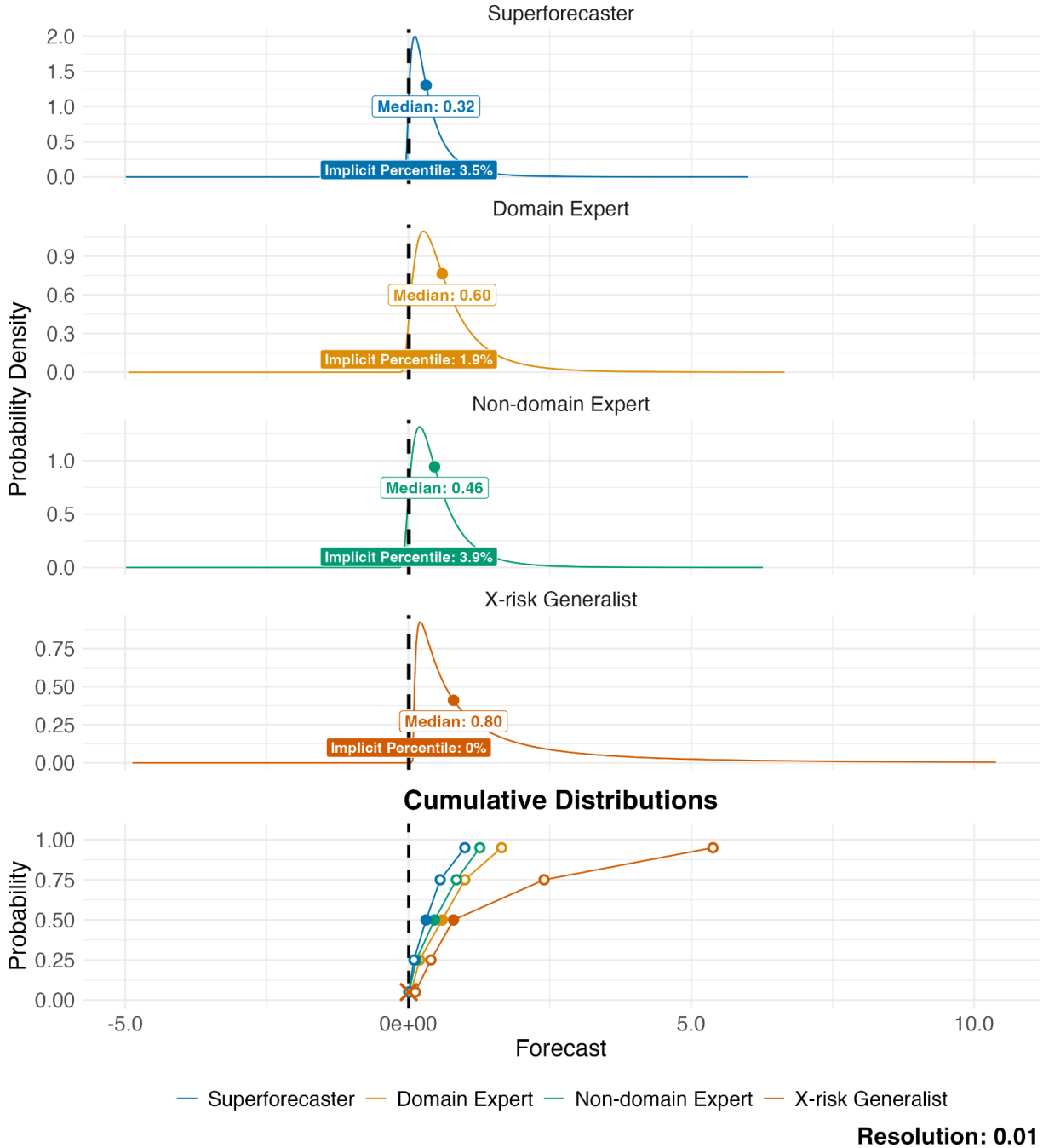
What percentage of the world's electricity will be provided by solar energy and wind energy combined in 2024?



Resolution: 15.0

Question 29

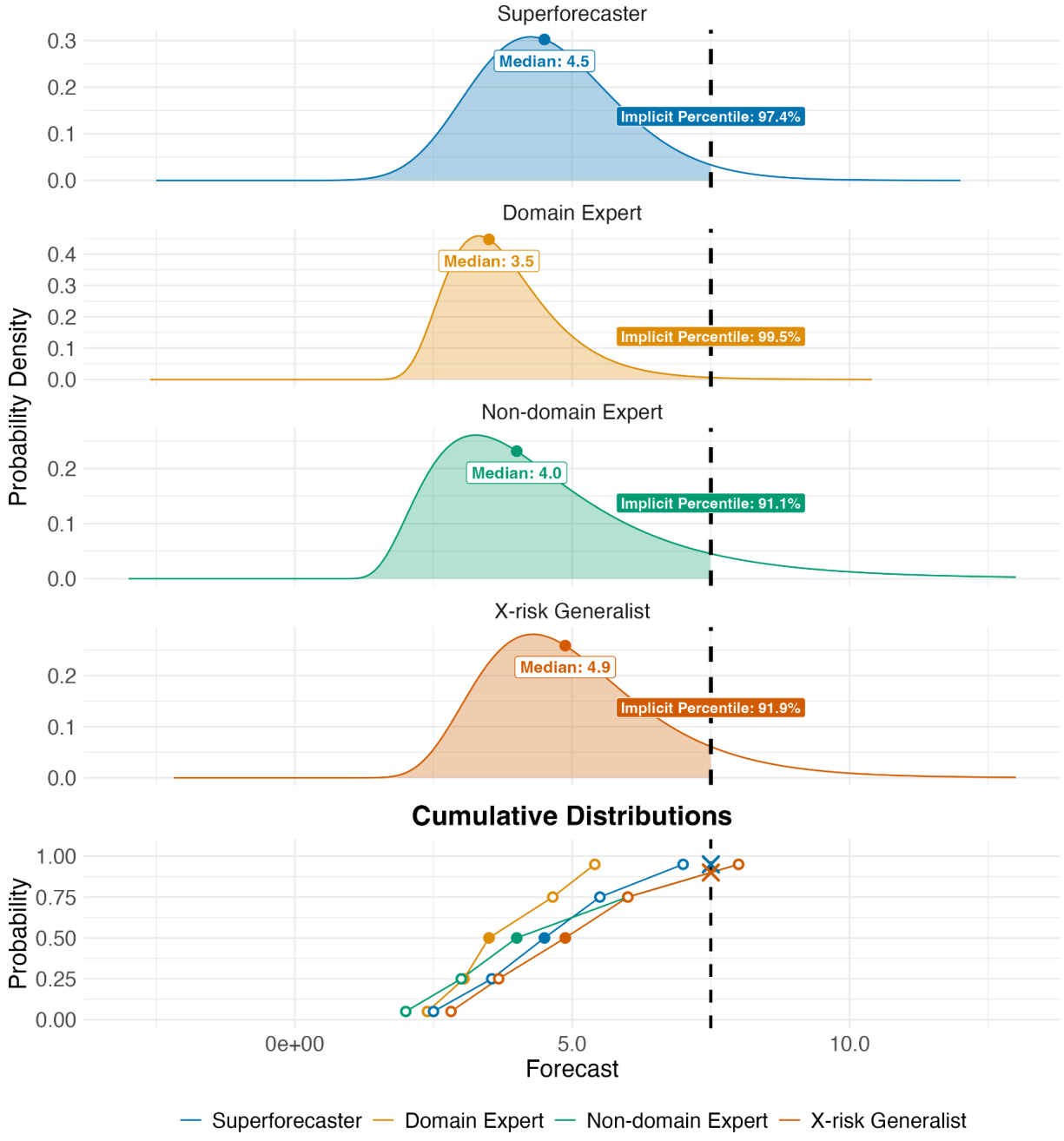
What will be the annual amount of CO2 captured and stored by direct air capture (in Mt CO2/year) in 2024?



Resolution: 0.01

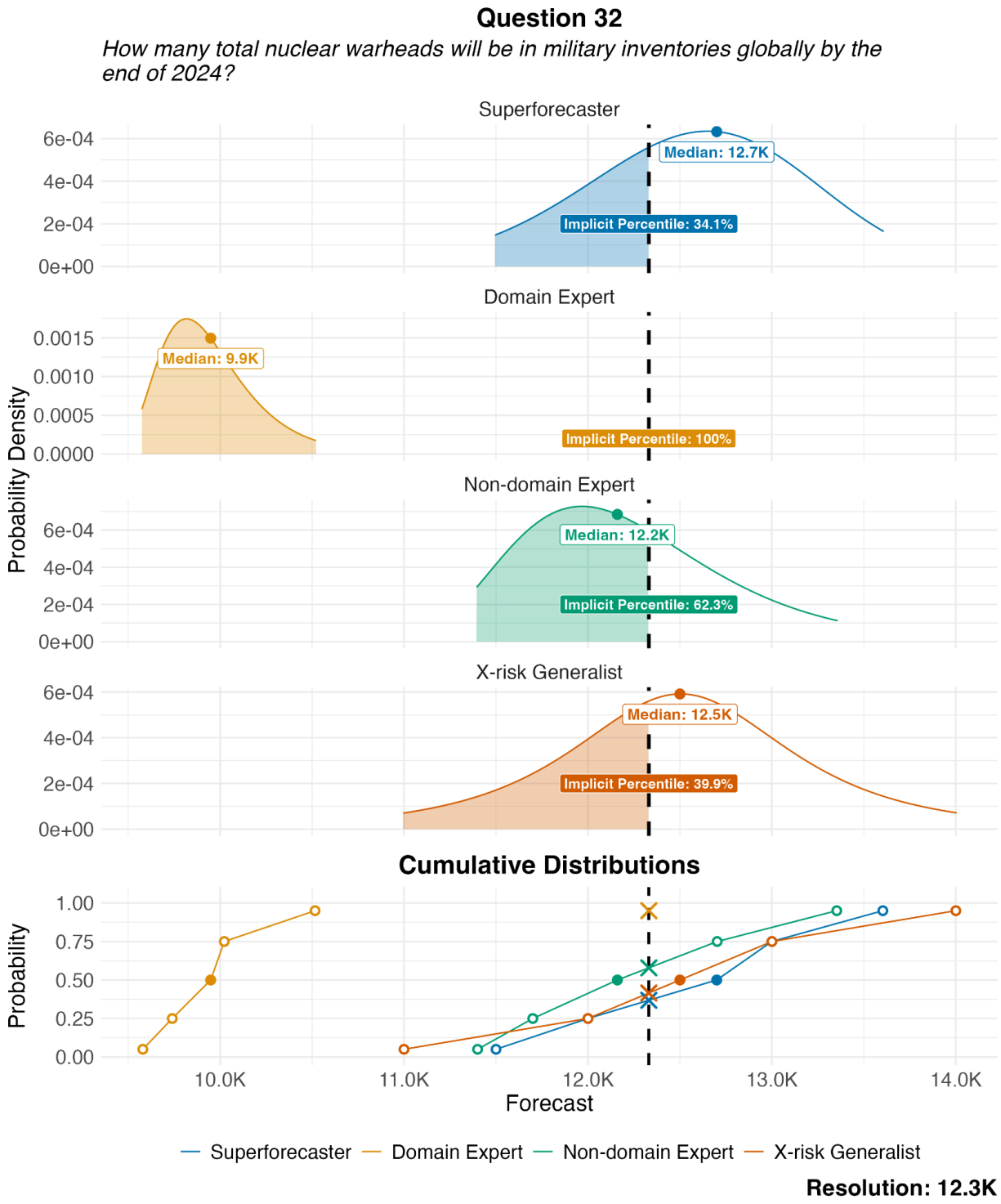
Question 30

How much will it cost to produce hydrogen from renewable electricity (in $ per kg of hydrogen) in 2024?



Resolution: 7.5

A5.3 Distributions for Nuclear-Related Questions



A5.4 Distributions for Biorisk-Related Questions