Final Report: Measuring Performance of LLM Forecasters for Quantitative Risk Modeling of Cyber Misuse

Jeff T. Mohl
jefftmohl@gmail.com

Madhav Khanal
mkhanal@rollins.edu

Jakub Kryś
jakub@safer-ai.org

Matt Smith
matt@safer-ai.org

Supervised Program for Alignment Research — Spring 2026

Abstract

Risk modelling is a common way of elucidating, estimating and managing risk across various safety-critical contexts, including AI-enabled cyberattacks. However, to derive quantitative estimates of real-life risk, current approaches often depend on expert judgment, which remains costly and time-consuming. In this work, we address this issue by developing a methodology for evaluating automated LLM "expert" forecasters that map benchmark performance to forecasts of real-world cyberattack capability uplift. Using forecasts over MITRE ATT&CK steps, we assess agreement between forecast variants under various experimental manipulations and find robust self-consistency across repeated runs, as well as broad invariance to alternative elicitation formats, expert persona prompts, and most prompt components, suggesting that LLM forecasters are not highly sensitive to these choices. Additionally, we evaluate LLM forecaster performance in a related cross-task prediction framework, finding that frontier models are capable of evaluating task difficulty and model performance to make accurate predictions on held out tasks. Together, these results suggest that LLM expert forecasters are a useful tool for rapid cyber risk assessments of future models and provide a framework for further improving these forecasters.

Keywords Forecasting · Cyber risk · AI safety

1 Introduction

AI enabled cyberattacks are an important risk vector for advanced AI systems, and have already resulted in direct harm (Hao et al., 2025). In prior work, SaferAI developed 9 cybersecurity risk models that leverage AI evaluation benchmarks as risk indicators to forecast the degree of cyber capability uplift from frontier models (Barrett et al., 2025b). Historically, the mapping between benchmark performance and forecast uplift has relied on human expert elicitation, i.e. a structured process in which domain experts give their best estimates of these quantitative values. This process is time consuming and expensive, which imposes limits on its utility. Instead, we would like to use automated LLM experts to elicit these forecasts for input into risk models, allowing for scalable monitoring of a wider range of models on shorter time scales. However, a methodology for measuring and evaluating the capability of LLM experts at this task is currently lacking. In this work, we develop one such methodology and leverage it to test the performance of simulated LLM experts in forecasting AI cyber capability uplift. We first measure the robustness and consistency of LLM forecasters conducting real-world risk assessments, and find that they produce consistent results

across runs and are robust to significant prompt changes. We then evaluate forecasters against ground-truth values on a proxy dataset, finding that they are able to synthesize information passed in-context to produce accurate predictions of difficulty of cybersecurity challenges and expected performance. Together, these results provide a framework for optimizing LLM forecasters for cybersecurity risk assessment and suggest that current frontier LLMs deliver valuable predictions even with minimal prompt optimization.

2 Establishing Forecaster Consistency

Dependable forecasts should, at minimum, demonstrate self-consistency when repeating those forecasts multiple times with identical data. We perform several experiments to establish the degree of self-consistency across multiple runs of LLM forecasters, and then leverage this consistency evaluation to explore the impact of various manipulations to the forecasting pipeline.

2.1 Forecast Elicitation for Consistency Experiments

The forecasts used in subsequent sections are elicited using the approach described in Barrett et al. (2025b). Briefly, the objective of these forecasts is to estimate the probability of success on a given step in a cyberattack killchain (drawn from the MITRE ATT&CK framework (MITRE. The mitre corporation, 2025a)), conditional on observed success on a provided cyber-relevant benchmark task. Instead of a single point estimate, several points on a probability distribution are elicited: 50th percentile (median), as well as 25th and 75th percentiles to reflect a credible uncertainty interval.

The LLM forecasting setup is complex and contains numerous features that could be experimentally manipulated. The most important components for the purposes of this work include:

- **Persona descriptions**: a set of custom prompts simulating various 'expert personas' during the estimation procedure. The inclusion of multiple personas aims to generate expert diversity and variance of opinions (Barrett et al., 2025a). This, in turn, is motivated by the observation that in forecasting tournaments, aggregating the predictions of multiple mediocre forecasters can approach the performance of top-calibrated 'superforecasters' (Tetlock and Gardner, 2016).
- **Baseline estimates**: estimates of the probability of success on each step *without* AI assistance (also as a 3-point distribution: median and the 25th/75th percentiles as confidence intervals).
- **AI capability analysis prompts**: a prompt instructing the forecaster on characterizing the capability of the capability of the observed AI system based on provided data.
- **Reasoning prompts**: a chain of custom prompts guiding the forecaster through the reasoning process, including detailed analysis of provided benchmark task and instructions for translating this analysis into a forecast.

2.2 Consistency Metrics

To develop a method for evaluating modifications to the forecasting procedure, we first define a metric that quantitatively captures the difference between various setups. For this purpose, we focus on the first-order and second-order Wasserstein distances (W1 and W2), which summarize the discrepancy between two distributions in terms of the amount of probability mass that must be shifted to transform one distribution into the other (Panaretos and Zemel, 2019). W1 (earthmover's distance) captures the average magnitude of this shift, while W2 places greater weight on larger

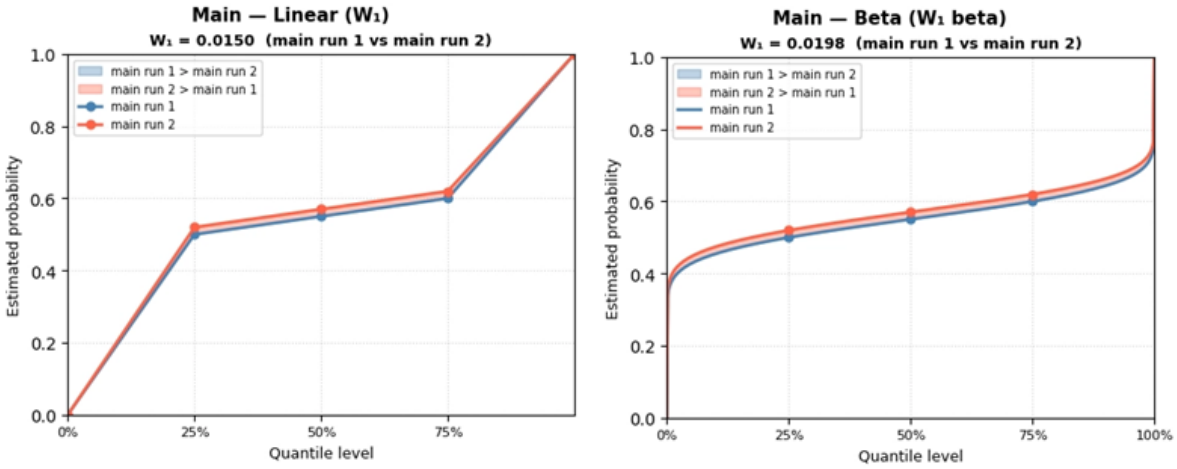

Figure 1: Example of W1 estimation for two consecutive runs of the same estimator settings. Left: estimation of W1 using a linear approach. Right: estimation using Beta distributions fit to the same elicited points.

discrepancies and is correspondingly more sensitive to occasional large deviations. Furthermore, we generate 95% confidence intervals by bootstrapping these distances across resampled sets of elicited points.

This metric serves two primary purposes:

- **Quantifying estimator consistency**: by comparing multiple runs with identical settings, we quantify the reliability of an estimator. Low run-to-run variability is desired, as this indicates the estimator predicts similar distributions when re-run with identical data.
- **Quantifying impact of experimental modifications**: if modifying some component of the estimator (e.g., prompt components) has no impact on consistency compared to the original setup, this suggests that the modification has minimal effect on the behavior of the estimator. This can be used to refine interventions that materially impact estimator performance.

The W1 and W2 distances are primarily computed by directly comparing elicited quantiles (see Fig. 1, left), as this removes an additional degree of uncertainty that would be introduced by fitting these points to an assumed distribution shape. However, we also fit the elicited points to a beta distribution before computing the distance using numeric integration (see Fig. 1, right). This provides a fairer comparison for manipulations which elicit a different number of point estimates, as the direct comparison requires linear interpolation to common points, which biases these comparisons towards larger distances.

2.3 Self-Consistency

We first compare the forecaster self-consistency across multiple runs of the main, unmodified elicitation approach. Elicitations for these plots are conducted with a single expert persona and three MITRE steps from a chosen cyberattack scenario. Each elicitation is repeated 20 times, resulting in 190 pairs for comparison within each task. Two LLM forecasters (Claude Sonnet 4.6 and GPT-5 mini) are tested across all steps (see Fig. 2)

While elicitations from GPT-5 mini had higher run-to-run average distance (indicating less agreement), both forecaster models demonstrated robust self-consistency. For reference, a W1 value

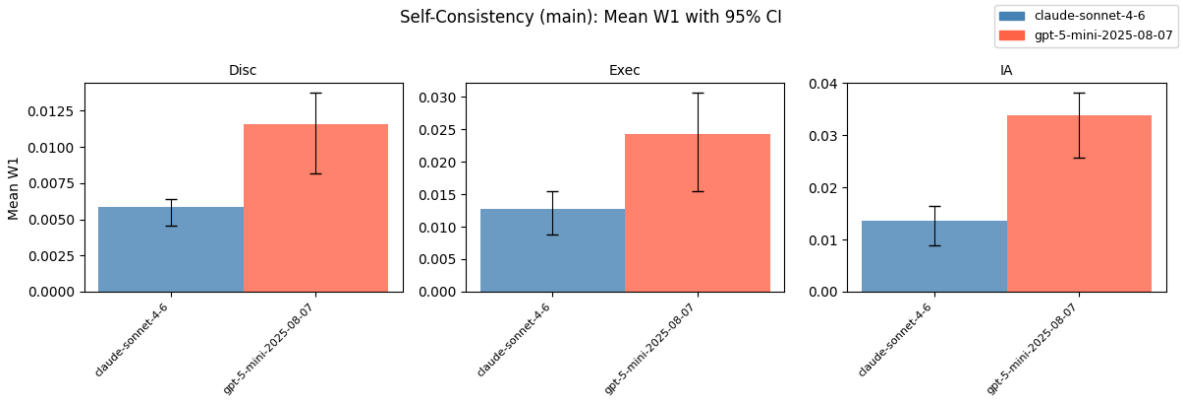

Figure 2: Mean W1 distance and 95% confidence intervals for three MITRE steps using two estimator models.

of 0.05 indicates that the probability mass was shifted an average of 5% across runs. This result is encouraging, as it suggests that there is not a large degree of stochasticity inherent to the model.

2.4 Elicitation Invariance

Aside from run-to-run consistency with identical settings, we also evaluate whether the forecasters remain statistically consistent when asked to elicit the same distribution in different ways. For instance, a forecaster asked to estimate the probability of failure instead of success should report a value that is comparable to $1 - \text{P(success)}$.

We generate 5 variations of our elicitation approach, including the 'negation' example above and four approaches where the elicited quantiles are different from the 25%, 50%, and 75% points used by default. All other conditions are kept identical to the 'main' runs. We vary quantiles both by number of points (e.g., eliciting deciles from 10-90%) and the location of points, including a pseudo-random selection of points (15%, 60%, and 77%). We then fit a Beta distribution to each elicitation and compute the W1 distances between each elicitation and the unmodified 'main' elicitations.

We find that the consistency across elicitation approaches is generally comparable to the internal consistency in the default approach (see Fig. 3), suggesting that forecasters are consistent in their estimations despite being required to provide their estimates in diverse ways. One possible exception is the elicitation of 5th, 50th, and 95th quantiles from Claude Sonnet 4.6, which meaningfully differs from the default approach across task steps.

2.5 Model Choice Explains More Variance than Expert Persona

The forecaster includes 10 expert personas, intended to create diversity of opinion, but it is not clear whether the 10 expert personas induce meaningful forecast variation compared with the variation induced by the choice of LLM forecaster. Since each forecast is represented as a probability distribution rather than a single scalar, we use Fréchet ANOVA (Dubey and Müller, 2019) to quantify variance, which extends ANOVA to metric-space objects by partitioning variation in distributional distances. Forecasts were represented as beta distributions fit to the elicited 25th, 50th, and 75th percentiles using least-squares regression, rejecting fits where any fitted quantile deviated from the target by more than 0.05, i.e., $|\text{CDF}(p_q) - q| > 0.05$, where p_q denotes the elicited forecast value

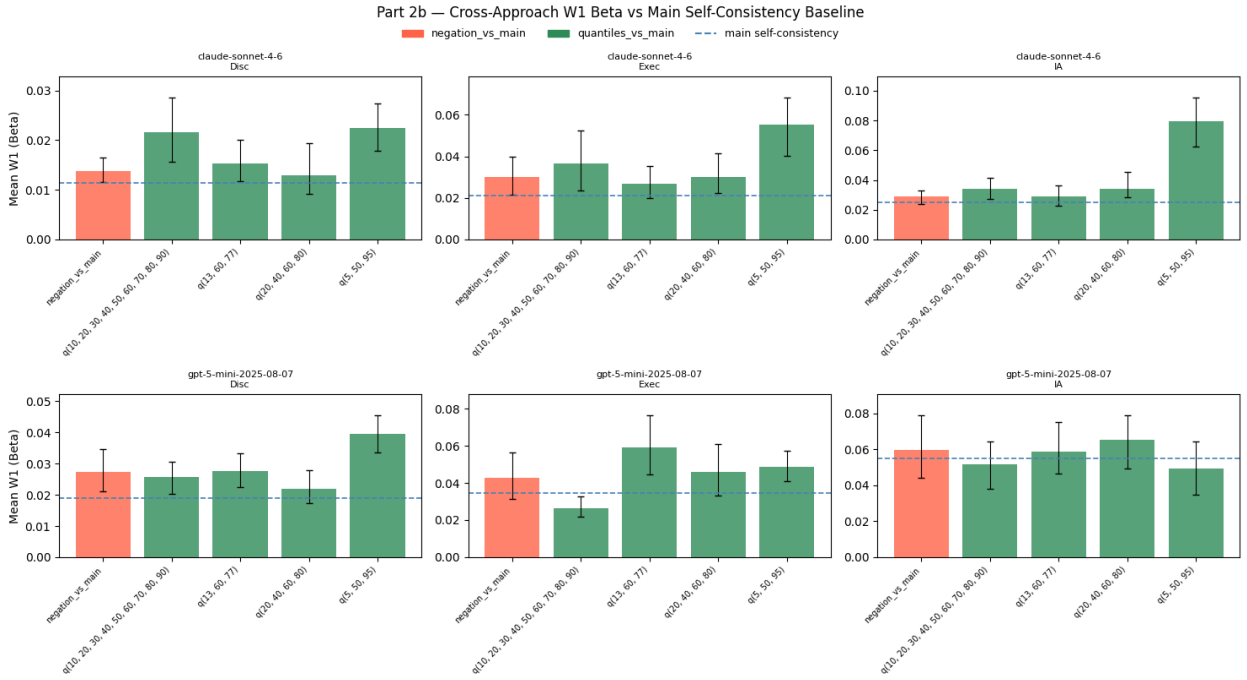

Figure 3: Mean W1 distance across approaches and 95% bootstrapped confidence intervals for three MITRE steps using two estimator models. W1 distance is computed between all pairs of the main approach runs and runs from various elicitation approaches. The blue reference line indicates the mean within-approach variance of the corresponding model from figure 2.

at quantile level q. We report the Fréchet intraclass correlation coefficient (ICC_F), which can be interpreted as the proportion of total distributional variance attributable to a grouping factor.

We test three LLM forecasters (Claude Sonnet 4.5, GPT-4o, Gemini 2.5 Pro) with 10 expert personas and three attack steps (10 runs per configuration). Persona explained a negligible fraction of variance across runs. Claude and GPT-4o as forecasters show no significant variance ($5.7\% < ICC_F < 13.9\%$, all $p > 0.05$), while Gemini produces more variance ($18.3\% < ICC_F < 24.9\%$, $p < 0.05$), but less than variance due to cross-model effects. Specifically, comparing LLM forecasters based on different AI models, we find that model choice explains 48%–69% of distributional variance (all $p < 0.001$), 3.5–12× more than persona. Fig. 4 and Fig. 5 show tight persona clustering within forecaster models and a clear separation between models.

2.6 Prompt Sensitivity

We also measure the W1 distance between 7 prompt variants and the unmodified control prompt (approx. 615 words, see Section 2.1). We use the following variants: full prompt, no baseline Confidence Interval (CI), no baseline median values, removing both CI and baseline, removing the AI capability analysis, removing the 3-phase reasoning scaffold, and removing all analysis + reasoning (reduced to 100 words). We find that removing the baseline estimate (currently estimated by humans) causes the largest shift (W1 = 0.154, 18.1× variance compared to the unmodified prompt), see Fig. 6. Trimming the entire reasoning and analysis produces a smaller distribution shift and more consistent responses (W1 = 0.016, 0.07× variance compared to the unmodified prompt).

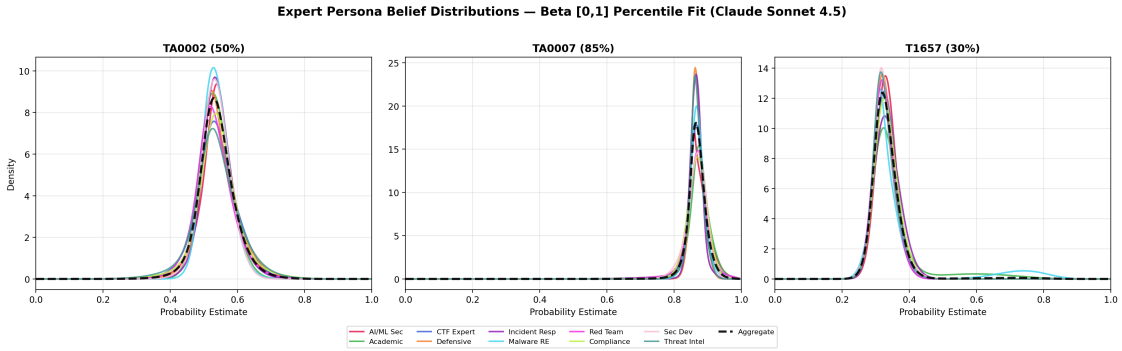

Figure 4: Probability distributions elicited from Claude Sonnet 4.5 as the forecaster show little variance across expert personas, for all three MITRE steps tested.

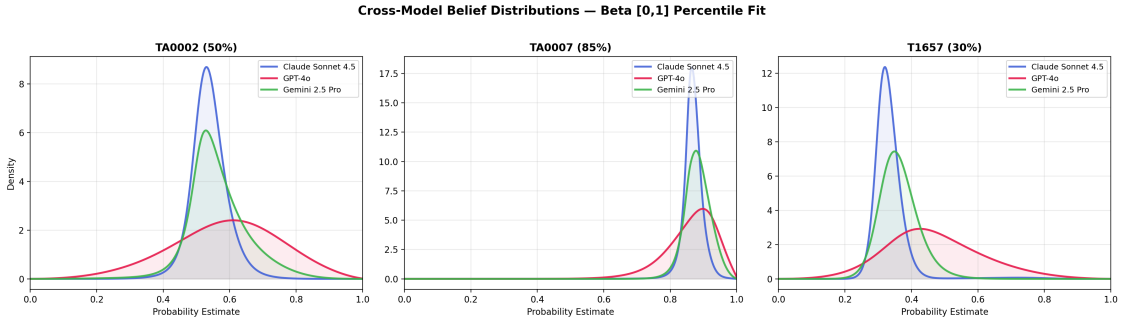

Figure 5: Probability distributions elicited from three different LLMs show clear separation.

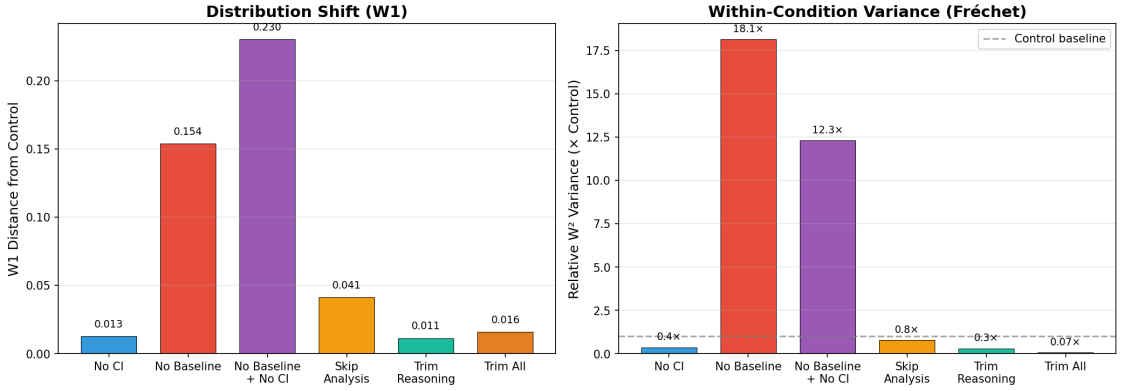

Figure 6: Impact of prompt manipulations on forecaster predictions. Left:impact of various prompt ablations in terms of the shift relative to the default prompting strategy. Right: variance within each prompt ablation. The gray reference line shows the within condition variance across multiple elicitations with the unmodified prompt.

2.7 Baseline Anchoring

To test whether the LLM forecasters anchor their estimates strongly on the baseline distributions supplied in-context, we remove the actual baseline and confidence intervals and replace them with fixed median probabilities from 10% to 90% (in 10% increments, 10 runs per baseline). Using Claude Sonnet 4.6 as the forecaster, the median elicited probability follows a roughly linear trend with the provided baseline, though this appears to encounter a ceiling effect at high baseline levels (Fig. 7). This suggests that LLM forecasters strongly anchor on the supplied baseline values.

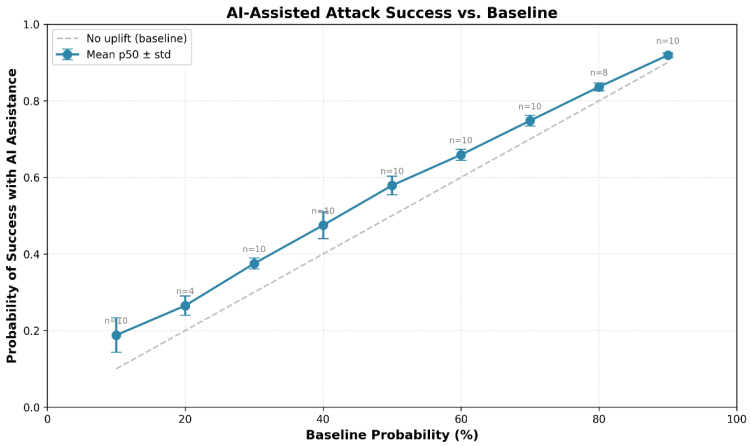

Figure 7: Effect of varying the median baseline value supplied to the LLM forecaster, with all other settings held constant. Across most baseline values, the LLM forecaster appears to report a fixed level of uplift. This is attenuated when the baseline value approaches 100% (ceiling effect).

3 Measuring Forecaster Accuracy on Ground Truth Proxies

One difficulty with empirically validating LLM forecaster performance is that reference outcomes are challenging to establish and may take months or years to occur. To address this issue, we implement a proxy experiment where the ground-truth values are taken from the performance of popular AI models on a variety of cybersecurity tasks. We then establish several baseline comparator approaches, and use these to evaluate the accuracy of several recently released frontier models as forecasters.

3.1 Proxy Forecasts using Intra-benchmark Prediction

We develop a modified forecaster pipeline similar to that described in Section 2.1 to elicit task level forecasts that can be evaluated against actual performance. We use an open-access dataset produced by Lyptus Research (Payne et al., 2026) as the source of ground truth for this experiment. This dataset, unlike many others which only report aggregate scores on the whole benchmark, provides scores for a set of 12 common LLMs on each of its tasks in addition to an empirical task difficulty rating in the form of human-generated first solve times (FST).

This proxy experiment is meant to resemble the primary objective of risk modelling, i.e. predicting AI performance on an attack step of unknown difficulty, based on this AI's performance on a related set of cybersecurity tasks. We split the dataset into 5 difficulty bins based on FST, obtaining approximately 50 tasks per bin. The objective of the forecaster is to use the context of AI performance

on a subset of difficulty bins to predict AI performance on a single task selected from a held out 'target bin'. Importantly, the information about the FST of this target task is *not* passed to the LLM estimator, similarly to how this estimator would have to independently estimate the difficulty of a real-life attack step.

For each forecast, the forecaster's context consists of the bin-level average performance of the examined AI model on 4 of the 5 difficulty bins, together with the descriptions of 3 representative tasks from each of these reference bins. The forecaster is then provided with the target task's description and asked to predict the probability of success on this task. Specifically, the forecaster gives its 25th, 50th, and 75th percentile estimates to establish the median performance estimate and a credible confidence range. For each bin, 5 target tasks were selected representing the a range of difficulties within the bin as defined by FST. Importantly, predictions for each target task are performed using a fully independent forecaster run.

3.2 Intra-benchmark Performance Evaluation

To evaluate forecaster performance, we compare the predicted pass rates to the actual pass rates for the AI model being examined. We use two performance metrics: single-point 50th percentile forecasts are evaluated using the Brier score:

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2\,,$$

where $i = 1, \ldots, N$ indexes benchmark tasks, N is the total number of evaluated tasks, p_i is the forecaster's predicted probability that the AI model passes task i, and $y_i \in \{0, 1\}$ is the observed outcome, with $y_i = 1$ indicating that the model passed and $y_i = 0$ indicating that it failed.

The complete beta distributions fitted to the 3 elicited estimates are evaluated using the Continuous Ranked Probability Score (CRPS):

$$\mathrm{CRPS}(F_i, y_i) = \int_{-\infty}^{\infty} \left(F_i(x) - \{x \geq y_i\}\right)^2 \, dx\,.$$

where F_i denotes the predictive cumulative distribution function for the pass rate on task i, obtained by fitting a Beta distribution to the three elicited estimates, and $\mathbf{1}\{x \geq y_i\}$ is an indicator function equal to 1 when $x \geq y_i$ and 0 otherwise. Both metrics are averaged across all target tasks and examined AI models to produce a single overall score for each forecaster.

These scores provide useful relative performance, but a comparison with a baseline is needed to evaluate performance in absolute terms. We implemented four such comparisons, representing increasing forecast precision:

- **Uninformed Estimates:** An uninformed forecaster that always produces a flat distribution (with a median of 0.5) on all tasks, resulting in a Brier score of 0.25 and CRPS of 0.5 for all cases.
- **Model Level Pass Rates:** Mean AI model performance across all tasks is used to predict performance on each individual task.
- **Model + Bin Pass Rates:** Mean AI model performance within each difficulty bin is used to predict individual task performance within that bin.
- **Item Level Response Logistic Fit:** A logistic curve between FST and task performance is fitted across all tasks for each AI model and used to predict individual task performance from the FST.

These baselines represent a continuum of informational content, from completely uninformed guessing to almost perfect knowledge of the correlation between FST and task-level performance.

3.3 Intra-benchmark Forecasting Performance

To assess LLM forecaster performance, we conduct a sweep using a set of 5 recently released LLMs (Fig. 8). We compare LLM forecasts against a set of increasingly well-informed baseline predictions ranging from 'uninformed' to 'perfect task-level difficulty knowledge'. All LLM forecasters outperform uninformed guessing, suggesting that their forecasts are genuinely informative. They also perform better than minimally informed baselines that account only for the average model score across all tasks. This suggests that all forecasters are capable of parsing some task-level difficulty information about the target task to better predict an AI's score on this task. However, there is substantial variability across LLM forecasters models, with the largest and most recently released LLMs performing best despite identical settings across runs. The best forecaster (GPT 5.5) performs as well as the maximally informed 'item response logistic fit', despite not having access to the task-level difficulty used for this fit.

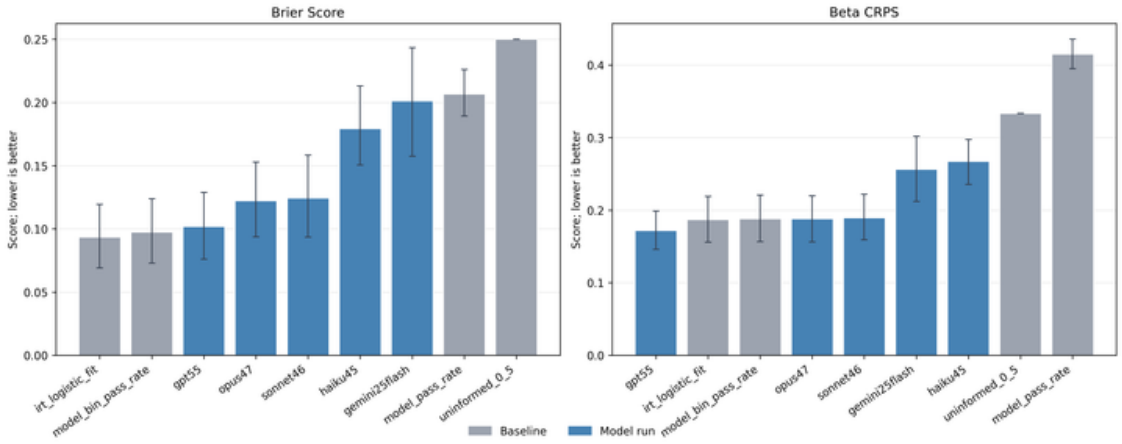

Figure 8: Performance of multiple LLM forecasters on the intra-benchmark experiment. Left: mean Brier score and bootstrapped 95% confidence intervals based on task-level forecasts. Right: mean CRPS and 95% bootstrapped confidence intervals for the same predictions. All LLM forecasters are evaluated on identical tasks (n = 190) and with identical settings. Note: some LLMs refused a number of API requests due to developers' guardrails. We only include tasks which were evaluated by all forecaster LLMs

4 Discussion

This work develops a methodology for evaluating LLM-based forecasters of cyber capability uplift and applies it to characterize the behavior of current frontier LLMs on this task. We focus on two primary primary components of this problem: understanding the reliability and consistency of LLM forecasters under various manipulations, and evaluating their performance on cybersecurity task prediction.

We find that repeated elicitation under identical conditions produces highly consistent distributions, with Wasserstein distances corresponding to roughly 5% probability mass displacement

between paired runs. This consistency extends across most experimental manipulations of the elicitation procedure, including: alternative distribution percentiles, reframing the elicitation question as P(failure) rather than P(success), different expert personas, and various ablations of the LLM forecaster's prompt. The primary exception to this rule is the baseline probability supplied to the forecaster (i.e., the probability of a successful completion of an attack step *wihout* AI assistance). Removing this baseline dramatically increases run-to-run variance, and artificially modifying its value produces a near-linear shift in the elicited probabilities. Together, these results describe a forecaster that is internally stable, but anchored heavily to the provided baseline values.

The contrast between the prompt structure and LLM choice carries practical implications for LLM elicitation design. Variance across the ten engineered expert personas was small and frequently statistically indistinguishable from noise, whereas variance across LLM choice was an order of magnitude larger and consistently significant. This suggests that ensembling across LLMs is likely to be more informative than ensembling across personas for capturing relevant uncertainty in these forecasts. Moreover, the engineering effort spent on persona design may have marginal returns relative to simply running multiple frontier LLMs in parallel.

Using a proxy experiment where ground-truth data is available, all tested LLM forecasters perform better than uninformed or naive approaches using only the model's mean performance, and the best-performing forecaster approaches the accuracy of a logistic fit with direct access to task-level FST. This indicates that, at least when forecasting tasks structurally similar to those already present in the evaluation context, LLM forecasters do extract meaningful difficulty signal from task descriptions, rather than merely averaging across reference performance. The substantial variability in performance across LLM forecasters, with newer and larger LLMs performing best, suggests that forecast quality scales with general capability and is likely to continue improving with AI progress.

Several limitations qualify these findings. First, self-consistency is necessary but not sufficient for forecast quality and other characteristics should be tested thoroughly as well. Next, this work does not establish accuracy at the underlying task of forecasting real-world cyber uplift, only at the proxy task of intra-benchmark performance prediction. The structural similarity between the proxy task and standard benchmark evaluations likely causes observed accuracy to overestimate performance on the genuine target of forecasting capability transfer to MITRE ATT&CK steps, where the available context is less informative about the target.

References

Stephen Barrett, Philip Fox, Joshua Krook, Tuneer Mondal, Simon Mylius, and Alejandro Tlaie. Assessing confidence in frontier ai safety cases, 2025a. URL https://arxiv.org/abs/2502.05791.

Steve Barrett, Malcolm Murray, Otter Quarks, Matthew Smith, Jakub Kryś, Siméon Campos, Alejandro Tlaie Boria, Chloé Touzet, Sevan Hayrapet, Fred Heiding, Omer Nevo, Adam Swanda, Jair Aguirre, Asher Brass Gershovich, Eric Clay, Ryan Fetterman, Mario Fritz, Marc Juarez, Vasilios Mavroudis, and Henry Papadatos. Toward quantitative modeling of cybersecurity risks due to ai misuse. *arXiv preprint arXiv:2512.08864*, 2025b. URL https://arxiv.org/abs/2512.08864.

Paromita Dubey and Hans-Georg Müller. Fréchet analysis of variance for random objects. *Biometrika*, 106(4):803–821, 10 2019. ISSN 0006-3444. doi: 10.1093/biomet/asz052. URL https://doi.org/10.1093/biomet/asz052.

Wei Hao, Van Tran, Vincent Rideout, Zixi Wang, AnMei Dasbach-Prisk, M. H. Afifi, Junfeng Yang, Ethan Katz-Bassett, Grant Ho, and Asaf Cidon. Do spammers dream of electric sheep? characterizing the prevalence of llm-generated malicious emails. In *Proceedings of the 2025 ACM Internet Measurement Conference*, IMC '25, page 192–203, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400718601. doi: 10.1145/3730567.3732922. URL https://doi.org/10.1145/3730567.3732922.

MITRE. The mitre corporation, 2025a. Mitre. the mitre corporation, 2025a. URL https://attack.mitre.org/.

Victor M. Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431, March 2019. ISSN 2326-831X. doi: 10.1146/annurev-statistics-030718-104938. URL http://dx.doi.org/10.1146/annurev-statistics-030718-104938.

Jack Payne, Jeremy Miller, and Sean Peters. Offensive cybersecurity time horizons. Research note, Lyptus Research, 4 2026. URL https://lyptusresearch.org/research/offensive-cyber-time-horizons.

Philip E. Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Random House Business Books, 2016.