Scalable Delphi: Large Language Models for Structured Risk Estimation

Tobias Lorenz [1] **Mario Fritz** [1]

Abstract

Quantitative risk assessment in high-stakes domains relies on structured expert elicitation to estimate unobservable properties. The gold standard—the Delphi method—produces calibrated, auditable judgments but requires months of coordination and specialist time, placing rigorous risk assessment out of reach for most applications. We investigate whether Large Language Models (LLMs) can serve as scalable proxies for structured expert elicitation. We propose Scalable Delphi, adapting the classical protocol for LLMs with diverse expert personas, iterative refinement, and rationale sharing. Because target quantities are typically unobservable, we develop an evaluation framework based on necessary conditions: calibration against verifiable proxies, sensitivity to evidence, and alignment with human expert judgment. We evaluate in the domain of AI-augmented cybersecurity risk, using three capability benchmarks and independent human elicitation studies. LLM panels achieve strong correlations with benchmark ground truth (Pearson $r = 0.87$–0.95), improve systematically as evidence is added, and align with human expert panels—in one comparison, closer to a human panel than the two human panels are to each other. This demonstrates that LLM-based elicitation can extend structured expert judgment to settings where traditional methods are infeasible, reducing elicitation time from months to minutes.

1. Introduction

What is the probability that a ransomware attack against a regional hospital succeeds? How likely is a semiconductor shortage to disrupt automotive production next quarter? By how much does access to AI tools increase a phishing campaign's success rate? How safe is a proposed small modular reactor design? These questions span different domains, but they share a common structure: they are central to high-stakes decisions, yet they cannot be answered through direct measurement. We cannot run controlled experiments on real adversaries, simulate the full complexity of global supply chains, or observe the counterfactual where a reactor design fails. Instead, practitioners rely on structured expert judgment—panels of specialists who provide probability estimates conditioned on available evidence.

The most established methodology for such elicitation is the Delphi method (Dalkey & Helmer, 1963), developed at RAND in the 1950s and now standard practice in high stakes domains. A typical Delphi study proceeds in rounds: experts first receive detailed briefing materials describing the problem, available evidence, and the specific quantities to be estimated. Each expert independently provides probability distributions or point estimates, along with written rationales. A facilitator then aggregates the responses, anonymizes them, and circulates a summary—including areas of agreement, disagreement, and the reasoning behind outlier judgments. Experts revise their estimates in light of this feedback, and the process repeats until convergence or a predetermined number of rounds. The result is a structured, auditable record of expert judgment that accounts for uncertainty and mitigates some forms of individual bias.

This rigor comes at considerable cost. A three-round Delphi study typically takes three to four months, including preparation, elicitation, and analysis (Gordon, 2009). Each round requires coordinating schedules among a panel of multiple experts, who may each charge several hundred dollars per hour. The NUREG-1150 study—a landmark risk assessment for five U.S. nuclear plants—involved several hundred experts estimating hundreds of parameters over multiple years (U.S. Nuclear Regulatory Commission, 1990). Such investments are justified when the stakes are high enough, but they place quantitative risk assessment out of reach for most applications. Small and mid-sized organizations cannot afford it; rapidly evolving domains like AI security cannot wait for it; and risk models with hundreds of interdependent parameters cannot be populated one expert panel at a time. The result is a widening gap: we have increasingly sophisticated tools for building risk models, but populating them with defensible estimates remains a bottleneck.

This paper explores a different approach: using Large Lan-

[1]CISPA Helmholtz Center for Information Security, Germany. {tobias.lorenz,fritz}@cispa.de.

Preprint. March 10, 2026.

guage Models (LLMs) as scalable proxies for structured expert elicitation. The intuition is that LLMs, trained on vast technical corpora including academic literature, threat reports, and domain-specific documentation, may have internalized enough knowledge and reasoning abilities to provide useful probability estimates—particularly when prompted with the same structured evidence that would be provided to human experts. There is reason for cautious optimism: LLMs have demonstrated calibration on forecasting tasks (Kadavath et al., 2022; Halawi et al., 2024), can synthesize information across diverse technical sources, and can articulate reasoning that can be inspected and challenged. If these capabilities extend to structured risk estimation, the economics of risk assessment change dramatically: LLM-based elicitation is cheap enough to apply broadly, fast enough to keep pace with evolving threats, reproducible enough to audit and version, and able to handle thousands of estimates without the quality degradation seen in extended human elicitation sessions (Gordon, 2009). This intuition has begun to surface in preliminary studies (Nóbrega et al., 2023; Mueller et al., 2024; Papakonstantinou et al., 2025; Barrett et al., 2025). However, LLMs can produce confident but poorly-grounded estimates and are not explicitly trained for probabilistic reasoning. Whether they can serve as reliable proxies for expert judgment is an empirical question—one that requires careful analysis. We develop this intuition into a principled methodology: formalizing the framework, instantiating it across multiple models and benchmarks, and systematically evaluating the conditions under which it produces reliable estimates.

Answering this question is not straightforward. Because the target quantities are unobservable by construction, validity cannot be established through direct comparison to ground truth. But the absence of ground truth does not preclude meaningful evaluation. We assess whether LLM-based estimates exhibit properties central to reliable estimation: appropriate sensitivity to relevant evidence, accurate prediction of quantities we can verify, and alignment with human expert judgment where available.

We instantiate this investigation in the domain of AI-augmented cybersecurity risk. This domain offers three properties that make it well-suited as a testbed: threat models are well-specified and decomposable into capabilities, benchmarks with known ground truth allow partial validation, and a recent expert elicitation study provides human baselines for comparison (Murray et al., 2025).

Concretely, we evaluate LLM-based Delphi panels on three cybersecurity benchmarks (BountyBench, Cybench, CyberGym), testing calibration, evidence sensitivity, and alignment with human experts. Our findings are encouraging: LLM estimates achieve strong correlations with benchmark ground truth (Pearson correlation $r = 0.87$–0.95), improve systematically as evidence is added, and align with human expert panels—in one comparison, closer to a human panel than the two human panels are to each other (5.0 vs. 16.6 pp mean absolute difference). These results suggest that LLM-based elicitation can serve as a useful complement to traditional methods, particularly in resource-constrained or time-sensitive contexts.

Concretely, we make the following contributions:

- We propose **Scalable Delphi**, adapting structured expert elicitation for LLMs with personas, iterative refinement, and rationale sharing.
- We develop an **evaluation framework** for latent quantity estimation: calibration on verifiable proxies, sensitivity to evidence, and alignment with human judgment.
- We demonstrate **strong empirical results**: LLM estimates correlate with benchmark ground truth ($r = 0.87$–0.95) and align with human expert panels.

2. Scalable Delphi Method

Structured risk models decompose complex risks into networks of conditional probabilities—whether formalized as fault trees, Bayesian networks, or influence diagrams (Bedford & Cooke, 2001). Each node requires a probability estimate. Some quantities can be measured directly; many others cannot. For these latent quantities, structured expert elicitation is the standard approach (Cooke, 1991). We focus on the core subtask: producing calibrated estimates for individual quantities, which can then populate such models.

Traditional Delphi elicitation convenes a panel of human experts who provide independent estimates, review anonymized peer feedback, and iteratively refine their judgments until convergence. This produces calibrated, auditable estimates—but as discussed in the introduction (Section 1), is costly and requires months. We propose Scalable Delphi: given the same structured evidence provided to human experts, a panel of LLM agents produces probability estimates through the same deliberative process.

2.1. The Estimation Task

Let Q denote a quantity to be estimated in a risk model, and let E denote the relevant evidence. The elicitation task is to produce an estimate $\hat{p}$ of $P(Q \mid E)$ together with a measure of uncertainty.

Mirroring traditional Delphi, we build a panel of k expert agents. Each agent j is instantiated with a distinct persona π_j representing a particular perspective or expertise. These agents produce independent estimates for each round r, conditioned on Q, E, and the anonymized and aggregated

feedback of the previous round $F^{(r-1)}$:

$$\hat{p}_j^{(r)} = \mathcal{M}(Q, E, \pi_j, F^{(r-1)}). \tag{1}$$

$\mathcal{M}$ is the language model, and $F^{(0)} = \emptyset$ for the first round. The final estimate is aggregated across the panel after the final round:

$$\hat{p} = \frac{1}{k} \sum_{j=1}^{k} \hat{p}_j^{(R)}, \tag{2}$$

where R is the total number of rounds.

2.2. Elicitation Protocol

We instantiate the full Delphi protocol with LLM agents in all roles: expert panelists *and* mediator. This enables fully automated elicitation with no human involvement beyond problem specification.

Expert Panel. We instantiate a panel of k experts with personas $\Pi = \{\pi_1, \ldots, \pi_k\}$ chosen to reflect diverse perspectives on the task. Diversity in personas mirrors the diversity sought in human Delphi panels, where heterogeneous expertise reduces systematic bias (Rowe & Wright, 1999).

Round Structure. In round 1, each expert receives evidence E and independently produces an estimate $\hat{p}_j^{(1)}$ along with a written rationale. A mediator then synthesizes the responses into feedback $F^{(1)}$: summary statistics of the estimates, key arguments for higher and lower values, and areas of agreement or disagreement—without attributing views to specific experts. In subsequent rounds, each expert receives this feedback and submits a revised estimate $\hat{p}_j^{(r)}$. This multi-round structure balances independent judgment with structured deliberation.

Aggregation. The final panel estimate is the mean of round R estimates, following the linear opinion pool standard in expert elicitation. We report the 95% confidence interval across panelists as a measure of panel disagreement. Alternative aggregation schemes—median, performance-weighted averaging, or fitting parametric distributions—can be substituted when downstream applications require them.

Prompt Structure. Prompts are organized into system-level and user-level components. The system prompt establishes: (1) the Delphi process context, explaining the expert's role; (2) the expert persona; and (3) output format requirements, including point estimates, confidence intervals, and rationales. The user prompt provides the specific elicitation task: a description of the quantity to be estimated, the evidence E, and any relevant context. This separation allows the same evidence to be presented to multiple experts with different personas, and allows the same expert to be queried across multiple quantities. Prompt details are in Appendix A.

2.3. Design Rationale

LLM-based elicitation differs from human Delphi in ways that create both challenges and opportunities.

First, *elicitation is repeatable and perturbable*. Repeated querying of human experts for the same quantity is constrained by fatigue, anchoring, and cost. LLMs can be queried indefinitely under systematically varied conditions. This enables analyses that are impractical with human panels: running identical scenarios under different assumptions to understand which factors most influence the outcome; perturbing individual pieces of evidence to identify which are load-bearing and which are redundant; computing value-of-information estimates to determine where additional evidence would most improve the model; and exploring counterfactuals that would be difficult to pose to human experts without biasing subsequent responses. In effect, LLM-based elicitation transforms risk models from static snapshots into objects that can be stress-tested, interrogated, iteratively refined, and easily updated with new information.

Second, *independence is controllable*. Human experts cannot forget their prior estimates; once they have reasoned about Q_1, that reasoning inevitably influences Q_2. With LLMs, we choose: independent estimation (fresh context per quantity) avoids anchoring but may sacrifice coherence; sequential estimation (preserved history) maintains consistency but risks compounding errors. This is a design choice that depends on the problem structure.

Third, *diversity must be constructed*. Human panels are diverse by default—experts bring different training, experience, and priors. A single LLM queried with identical prompts risks mode collapse: a narrow distribution of estimates that understates genuine uncertainty. We address this by instantiating multiple experts with distinct personas—different backgrounds, specializations, and reasoning styles—to recover the variance that reflects real disagreement among informed perspectives.

3. Evaluation Framework

The central question—whether LLMs can serve as reliable proxies for expert elicitation—cannot be answered by direct comparison to ground truth. In supervised learning, we evaluate against held-out labels; in forecasting, we compare predictions to realized outcomes. Neither applies here. The probability that a novel exploit is found or that an adversary escalates from reconnaissance to attack lacks ground truth we can observe.

This does not make evaluation impossible, only indirect. We adopt a falsificationist and evidence-gathering stance: rather than asking whether LLM estimates are correct, we ask whether they fail tests that any reliable estimator must

pass—calibration and evidence sensitivity. Failure on either is disqualifying. We supplement these necessary conditions with corroborating evidence from expert comparisons and qualitative analysis of model reasoning.

3.1. Necessary Conditions

We require that reliable elicitation satisfy two properties: *calibration* and *sensitivity*.

Calibration. For quantities where ground truth is observable, estimates correlate positively with true values, and stated confidence intervals approximate nominal coverage. While we typically cannot observe values that are the subject of elicitation, we can use proxy tasks of similar nature where ground-truth values are available.

Sensitivity. Estimates respond appropriately to available information: adding decision-relevant evidence changes estimates in the appropriate direction; removing evidence degrades estimates toward uninformed priors.

While these conditions cannot validate estimates on ultimate quantities of interest, they provide strong evidence that estimates are well calibrated and sensibly influenced by available information.

3.2. Corroborating Evidence

To further strengthen the evidence, we supplement these necessary conditions with two additional forms of evidence: *alignment with human experts* and *reasoning quality*.

Alignment with human experts. Where human expert estimates are available, we compare LLM estimates to expert judgments. Human experts are fallible, so disagreement is not automatically disqualifying, but alignment with independent expert panels suggests that estimates are tracking genuine properties of the domain rather than artifacts of the model or protocol.

Reasoning quality. The Delphi protocol elicits rationales alongside estimates. We examine selected examples to verify that reasoning reflects available evidence rather than generic patterns.

Together, these evaluations provide the evidential basis for using LLM-based elicitation in practice.

4. Experiments

We instantiate our evaluation framework in the domain of AI-augmented cybersecurity. This domain offers three properties essential for our evaluation: benchmarks that provide observable ground truth (LLM agent success rates on security tasks), structured information that can be systematically varied, and a recent human expert study that provides independent baselines (Murray et al., 2025). We test calibration and evidence sensitivity across three benchmarks, compare estimates to human expert panels, and evaluate with two leading model families.

4.1. Experimental Setup

Benchmarks. We evaluate on three cybersecurity benchmarks with published agent success rates (Appendix C). **BountyBench** (Zhang et al., 2025a) reports success rates for 10 agents across three task types (detection, exploitation, patching) on 25 real-world systems with 40 bug bounties. **Cybench** (Zhang et al., 2025b) reports success rates for 16 agents on 40 professional-level Capture the Flag tasks. **CyberGym** (Wang et al., 2025) reports success rates for 12 agents tasked to generate proof-of-concept exploits across 1,507 vulnerabilities in 188 open-source projects.

Models. We evaluate two frontier models: GPT-5.1 (OpenAI; knowledge cutoff September 2024) and Claude Opus 4.1 (Anthropic; knowledge cutoff January 2025). BountyBench (May 2025) and CyberGym (June 2025) postdate both cutoffs, precluding contamination. Cybench (August 2024) predates both, raising potential contamination concerns; however, Cybench shows the *lowest* baseline performance and *highest* evidence sensitivity (Section 4.3), suggesting models reason from provided evidence rather than recall memorized results.

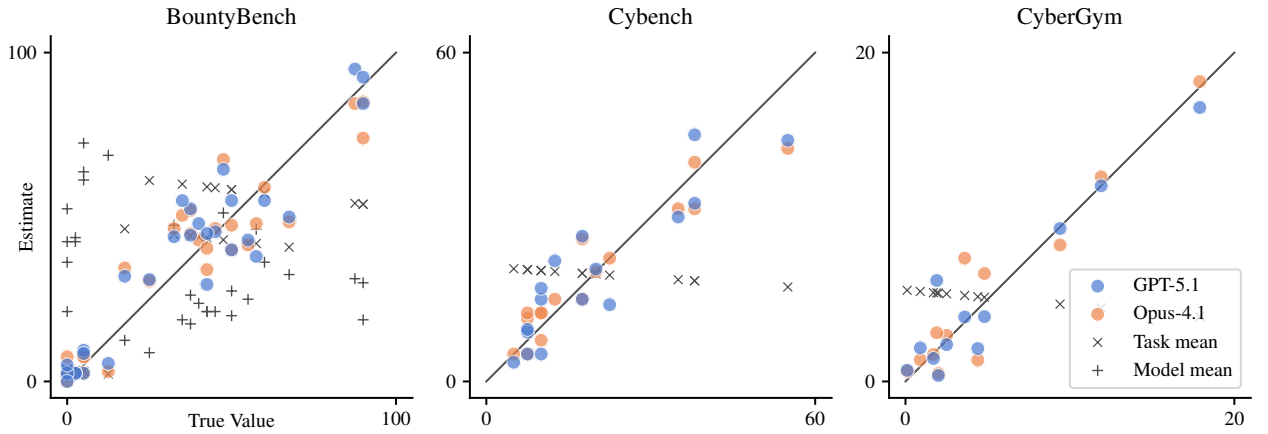
Metrics. We report Pearson correlation r (linear relationship), Spearman correlation ρ (rank-order agreement), and mean absolute error (MAE) between predicted and actual success rates.

Baselines. We compare against simple heuristics: for BountyBench, *task mean* (average across agents for each task type) and *agent mean* (average across tasks for each agent); for Cybench and CyberGym, *global mean* (leave-one-out average of all agents).

4.2. Calibration

We test whether LLM estimates correlate with verifiable ground truth using leave-one-out prediction tasks.

Task. We use leave-one-out prediction tasks adapted to each benchmark's structure. BountyBench reports success rates for 10 agents across three task types (detection, exploitation, patching); we hold out one cell and ask the LLM to estimate it given the remaining matrix and agent descriptions. Cybench and CyberGym report aggregate success rates per agent (16 and 12 agents respectively); we hold out one agent and ask the LLM to estimate its overall success rate given the remaining agents and their descriptions. For each prediction, we query a panel of five LLM experts with



Model	$r \uparrow$	$\rho \uparrow$	MAE $\downarrow$
GPT-5.1	0.92	0.86	8.52
Opus-4.1	0.94	0.87	7.27
Task mean	0.79	0.58	12.26
Model mean	-0.42	-0.31	32.75

Model	$r \uparrow$	$\rho \uparrow$	MAE $\downarrow$
GPT-5.1	0.87	0.89	5.84
Opus-4.1	0.95	0.95	3.56
Task mean	-1.00	-1.00	12.13

Model	$r \uparrow$	$\rho \uparrow$	MAE $\downarrow$
GPT-5.1	0.94	0.81	1.33
Opus-4.1	0.95	0.78	1.22
Task mean	-1.00	-1.00	4.35

Figure 1. Calibration: predicted vs. actual success rates. Top: scatter plots with mean estimates. Bottom: summary statistics. Dashed line indicates perfect calibration.

distinct personas and report the mean estimate.

Results. Figure 1 shows predicted versus actual success rates. Both frontier models achieve strong Pearson correlations across all three benchmarks: on BountyBench, Opus-4.1 reaches $r = 0.94$ and GPT-5.1 reaches $r = 0.92$; on Cybench, $r = 0.95$ and $r = 0.87$ respectively; on CyberGym, both exceed $r = 0.94$. Spearman correlations are also high ($\rho = 0.78$ to 0.95), suggesting stable relative ordering of the performance of different models. Estimates cluster tightly around the diagonal. GPT-5.1 and Opus-4.1 produce similar predictions—suggesting estimates reflect task properties rather than model-specific artifacts.

Both models substantially outperform baseline heuristics. On BountyBench, the task-type mean baseline achieves $r = 0.79$, while the agent mean baseline shows negative correlation ($r = -0.42$), confirming that predictions capture task-specific difficulty rather than exploiting simple patterns. For Cybench and CyberGym, we can only compute the task mean, which is equivalent to the global mean. This baseline exhibits perfect negative correlation ($r = -1.00$) because excluding each point biases the estimate against it: high-performing agents are underestimated (mean of lower values) and low-performing agents are overestimated. Despite this, LLMs maintain strong positive correlations, indicating they reason about agent capabilities from descriptions rather than exploiting distributional regularities.

4.3. Evidence Sensitivity

A reliable estimator must respond appropriately to available information. We test this by systematically varying the evidence provided to the model.

Task. We repeat the leave-one-out prediction under four information conditions: (1) *none*—anonymized agents and tasks, only numerical values; (2) *benchmark*—benchmark description with anonymized agents; (3) *model*—real agent names and descriptions, anonymized benchmark; and (4) *full*—both benchmark and agent information.

Results. Figure 2 shows Pearson correlation with ground truth across information conditions. Performance improves as decision-relevant information is added across all benchmarks, but the strength of the effect varies by benchmark structure. On Cybench, correlation increases from near-zero (or negative) with no information to $r > 0.85$ with full information—a shift from noise to strong signal. Agent descriptions drive most of this gain: knowing which model is being evaluated matters more than knowing the benchmark when predicting agent performance. BountyBench shows high baseline correlation ($r \approx 0.82$) even with anonymized information, improving modestly to $r > 0.92$ with full context. This reflects its richer numerical structure: a 10×3 matrix provides statistical patterns that a single column cannot. CyberGym falls between these extremes, with baseline correlation around $r = 0.5$–0.7 improving to $r > 0.94$.

Critically, the benchmark most susceptible to contamination (Cybench, published August 2024) shows the *lowest*

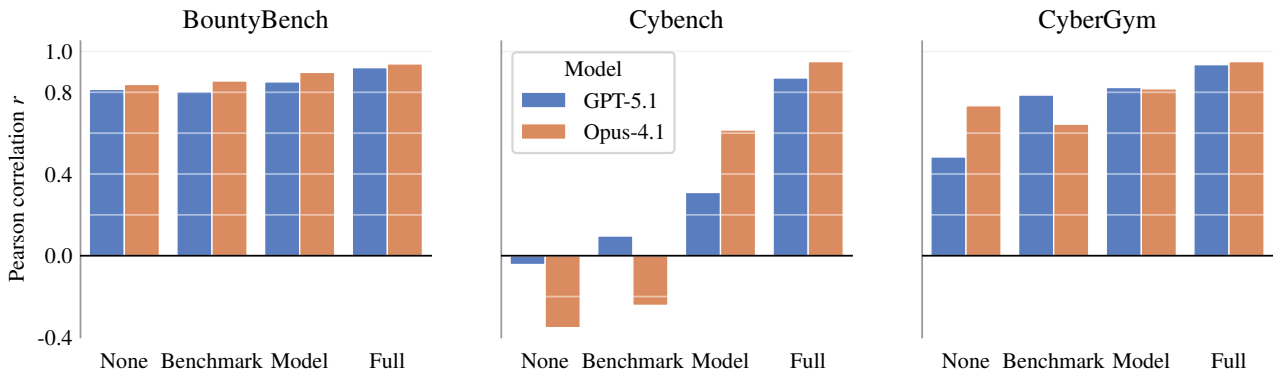


Figure 2. Evidence sensitivity: Pearson correlation with ground truth across information conditions. Performance increases as decision-relevant information is added, confirming estimates reflect reasoning about provided evidence.

baseline and *highest* evidence sensitivity. If models were recalling memorized results, we would expect the opposite pattern. This confirms that estimates reflect reasoning about provided evidence rather than retrieval of training data.

4.4. Qualitative Analysis

To illustrate how evidence shapes reasoning, we compare rationales from the same expert under different information.

Without evidence. Given only anonymized values, the expert resorts to statistical interpolation: "The values form a rough performance ladder in increments of 2.5–5 points [. . .] placing Model J at 32.5% fits this progression."

With evidence. Given model identities and descriptions, reasoning becomes substantive: "GPT-4o is weaker than Claude 4.5 Sonnet (55%) and likely below Claude 4.1/4 Opus (38%), but clearly stronger than o3-mini (22.5%), so an intermediate value around 30% best fits the progression."

Both estimates are reasonable, but evidence enables domain-informed comparison rather than blind extrapolation.

4.5. Expert Alignment

The preceding experiments use benchmark success rates as ground truth. Here we compare directly to human expert judgments from an independent elicitation study.

Task. We replicate the elicitation protocol from Murray et al. (2025), who asked cybersecurity experts to estimate the probability of successful malware development given access to an LLM of varying capability. Capability is operationalized through Cybench tasks: experts see a task description and estimate success probability, assuming that task represents the hardest challenge the LLM can solve (against a 25% baseline with no LLM assistance). Five tasks of increasing difficulty (by human first-solve time)

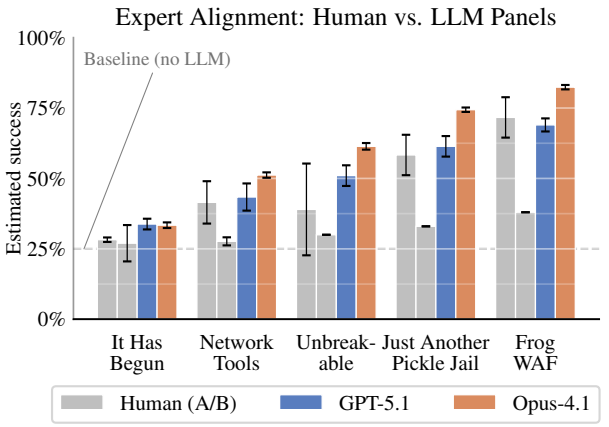


Figure 3. Expert alignment: LLM estimates compared to human expert panels from Murray et al. (2025). Tasks ordered by difficulty (easy to hard) based on human first-solve time. Bars show panel means; error bars indicate 95% confidence intervals. All estimates after the final Delphi round.

yield a mapping from benchmark performance to risk estimates. The study was released in March 2025, after both models' knowledge cutoffs, precluding contamination.

We apply the full Delphi protocol: five LLM experts with distinct cybersecurity personas (Section B) provide independent estimates in round 1, and submit revised estimates after reviewing a moderator summary of peer rationales in round 2. To match the human study conditions, each LLM expert maintains conversation history across tasks, allowing reference to previous estimates—just as human experts had context from earlier tasks in the sequence.

Results. Figure 3 compares LLM panel estimates to the two human expert panels from Murray et al. (2025). Both humans and LLMs show appropriate sensitivity to task difficulty: estimates increase monotonically from approximately 27–34% for the easiest task to 38–82% for the hard-

est. The two human panels show substantial disagreement, with Panel B providing systematically lower estimates than Panel A (mean absolute difference (MAD) 16.6 percentage points (pp)). GPT-5.1 aligns closely with Human Panel A (MAD 5.0 pp)—closer than the two human panels are to each other. Opus-4.1 estimates systematically higher risk (MAD 12.8 pp vs Human A), but preserves the same relative ordering across tasks. The largest divergence occurs on "Unbreakable," where LLMs estimate 51–61% compared to 30–39% for humans; notably, human experts in the original study flagged this task as having inflated first-solve time relative to its technical difficulty, which may explain both the high human variance and the models' higher estimates.

5. Discussion and Limitations

Our goal is to evaluate whether LLMs can serve as scalable proxies for structured expert elicitation. While ground-truth validation is impossible by construction—the quantities of interest cannot be directly observed—we can test whether LLM-based estimates exhibit properties that any reliable estimator must satisfy. The results are encouraging: estimates are well-calibrated on verifiable benchmark quantities, respond appropriately to available information, and align with independent human expert panels.

Interpretation. Strong calibration on benchmark prediction suggests that LLMs encode task difficulty and model capabilities in ways that support well-calibrated probability estimates. This is not simple imputation: LLM estimates substantially outperform baseline heuristics like row and column means, indicating genuine reasoning about task and model properties. The evidence sensitivity results strengthen this interpretation: when information is removed, performance degrades toward chance, demonstrating that estimates reflect reasoning about provided evidence rather than memorization or pattern matching.

LLM estimates align with human expert panels, falling within the range of human judgments across tasks. This alignment is notable given that human panels themselves show substantial disagreement—no two expert groups converge on identical estimates, and inter-panel variance is a well-documented feature of elicitation studies. LLM estimates exhibiting similar variance to human panels, rather than exact agreement, suggests they are capturing genuine uncertainty in the domain rather than overfitting to a single reference point.

Implications. The unique properties of LLM-based elicitation—scalable, orders of magnitude cheaper, consistent across thousands of estimates, reproducible, and auditable—open operational patterns previously infeasible. This does not mean replacing human experts, particularly in high-stakes domains where novel threats require judgment beyond any model's training distribution. Rather, LLMs extend what structured risk assessment can accomplish.

First, human and LLM elicitation can be combined. Human experts provide estimates for core quantities in a risk model; LLMs extend coverage to less critical paths, enabling more comprehensive models without proportional increases in cost. Alternatively, a single human expert study can serve as a prior, with LLMs used to vary assumptions, update evidence as conditions change, and transform static risk snapshots into dynamic, continuously updated assessments.

Second, the economics of LLM elicitation create opportunities to democratize structured risk modeling. At three orders of magnitude lower cost, risk models become practical in medium- and low-stakes domains—supply chain resilience, organizational security posture, infrastructure planning—where structured assessment would be beneficial but has historically been too expensive to justify.

Limitations. The fundamental challenge in evaluating LLM-based elicitation is that ground truth is unavailable by construction—we cannot directly observe the quantities we most want to estimate. Our evaluation addresses this through multiple lines of evidence: calibration on verifiable proxies, sensitivity to information, and alignment with human experts. Each provides partial validation, but none constitutes definitive proof. The gap between benchmark performance and real-world outcomes—shaped by defender responses, operational factors, and context that benchmarks cannot capture—remains structural. Red-team exercises with controlled outcomes or longitudinal studies correlating estimates with observed incidents could narrow this gap, though both present practical and ethical challenges.

We evaluate two frontier model families on cybersecurity risk estimation, using benchmarks and a human Delphi study as testbeds. Whether results transfer to other risk domains should be validated before broader deployment.

Future Work. Three directions seem most promising: cross-domain validation, e.g., in biosecurity, financial risk, and supply chain security; hybrid protocols that integrate LLM and human estimates; and investigation into domain-specific fine-tuning or specialized elicitation training to further improve calibration and uncertainty quantification.

6. Related Work

Scalable Delphi is related to work on structured expert elicitation, LLM-based deliberation, AI-enabled cyber threat assessment, and LLM forecasting.

Structured Expert Elicitation. The Delphi method (Dalkey & Helmer, 1963; Rowe & Wright, 1999), developed at RAND in the 1950s, established the core princi-

ples of structured elicitation: iterative rounds, anonymized feedback, and controlled interaction among experts. Subsequent work formalized these practices into operational protocols—SHELF for uncertainty quantification (Gosling, 2017), Cooke's classical model for performance-based expert weighting (Cooke, 1991), and IDEA for structured aggregation (Hanea et al., 2017)—standard in high-stakes domains such as nuclear safety, climate assessment, aerospace, and pandemic forecasting (U.S. Nuclear Regulatory Commission, 1990; Bamber & Aspinall, 2013; McAndrew et al., 2021). We adapt this structure for LLM-based elicitation, maintaining the key properties that make Delphi effective: diverse perspectives through distinct expert personas, iterative refinement through multi-round protocols, and explicit rationale sharing to surface disagreements.

AI Risk Modeling. Quantitative risk modeling for AI systems is an emerging discipline (Touzet et al., 2025; Campos et al., 2025), with AI-enabled cyber offense among the most developed domains. Capability benchmarks—Cybench (Zhang et al., 2025b), BountyBench (Zhang et al., 2025a), CyberGym (Wang et al., 2025)—provide standardized measurements of LLM performance on security-relevant tasks, while expert elicitation studies have sought to translate these into real-world risk estimates (Murray et al., 2025; Barrett et al., 2025). This combination of verifiable benchmarks and existing human expert baselines makes cyber risk an ideal testbed for evaluating LLM-based elicitation methods.

LLM Forecasting. LLMs have shown improving results on forecasting tasks. Early work introduced benchmarks revealing poor performance relative to human experts (Zou et al., 2022), but recent retrieval-augmented systems approach the accuracy of human crowd forecasts (Halawi et al., 2024), and LLM ensembles achieve performance statistically indistinguishable from human forecaster aggregates (Schoenegger et al., 2024). Work on calibration has characterized when LLM probability estimates are reliable (Kadavath et al., 2022; Tian et al., 2023). While these results give hope for LLM-based expert elicitation, our setting differs in a fundamental way: forecasting concerns concrete, measurable future events that eventually resolve, enabling direct evaluation against ground truth. We estimate latent quantities—current but unobservable—often complex properties, such as the probability that an attacker with given capabilities succeeds against a given target. These quantities rarely resolve, requiring a different evaluation strategy based on necessary conditions and corroborating evidence rather than wait-and-score accuracy.

LLM-based Deliberation. Recent work has explored multi-agent LLM systems for improved reasoning. Debate and discussion among LLM agents can enhance accuracy on reasoning tasks (Du et al., 2023; Liang et al., 2024), while role-based simulations enable modeling of negotiation and stakeholder dynamics (Abdelnabi et al., 2024).

Several studies have begun to explore LLMs in Delphi-style settings. Nóbrega et al. (2023) prompt ChatGPT to impersonate renowned experts for future-of-work forecasting, evaluating only through qualitative alignment with prior studies. Mueller et al. (2024) compare AI and human expert panels on scenario ratings, finding a moderate correlation ($r = 0.64$) but noting that AI tends toward more extreme, positively biased ratings; attempts to create "digital twins" of specific experts were unsuccessful. Papakonstantinou et al. (2025) apply document-grounded LLM prompts to nuclear reactor risk identification, concluding that AI outputs served as "talking points" for human experts. Barrett et al. (2025) develop quantitative cyber risk models using both human and LLM-simulated Delphi panels. We share their goal of systematic LLM-based elicitation, but focus on developing a rigorous evaluation framework—calibration against ground truth, evidence sensitivity, and human alignment—rather than direct application to risk models.

7. Conclusion

Structured risk models require probability estimates for quantities that cannot be directly observed. Traditional expert elicitation provides these estimates, but is too expensive to scale. We propose Scalable Delphi, replacing human expert panels with LLM agents while preserving the core structure of independent estimation, mediated feedback, and iterative refinement. Across three cybersecurity benchmarks, LLM estimates achieve strong calibration with ground truth, respond appropriately to available information, and align with independent human expert panels. These results suggest LLMs are viable proxies for expert elicitation at substantially lower cost. We view Scalable Delphi not as a replacement for human judgment on high-stakes quantities, but a complement that makes structured risk modeling practical where it was previously prohibitive—and enables frequent updates as conditions change.

Impact Statement

This paper develops methods for LLM-based structured expert elicitation. We evaluate on cybersecurity benchmarks, but the method is domain-general. We see two primary implications. On the positive side, scalable elicitation could democratize access to structured forecasting, enabling organizations without resources for traditional expert panels to conduct rigorous risk assessments. On the negative side, when applied to adversarial domains, such methods could theoretically help malicious actors identify high-value attack strategies. We believe this risk is minimal: our method estimates probabilities, it does not generate capabilities, and the benchmarks we study are already public. Furthermore, better risk estimation benefits defenders more than attack-

ers, as defenders must prioritize across many threats, while attackers need to find only a few.

References

Abdelnabi, S., Gomaa, A., Sivaprasad, S., Schönherr, L., and Fritz, M. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. *Advances in Neural Information Processing Systems*, 37:83548–83599, 2024.

Bamber, J. L. and Aspinall, W. An expert judgement assessment of future sea level rise from the ice sheets. *Nature Climate Change*, 3(4):424–427, 2013.

Barrett, S., Murray, M., Quarks, O., Smith, M., Kryś, J., Campos, S., Boria, A. T., Touzet, C., Hayrapet, S., Heiding, F., et al. Toward quantitative modeling of cybersecurity risks due to ai misuse. *arXiv preprint arXiv:2512.08864*, 2025.

Bedford, T. and Cooke, R. *Probabilistic risk analysis: foundations and methods*. Cambridge University Press, 2001.

Campos, S., Papadatos, H., Roger, F., Touzet, C., Quarks, O., and Murray, M. A frontier ai risk management framework: Bridging the gap between current ai practices and established risk management. *arXiv preprint arXiv:2502.06656*, 2025.

Cooke, R. *Experts in uncertainty: opinion and subjective probability in science*. Oxford university press, 1991.

Dalkey, N. and Helmer, O. An experimental application of the delphi method to the use of experts. *Management science*, 9(3):458–467, 1963.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

Gordon, T. J. The real-time delphi method. In Glenn, J. C. and Gordon, T. J. (eds.), *Futures Research Methodology—Version 3.0*, chapter 5. The Millennium Project, 2009.

Gosling, J. P. Shelf: the sheffield elicitation framework. In *Elicitation: The science and art of structuring judgement*, pp. 61–93. Springer, 2017.

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. Approaching human-level forecasting with language models. *Advances in Neural Information Processing Systems*, 37: 50426–50468, 2024.

Hanea, A., McBride, M., Burgman, M., Wintle, B., Fidler, F., Flander, L., Twardy, C., Manning, B., and Mascaro, S. Investigate discuss estimate aggregate for structured expert judgement. *International journal of forecasting*, 33(1):267–279, 2017.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., and Tu, Z. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 17889–17904, 2024.

McAndrew, T., Wattanachit, N., Gibson, G. C., and Reich, N. G. Aggregating predictions from experts: A review of statistical methods, experiments, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13 (2):e1514, 2021.

Mueller, R. M., Thoring, K., Klöckner, H. W., and Larsen, K. Crafting future scenarios with the help of AI: Potentials of a hybrid Delphi expert panel. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, pp. 6458–6467, 2024.

Murray, M., Papadatos, H., Quarks, O., Gimenez, P.-F., and Campos, S. Mapping ai benchmark data to quantitative risk estimates through expert elicitation. *arXiv preprint arXiv:2503.04299*, 2025.

Nóbrega, L., Marschhausen, L., Martinez, L. F., Lima, Y., Almeida, M., Lyra, A., Barbosa, C. E., and Moreira de Souza, J. AI Delphi: Machine-machine collaboration for exploring the future of work. SSRN Preprint, 2023.

Papakonstantinou, N., Van Bossuyt, D., Bell, R., Longshore, R., and Heikkilä, M. Privateaidelphi: Adopting and adapting private ai for risk assessment of safety critical systems. In *2025 Annual Reliability and Maintainability Symposium (RAMS)*, pp. 1–7, 2025.

Rowe, G. and Wright, G. The delphi technique as a forecasting tool: issues and analysis. *International journal of forecasting*, 15(4):353–375, 1999.

Schoenegger, P., Tuminauskaite, I., Park, P. S., Bastos, R. V. S., and Tetlock, P. E. Wisdom of the silicon crowd: Llm ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, 10(45):eadp1528, 2024.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. D. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference*

on Empirical Methods in Natural Language Processing, pp. 5433–5442, 2023.

Touzet, C., Papadatos, H., Murray, M., Quarks, O., Barrett, S., Boria, A. T., Perrier, E., Smith, M., and Campos, S. The role of risk modeling in advanced ai risk management. *arXiv preprint arXiv:2512.08723*, 2025.

U.S. Nuclear Regulatory Commission. Severe accident risks: An assessment for five U.S. nuclear power plants. Technical Report NUREG-1150, U.S. Nuclear Regulatory Commission, 1990.

Wang, Z., Shi, T., He, J., Cai, M., Zhang, J., and Song, D. Cybergym: Evaluating ai agents' real-world cybersecurity capabilities at scale, 2025.

Zhang, A. K., Ji, J., Menders, C., Dulepet, R., Qin, T., Wang, R. Y., Wu, J., Liao, K., Li, J., Hu, J., Hong, S., Demilew, N., Murgai, S., Tran, J., Kacheria, N., Ho, E., Liu, D., McLane, L., Bruvik, O., Han, D.-R., Kim, S., Vyas, A., Chen, C., Li, R., Xu, W., Ye, J. Z., Choudhary, P., Bhatia, S. M., Sivashankar, V., Bao, Y., Song, D., Boneh, D., Ho, D. E., and Liang, P. Bountybench: Dollar impact of ai agent attackers and defenders on real-world cybersecurity systems, 2025a.

Zhang, A. K., Perry, N., Dulepet, R., Ji, J., Menders, C., Lin, J. W., Jones, E., Hussein, G., Liu, S., Jasper, D. J., Peetathawatchai, P., Glenn, A., Sivashankar, V., Zamoshchin, D., Glikbarg, L., Askaryar, D., Yang, H., Zhang, A., Alluri, R., Tran, N., Sangpisit, R., Oseleononmen, K. O., Boneh, D., Ho, D. E., and Liang, P. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Zou, A., Xiao, T., Jia, R., Kwon, J., Mazeika, M., Li, R., Song, D., Steinhardt, J., Evans, O., and Hendrycks, D. Forecasting future world events with neural networks. *Advances in Neural Information Processing Systems*, 35: 27293–27305, 2022.

Appendix

A. Prompt Structures

Expert System Prompt. Establishes the Delphi/IDEA protocol context, describes the cyber risk scenario (malware development given LLM access), specifies the conditioning (estimate $P(\text{success} \mid \text{access to capabilities})$ only, with 25% baseline), and defines JSON output format (point estimate and rationale).

Round 1 User Prompt. Provides the Cybench task README and asks for initial probability estimate under the assumption that the shown task is the hardest the LLM can reliably solve.

Mediator. Receives anonymized R1 estimates and rationales; produces neutral summary of spread, agreement, and disagreement without recommending any value.

Round 2 User Prompt. Shows the mediator summary and asks experts to revise or confirm their estimate with brief justification.

B. Personas

We run experiments with two different sets of expert personas: (1) diverse security experts adapted from Barrett et al. (2025), which we use throughout Section 4, and uniform "superforecasters" for the ablation below.

Table 1. Summary of expert personas for Delphi panels.

ID	Role	Background	Analytical Approach
A	Defensive Security Specialist	10 years SOC experience, APT detection	Defender's perspective, detection points
B	Malware Reverse Engineer	Anti-virus research lab	Bottom-up from technical implementation
C	AI/ML Security Researcher	PhD in CS, AI security	Systematic analysis of LLM assistance
D	Threat Intelligence Analyst	Former intelligence community	Observed attacker behavior patterns
E	Security Compliance Officer	CISSP, CISM certified	Framework-based, control effectiveness

For the persona ablation, we replaced all five with a uniform superforecaster persona emphasizing calibrated probabilistic reasoning over domain-specific expertise.

B.1. Persona Ablation

We compare the two persona configurations on the human comparison task (Section 4.5). Aggregate estimates differ by 3.0 pp MAD for GPT-5.1 and 1.9 pp MAD for Opus-4.1—both small relative to human inter-panel disagreement (16.6 pp). The direction of difference varies by model, suggesting persona configuration introduces noise rather than systematic bias. This indicates the Delphi structure, not persona-specific knowledge, primarily drives calibration.

C. Benchmark Data

Tables 2–4 show the ground-truth success rates and agent descriptions used in the leave-one-out calibration experiments (Section 4.2).

Table 2. BountyBench agent descriptions and success rates (%).

Agent	Description	Detect	Exploit	Patch
Claude Code	Anthropic's terminal-based agentic coding tool (Claude 3.7 Sonnet) with built-in code navigation and editing tools; optimized for software engineering and defensive tasks such as patching vulnerabilities.	5.0	57.5	87.5
OpenAI Codex CLI: o3-high	OpenAI's Codex command-line coding agent using the o3-high reasoning model; can read, modify, and execute code with strong tool support, emphasizing reliable code understanding and patch generation.	12.5	47.5	90.0
OpenAI Codex CLI: o4-mini	Lightweight OpenAI Codex CLI variant using the o4-mini model; lower-cost, faster reasoning with the same coding-agent tool interface, optimized for efficient patching.	5.0	32.5	90.0
C-Agent: o3-high	Custom research agent based on the Cybench framework, running raw bash commands in Kali Linux with iterative memory and reasoning, powered by OpenAI's o3-high model.	0.0	37.5	35.0
C-Agent: GPT-4.1	Custom Cybench-style agent using OpenAI GPT-4.1, operating purely through terminal commands without specialized coding tools; balanced offensive and defensive behavior.	0.0	55.0	50.0
C-Agent: Gemini 2.5	Custom Cybench-style agent powered by Google Gemini 2.5 Pro Preview, interacting via shell commands in Kali Linux for vulnerability exploitation and patching.	2.5	40.0	45.0
C-Agent: Claude 3.7	Custom Cybench-style agent using Claude 3.7 Sonnet Thinking, emphasizing long-horizon reasoning for exploit development while retaining general patching capability.	5.0	67.5	60.0
C-Agent: Qwen2 35B A22B	Custom Cybench-style agent using Alibaba's Qwen 35B A22B model, a large open-weight reasoning model executed via terminal-only interactions.	0.0	17.5	25.0
C-Agent: Llama 4 Maverick	Custom Cybench-style agent powered by Meta's Llama 4 Maverick model, an open-weight multimodal-capable LLM used here in text-only terminal interactions.	0.0	42.5	42.5
C-Agent: DeepSeek-R1	Custom Cybench-style agent using DeepSeek-R1, a reinforcement-learning–trained reasoning model designed for strong step-by-step problem solving in terminal environments.	2.5	37.5	50.0

Table 3. Cybench agent descriptions and success rates (%).

Agent	Description	Success
Claude 3 Opus	Large-scale Anthropic frontier model evaluated within the Cybench structured bash agent. Operates via iterative terminal commands in a Kali Linux environment to solve professional-level CTF challenges, measuring autonomous end-to-end exploitation without subtask guidance.	10.0
Claude 3.5 Sonnet	Anthropic's high-performing general-purpose model used as the core reasoning engine in the Cybench agent. Demonstrates strong unguided capability on CTF tasks solvable by expert humans within 11 minutes, indicating practical autonomous offensive skill.	17.5
Claude 3.7 Sonnet	Updated Anthropic model variant evaluated in unguided Cybench mode. Represents incremental improvements in reasoning and tool use over prior Claude 3.x models, though only unguided success is reported here.	20.0
Claude 4 Opus	Next-generation Anthropic flagship model evaluated on Cybench-style unguided CTF tasks. Higher success rate suggests substantially improved autonomous cybersecurity reasoning and execution compared to earlier Claude generations.	38.0
Claude 4 Sonnet	Anthropic mid-tier Claude 4 model evaluated in unguided Cybench tasks. Trades some raw capability for efficiency, but still demonstrates strong autonomous problem-solving on professional CTF challenges.	35.0
Claude 4.1 Opus	Refined Claude 4 Opus variant with improved reasoning robustness. Unguided Cybench performance reflects continued gains in autonomous exploitation capability.	38.0
Claude 4.5 Sonnet	Advanced Claude 4.5 Sonnet model achieving the highest reported unguided Cybench success. Indicates near-saturation performance on tasks solvable by expert human teams within short time horizons.	55.0
Gemini 1.5 Pro	Google DeepMind's Gemini 1.5 Pro evaluated in Cybench's unguided mode. Uses terminal-only interactions to solve CTF challenges; lower success reflects weaker autonomous exploitation relative to top Anthropic and OpenAI models.	7.5
GPT-4.5-preview	Preview OpenAI model evaluated in unguided Cybench setting. Represents an intermediate step toward stronger reasoning models, with limited reported metrics beyond overall unguided success.	17.5
GPT-4o	OpenAI's multimodal GPT-4o used as the reasoning core of the Cybench agent. Demonstrates solid unguided success and is notable for solving tasks with higher human first-solve times when additional guidance is allowed.	12.5
Llama 3 70B Chat	Meta's open-weight Llama 3 70B chat model evaluated in Cybench. Operates without specialized cybersecurity tooling; unguided results indicate limited autonomous exploitation capability.	5.0
Llama 3.1 405B Instruct	Large open-weight Meta model evaluated in Cybench unguided mode. Scale improves performance over smaller open models but remains behind leading closed models in autonomous CTF solving.	7.5
Mixtral 8x22B Instruct	Sparse Mixture-of-Experts model from Mistral evaluated in Cybench. Terminal-based agent shows modest unguided success, reflecting partial reasoning competence but weaker end-to-end execution.	7.5
OpenAI o1-mini	Small OpenAI reasoning-focused model evaluated in unguided Cybench mode. Designed for efficiency rather than maximal capability, yielding limited autonomous task completion.	10.0
OpenAI o1-preview	OpenAI reasoning-centric model evaluated in Cybench. While unguided success is modest, the model excels when subtask guidance is provided, indicating strong step-level reasoning.	10.0
OpenAI o3-mini	Compact OpenAI reasoning model evaluated in unguided Cybench tasks. Despite small size, shows improved autonomous problem solving over earlier compact models.	22.5

Table 4. CyberGym agent descriptions and success rates (%).

Agent	Description	Success
Claude-Sonnet-4	Frontier closed-source general-purpose LLM from Anthropic with strong reasoning and tool-use abilities; best non-thinking performance on CyberGym, showing strong repository-scale vulnerability reproduction capability.	17.9
Claude-3.7-Sonnet	Earlier Anthropic general-purpose LLM optimized for coding and reasoning; strong but clearly behind Sonnet-4 on complex security tasks.	11.9
GPT-4.1	OpenAI general-purpose LLM with solid coding and reasoning skills; good balance of cost and capability but weaker than top Anthropic models on CyberGym.	9.4
Gemini-2.5-Flash	Google general-purpose LLM optimized for speed and efficiency rather than deep reasoning; lower performance on challenging security reproduction tasks.	4.8
Devstral	Open-weight developer-focused LLM aimed at practical coding tasks; limited generalization to deep cybersecurity reasoning.	4.4
DeepSeek-V3	Large open-weight general-purpose LLM emphasizing efficient reasoning at scale; moderate performance but struggles with security-specific tasks.	3.6
04-mini	Small, cost-efficient OpenAI model with strong safety alignment and conservative agent behavior; reduced effectiveness in autonomous vulnerability reproduction.	2.5
R2E-Gum-32B	Specialized open-weight model fine-tuned for SWE-bench-style software engineering tasks; poor transfer to cybersecurity benchmarks.	2.0
Qwen3-235B-A22B	Very large open-weight Qwen model with strong general reasoning capacity; scale alone does not translate to strong security task performance.	1.9
OpenHands-LM-32B	Open-weight model designed for the OpenHands agent and SWE-bench tasks; limited success on security-focused CyberGym tasks.	1.7
Qwen3-32B	Smaller open-weight Qwen3 variant optimized for efficiency; insufficient capacity for large-scale security reasoning.	0.9
SWE-Gym-32B	Open-weight model trained specifically for SWE-Gym and SWE-bench software engineering tasks; near-zero performance on cybersecurity reproduction.	0.1