Know-It-All?: A Human-Calibrated LLM Benchmark for Cybersecurity Knowledge

Athanasios Theocharis, Armin Buescher, Omer Akgul, Dario Pasquini,
Oystein Fladby, Dan Marino, Christopher Gates, Petros Efstathopoulos
RSAC
{*thanos.theocharis, armin.buescher, omer.akgul, dario.pasquini,*
oystein.fladby, dan.marino, chris.gates, petros.efstathopoulos}*@onersac.com*

***Abstract*—Large Language Models (LLMs) have rapidly advanced in cybersecurity knowledge, yet large-scale standardized benchmarking against cybersecurity professionals remains limited. To address this gap, we introduce a human-curated and difficulty-calibrated cybersecurity benchmark covering 21 subtopics. To enable a practical human performance baseline (e.g., reducing fatigue and attention effects), we design a gamified platform and deploy it at a large cybersecurity conference. Across all topics and difficulty bands, we find that LLMs (n=39) outperform humans (n=279) with an average failure rate of 12.60% vs. 32.87%. The largest difference was observed in domains that require precise technical recall, such as applied cryptography and infrastructure security. We discuss implications for cybersecurity work and LLM benchmarking.**

***Index Terms*—Large language models, LLM benchmarking, human evaluation**

1. Introduction

Large Language Models (LLMs) are becoming increasingly capable at specialized tasks, including cybersecurity analysis and decision support. Standardized cybersecurity benchmarks, however, remains limited and ad-hoc, complicating evaluation, procurement, and safe deployment in security-critical contexts [1], [2]. This necessitates the development of rigorous, broad-ranging, standardized benchmarks. Such benchmarks could highlight which models work best under different cyber tasks (e.g., vulnerability detection), leading to better task-model fit and overall performance in cybersecurity AI products.

Existing benchmarking efforts make important progress but exhibit (one or more of) three main limitations: 1) narrow focus [3], [4], [5], [6], [7] 2) no human review of LLM generated questions [8], [9], 3) no [4], [8], [9] or limited ($n \leq 30$) human baseline [5], [10].

The most widely used (e.g., by [1], [2]) benchmark, CTI-Bench [4], is focused only on threat intelligence and lacks a human baseline. Further, our qualitative review shows several questions that are not directly security-related (e.g., *"In terms conducting data correlation using statistical data analysis, which data correlation technique is a nonparametric analysis, which measures the degree of relationship between two variables?"*).

Most relevant to our work, [10] develop a cyber focused benchmark using LLMs and retrieval on a set of cyber-related documents and include a human baseline on 80 questions, but with a limited participant pool (n=30).

In contrast, in this work, after creating an LLM and human reviewed benchmark of 625 questions across 21 subtopics, we run the largest scale knowledge-based benchmarking study in the security domain. We measure the performance of 39 LLMs and establish the largest human baseline (n=279) in this domain ($\sim 9\times$ more than [10]) due to our gamified approach. Our benchmark involved two distinct groups of security professionals: (i) a team of *domain experts* (n=10, the authors and colleagues, each with $\geq$3 years of cybersecurity experience) who curated, reviewed, and calibrated the benchmark questions, and (ii) *cybersecurity professionals* at RSAC 2025 who played the gamified quiz and provided the human baseline data (n=279). We find that LLMs outperform cybersecurity professionals across nearly all topic categories and difficulty bands of our on cyber-knowledge benchmark and discuss implications of our work.

Among the 21 topics covered, our benchmark has particular relevance to software security, with 118 questions directly spanning DevSecOps & Application Security, Product Security, Open Source Tools, and Supply Chain Security that evaluate knowledge required for secure development and vulnerability management. These are augmented by adjacent questions on supply chain policy, vulnerability management and patching, and cloud and container security.

The paper is organized as follows: section 2 describes how we create the benchmark, section 4 details the game used to establish a human baseline, section 5 describes the results, and in section 6 we discuss the implications of this work. The benchmark is made public as part of this publication.[1]

2. Developing the benchmark

Figure 1 shows an overview of the main components of the benchmark developments process, which we detail in this section.

1. Available at https://osf.io/v72hc/overview?view_only=8fe8b0f798774dd8a66082611f48b989

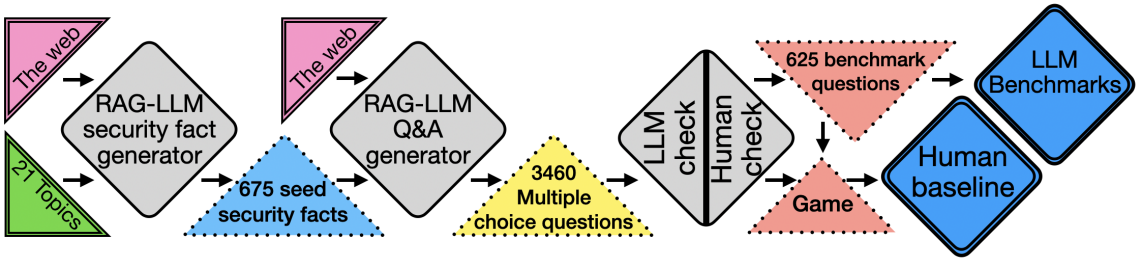


Figure 1: Question generation, benchmarking, and human baseline pipeline.

Topic Selection and Coverage. The 21 benchmark topics were derived from the official taxonomy of the RSAC 2025 Conference (the largest global gathering in cybersecurity, and the venue we collected human data through). Together, these 21 topics provide nearly complete coverage of the cybersecurity domain and accurately capture the background and expertise of the conference attendees (i.e., the participants in our study). Topics are visible in Figure 3. Grouped under higher-level categories, the most common are: Governance and Compliance (Policy & Government, Law, Privacy, Risk Management & Governance, Business Perspectives), Software Security (DevSecOps & Application Security, Product Security, Open Source Tools, Protecting Data & the Supply Chain Ecosystem), and Security Operations & Defense (Analytics Intelligence & Response, Hackers & Threats, Security Strategy & Architecture, Technology Infrastructure & Operations). Each topic seed served as an initial prompt for the Q&A generation pipeline, ensuring that the resulting benchmark questions align with real-world security knowledge.

Question generation. For each of the 21 domains, we used a retrieval-augmented generation (RAG) pipeline on the whole web to collect short, source-grounded seed facts that constrain question content. Each seed represents a general subtopic within a broader domain. For example, a seed generated for the domain *"Analytics Intelligence & Response"* is *"Managing Alert Fatigue"*.

From each seed, we use a RAG-based web pipeline to gather resources tailored to that seed, explicitly avoiding commercial or vendor-oriented material. For seeds for which we can retrieve sufficient context (at least five relevant web pages), we use an LLM to synthesize one or more grounded four-option multiple-choice questions (MCQs) with a single correct answer. To increase the likelihood of generating challenging items, after question generation, we employ a second LLM to produce three semantically plausible distractors. Prompts explicitly forbade unsupported content and required correctness to be decidable from the seed evidence. The LLMs used for synthetically generating questions and answers were `o1-2024-12-17`, `gemini-2.5-pro-exp-03-25`, and `claude-3-7-sonnet-20250219`.

Curation. The heavy use of LLMs might have introduced unique biases and artifacts in the candidate questions generated. Thus, to reduce these issues, all questions and answer sets underwent (i) an automated LLM pass to flag multiplicity, null-answer, contradiction, or ambiguity errors; and (ii) review by the domain expert team of 10 (the authors and colleagues) with the possibility of revision or removal of each item.

$\leq$ **4B** Param.	$>$ **4B** $\leq$ **15B** Param.	$>$ **15B** Param.
qwen1.5-0.5b-chat	llama-7b	mistral-small-3.1-24b-inst.
qwen2-0.5b-instruct	qwen1.5-7b-chat	qwen2.5-32b-instruct
phi-1.5	qwen2.5-3b-instruct	qwen2.5-72b-instruct
qwen2.5-0.5b-instruct	mistral-7b-instruct-v0.2	mistral-large-instruct-2411
qwen1.5-1.8b-chat	mistral-7b-instruct-v0.3	llama-3.3-70b-instruct
llama-3.2-1b-instruct	llama-3.1-8b-instruct	gemma-3-27b-it
gemma-2-2b-it	qwen2-7b-instruct	gpt-5.2
qwen2.5-1.5b-instruct	qwen2.5-7b-instruct	-
qwen2-1.5b-instruct	gemma-2-9b-it	-
phi-3.5-mini-instruct	mistral-7b-v0.1	-
phi-4-mini-instruct	phi-4	-
phi-3-mini-4k-instruct	gemma-7b-it	-
llama-3.2-3b-instruct	gemma-3-4b-it	-
-	llama-2-13b-chat	-
-	qwen3-8b-instruct	-
-	ministral-3-8b-instruct	-
-	gemma-3-12b-it	-
-	qwen3-14b-instruct	-
-	ministral-3-14b-instruct	-

TABLE 1: Evaluated LLMs grouped by size category.

Questions were then assigned difficulty scores using a model calibrated with security-expert preferences. Each question is evaluated by an LLM on four factors, each scored from 1 to 5: *cognitive complexity*, *knowledge specificity*, *trickiness*, and *multi-domain integration*. We then involved the domain expert team of 10 security professionals (the authors and a few colleagues, with a minimum of three years of cybersecurity experience) who provided pairwise difficulty preferences between questions. This human feedback was used to learn a set of weights applied to the LLM-generated factor scores, such that their weighted sum yields a difficulty score that best aligns with expert preferences. Figure 5 shows that human success rates decrease as hardness increases, validating our difficulty assignments. An intermediary step contained 3,460 validated MCQs spanning all topics and difficulty bands. The final benchmark contains a topic and difficulty balanced subset of 625 MCQs, individually re-verified by the domain expert team. This final dataset was used for both benchmarking LLMs and obtaining the human baseline.

3. Benchmarking LLMs

We evaluated a broad spectrum of openly available language models spanning multiple architectures, parameter scales, and release periods to compare across multiple variants. We primarily evaluated publicly released, locally executable models to ensure reproducibility; we additionally included GPT-5.2 to broaden coverage and use as a reference point, noting that such models may change behavior due to quantization, watermarking, and system prompt changes invisible to end-users [11]. In practice, newer releases frequently outperformed older counterparts of similar size, while several domain-fine-tuned models lagged behind their generic versions, highlighting that recency and dataset quality often outweighed narrow specialization.

During evaluation, all models were evaluated zero-shot with a fixed prompt template and randomized option order. Outputs were constrained to a single-letter (A–D), reducing verbosity-induced variance.

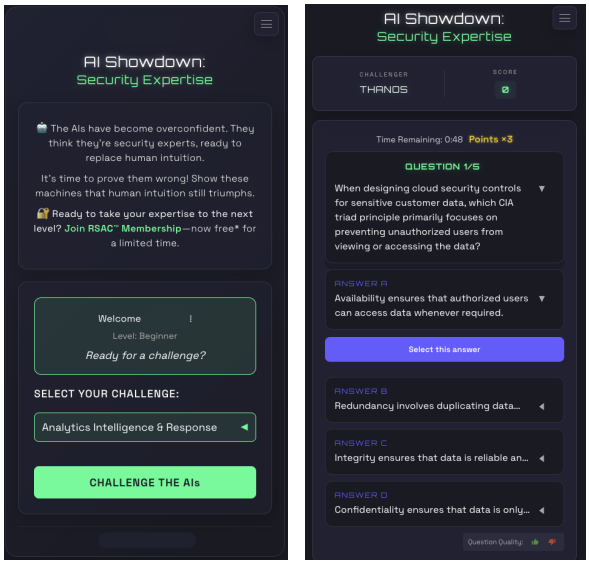
4. Human Baseline via Gamified Approach

We measured human performance with a gamified quiz at a major cybersecurity conference. Here, we describe the process.

Human–Model Comparison Methodology. To obtain a realistic, controlled baseline of human cybersecurity knowledge on the *same multiple-choice items* used for model benchmarking, we deployed a game during RSAC 2025 as part of a research agreement, ensuring participants would have cybersecurity-relevant background.[2] The game was designed as a timed, mobile/online, multiple-choice quiz with four possible answers, randomized item ordering, and no backtracking. Disabled copy/paste and time limits minimized chances of using online help. While participants answered questions, the platform recorded per-question correctness and response latency. We defined accuracy as the proportion of correctly selected options and *failure rate* as 1 - accuracy, and we aggregated responses into topic and difficulty based (1-10 scale) failure rates. Humans and models were compared on the same items.

RSAC Quiz Environment and Game Design. Substantial empirical evidence has built up on the engagement and cognitive benefits of gamification (points, taunts, leaderboards) on a large variety of tasks [14], [15]. As such, the system was designed to resemble a security challenge rather than a traditional survey, increasing engagement while ensuring high-quality measurement.

When players accessed the platform, they were shown an introductory interface presenting the game narrative (taunting security experts that AI would outperform them) and



(a) Introductory game UI. (b) Response collection UI.

Figure 2: Screenshots from the game UI.

instructions. During signup, users were asked for consent to collect non-identifying metadata, including country of residence, years of cybersecurity experience, and self-declared expertise level. Then, players selected a challenge category corresponding to one of the 21 benchmark topics. A leaderboard provided real-time cumulative participant scores, further encouraging engagement. The game is available at https://haiqu.onersaclabs.com.

Each session consisted of five multiple-choice questions sampled from the benchmark pool, selected by topic and difficulty. The quiz interface (Fig. 2) presented one question at a time, with four fixed options (A–D), a countdown timer, and no backtracking. This prevented iterative external lookup and incentivized engagement. The game recorded both accuracy and per-item response latency. After each response, users were shown whether their selection was correct, a brief explanatory note, providing an educational element, and a taunt or celebratory note that aimed to entertain and emphasize the human-versus-ai element of the game.

Participants could repeatedly play the game with a different set of questions. A separate profile page allowed returning players to view their accumulated statistics, including accuracy and scores distribution across topics.

Participation This design enabled rapid, large-scale collection of answers from 279 cybersecurity professionals at RSAC 2025. Participants collectively submitted 2439 answers to 559 (out of 625) questions spanning all 21 topics. RSAC attendees have an average of 10 years of cybersecurity industry experience [13], making our participants a well-experienced sample. However, we note that our participant pool is a convenience sample of conference attendees who voluntarily opted into the game; it is not a random draw from all cybersecurity practitioners, and individual-level expertise

2. RSAC Conference attracts approximately 44,000 cybersecurity professionals from around the world [12]. According to the conference organizers, session attendees have an average of 10 years of security industry experience, and the audience is described as "well beyond entry-level," with intermediate-to-advanced content most well-received [13]. Attendees include CISOs, security architects, compliance directors, DevSecOps leads, researchers, policymakers, and vendors.

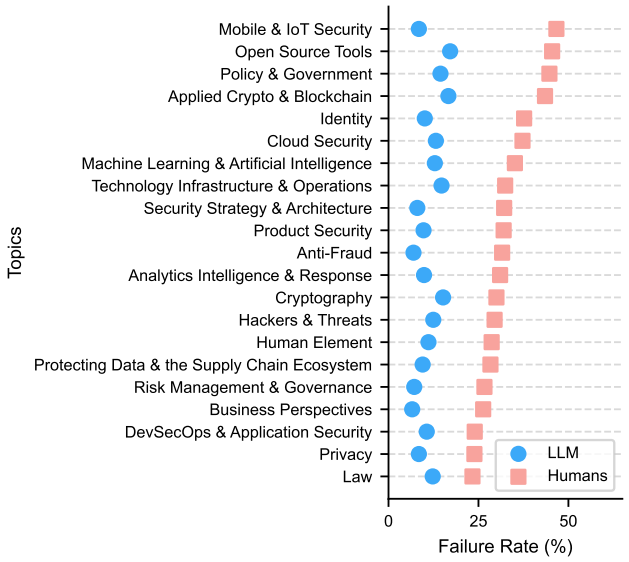


Figure 3: LLMs vs. Humans failure rate by topic.

may vary.

Ethics Our study received ethics and legal approval from our institution. All participants were informed about data collection for this study and an opt-out option was given. The questions were vetted for potentially offensive content during the human review stage.

5. Results

Overall performance. Across 21 topical categories, LLMs (n=39) consistently outperformed human cybersecurity professionals (n=279) in correctly answered questions (see Figure 3 and Figure 5). Model failure rates ranged from **6.7–17.0%** (median **10.8%**, mean **12.6%**), whereas human failure rates spanned roughly **19–50%**. This pattern held uniformly: no category showed humans outperforming models, and notably, the best human category (Law at 19.0%) performed worse than the worst LLM category (Open Source Tools at 17.0%).

Topic-wise analysis. Lowest model failure rates were observed in *Business Perspectives* (**6.7%**), *Anti-Fraud* (**6.9%**), *Risk Management & Governance* (**7.1%**), *Security Strategy & Architecture* (**8.2%**), and *Mobile & IoT Security* (**8.5%**). Most challenging for models were *Open Source Tools* (**17.0%**), *Applied Crypto & Blockchain* (**16.7%**), *Cryptography* (**15.9%**), *Technology Infrastructure & Operations* (**15.0%**), and *Policy & Government* (**14.9%**). The largest absolute gap appeared in *Mobile & IoT Security* (humans $\approx$**50%** vs. models **8.5%**).

Comparing extremes. At the individual level, top humans matched top models, with both achieving 100% on at least one category. However, humans who reached perfect scores did so only sporadically and in areas aligned with their personal expertise, whereas LLMs achieved perfect performance across a broader range of topics, suggesting a more generalized model knowledge.

Conclusions. On this curated, difficulty-calibrated knowledge benchmark, LLMs surpassed participating cybersecurity professionals on multiple-choice cybersecurity knowledge questions across *all* topic categories and across *all* hardness categories we studied. Model errors were tightly dispersed across topics, suggesting broad coverage rather than niche topic expertise. Human errors varied more substantially, with steeper failures clustering in technical subdomains (e.g., *Applied Cryptography*) and evolving practices (e.g., *Policy and Government*).

6. Discussion

Our findings show that LLMs are increasingly more capable in cybersecurity knowledge. Further, on average, models significantly outperform humans on these questions. However, models are not perfect, showing the importance of model-task match in cyber products.

Limitations Our study has limitations. First, the multiple-choice (MCQ) format may not fully capture cybersecurity judgment in all settings, and overrepresent knowledge-based tasks. More broadly, the MCQ format may structurally favor LLMs relative to humans. MCQs test *recognition*, selecting the correct answer from presented options, rather than *recall*, generating answers from memory [16]. MCQ formats can inflate LLM (and potentially human) scores due to this fact; when the same questions are posed in open-ended format, LLM accuracy drops substantially [17]. Furthermore, the conditions under which humans and LLMs answered questions were asymmetric: human participants answered under time pressure at a conference while LLMs answered without time constraints, however, LLM responses are near instantaneous. In practice, cybersecurity professionals routinely consult documentation and collaborate with peers. Our results are therefore more reflective of cybersecurity knowledge under constrained testing conditions than of operational cybersecurity skill, and the observed performance gap may narrow under open-ended, tool-assisted, or collaborative settings. Second, like prior work (e.g., [10]) we make heavy use of LLMs in our question generation pipeline (factoids, Q&A generation, filtering, difficulty calculation), potentially creating questions that LLMs might have an advantage in answering [18]. We partially mitigate MCQ and generation concerns by covering a broad range of topics, randomizing option order, and conducting multiple rounds of human expert review to ensure technical correctness while reducing generation artifacts and ambiguity. Third, our participant sample is not a random draw from all practitioners, and self-reported data may introduce measurement bias. While RSAC attendees are generally senior cybersecurity professionals (averaging 10 years of industry experience [13]), the conference population also includes vendors, students (e.g., RSAC College Day), and executives, and our game participants may not be representative of the broader conference population or of cybersecurity professionals as a

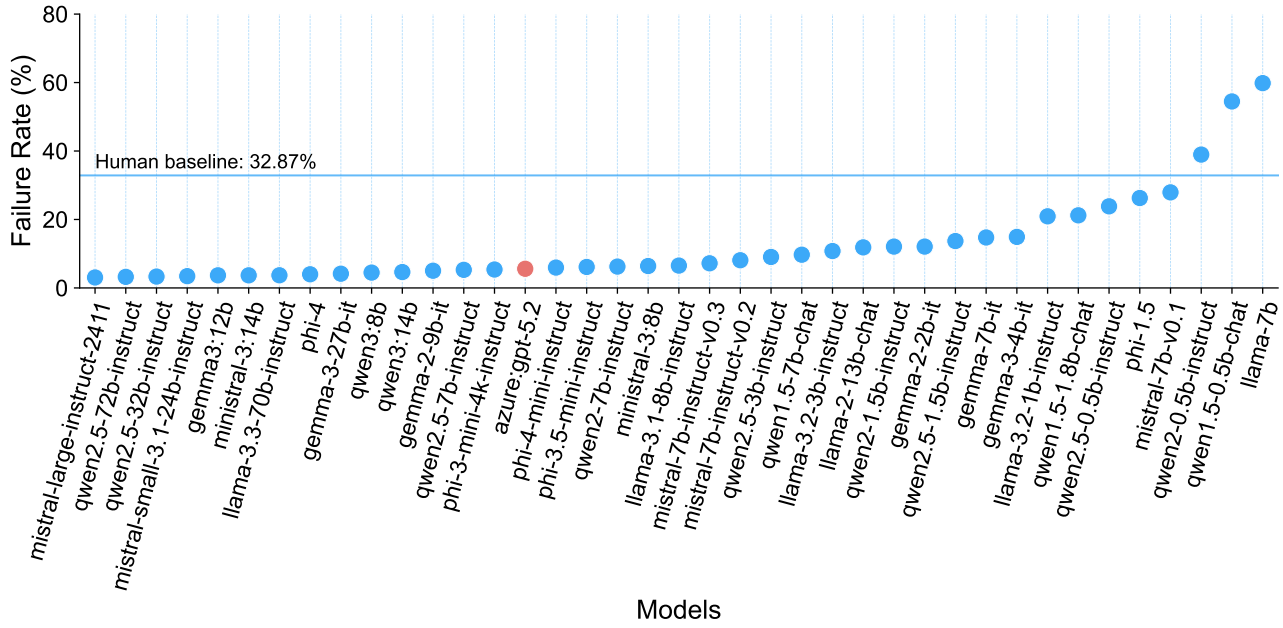


Figure 4: Models' failure rates on the benchmark. All are open-weight except GPT-5.2 (red), which is shown for reference.

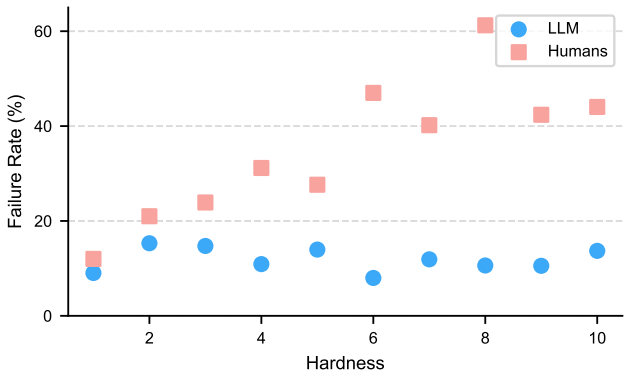


Figure 5: LLMs vs. Humans failure rate by hardness.

whole. However, the sample is large ($n = 279$) and drawn from the world's largest cybersecurity conference, providing a practical and informative estimate of human performance on this benchmark.

Need for skill-based benchmarking. As cyber-capable and cyber-focused LLMs become more widespread ([1], [2]), the need for benchmarking such models increases substantially. Like the benchmark introduced in this work, most benchmarks are primarily knowledge based questionnaires. While this testing strategy satisfies the crucial need for fast testing and practical human benchmarking, it diverges from some real world cyber tasks (e.g., penetration testing) in important ways [19]. However, skill based benchmarks (e.g., [3], [6], [7]) do not measure cyber capabilities holistically, knowledge-based cyber work is still highly prevalent, suggesting the need for hybrid benchmarks. Despite these limitations, our methodology for benchmark construction, human difficulty calibration, and scalable collection of human comparison data provides a practical foundation for future cybersecurity evaluation, including extensions to more open-ended and task-based settings.

How to choose which model to use. The choice of models in security products should be guided by task- and domain-specific evaluation of the models. While one model could be particularly promising in security-oriented code review, another might be more capable in compliance review. Capability evaluation might also support model routing. Simpler tasks could be routed to smaller but less capable models, while complex tasks are assigned to the most capable ones.

References

[1] E. Burzstein and M. Tishchenko. (2025) Google announces sec-gemini v1, a new experimental cybersecurity model. Google Online Security Blog. Accessed November 12, 2025. [Online]. Available: https://security.googleblog.com/2025/04/google-launches-sec-gemini-v1-new.html

[2] M. Levi, D. Ohayon, A. Blobstein, R. Sagi, I. Molloy, and Y. Allouche, "Toward cybersecurity-expert small language models," *preprint arXiv:2510.14113*, 2025.

[3] A. K. Zhang, N. Perry, R. Dulepet, J. Ji, C. Menders, J. W. Lin, E. Jones, G. Hussein, S. Liu, D. Jasper, P. Peetathawatchai, A. Glenn, V. Sivashankar, D. Zamoshchin, L. Glikbarg, D. Askaryar, M. Yang, T. Zhang, R. Alluri, N. Tran, R. Sangpisit, P. Yiorkadjis, K. Osele, G. Raghupathi, D. Boneh, D. E. Ho, and P. Liang, "Cybench: A framework for evaluating cybersecurity capabilities and risks of language models," 2025. [Online]. Available: https://arxiv.org/abs/2408.08926

[4] M. T. Alam, D. Bhusal, L. Nguyen, and N. Rastogi, "Ctibench: A benchmark for evaluating llms in cyber threat intelligence," *Advances in Neural Information Processing Systems*, 2024.

[5] D. Bhusal, M. T. Alam, L. Nguyen, A. Mahara, Z. Lightcap, R. Frazier, R. Fieblinger, G. L. Torales, B. A. Blakely, and N. Rastogi, "Secure: Benchmarking large language models for cybersecurity," in *2024 Annual Computer Security Applications Conference (ACSAC)*. IEEE, 2024.

[6] M. Vetzler, N. Ohfeld, and A. Schindel, "Introducing AI cyber model arena: A real-world benchmark for AI agents in cybersecurity," Wiz Blog, Feb. 2026. [Online]. Available: https://www.wiz.io/blog/introducing-ai-cyber-model-arena-a-real-world-benchmark-for-ai-agents-in-cybersec

[7] b3rt0ll0, "Benchmarking AI security: Inside the new HTB AI range," Hack The Box Blog, Dec. 2025. [Online]. Available: https://www.hackthebox.com/blog/ai-range-and-model-evaluations

[8] G. Li, Y. Li, W. Guannan, H. Yang, and Y. Yu, "Seceval: A comprehensive benchmark for evaluating cybersecurity knowledge of foundation models," https://github.com/XuanwuAI/SecEval, 2023.

[9] Z. Liu, "Secqa: A concise question-answering dataset for evaluating large language models in computer security," *preprint arXiv:2312.15838*, 2023.

[10] N. Tihanyi, M. A. Ferrag, R. Jain, T. Bisztray, and M. Debbah, "Cybermetric: a benchmark dataset based on retrieval-augmented generation for evaluating llms in cybersecurity knowledge," in *2024 IEEE International Conference on Cyber Security and Resilience (CSR)*. IEEE, 2024.

[11] I. Gao, P. Liang, and C. Guestrin, "Model equality testing: Which model is this api serving?" *preprint arXiv:2410.20247*, 2024.

[12] RSAC, "Frequently asked questions," 2025, accessed: 2026-03-30. [Online]. Available: https://www.rsaconference.com/about/faq

[13] RSAC, "Call for submissions faq," 2025, accessed: 2025-11-17. [Online]. Available: https://www.rsaconference.com/usa/call-for-submissions/faq

[14] J. Hamari, J. Koivisto, and H. Sarsa, "Does gamification work?–a literature review of empirical studies on gamification," in *2014 47th Hawaii international conference on system sciences*. Ieee, 2014.

[15] J. A. Yip, M. E. Schweitzer, and S. Nurmohamed, "Trash-talking: Competitive incivility motivates rivalry, performance, and unethical behavior," *Organizational Behavior and Human Decision Processes*, 2018.

[16] M. E. Martinez, "Cognition and the question of test item format," *Educational Psychologist*, vol. 34, no. 4, pp. 207–218, 1999.

[17] A. Myrzakhan, S. M. Bsharat, and Z. Shen, "Open-LLM-leaderboard: From multi-choice to open-style questions for LLMs evaluation, benchmark, and arena," *arXiv preprint arXiv:2406.07545*, 2024.

[18] A. Panickssery, S. R. Bowman, and S. Feng, "Llm evaluators recognize and favor their own generations," in *NeurIPS*, 2024.

[19] Cybersecurity and Infrastructure Security Agency, "Cybersecurity talent identification and assessment," U.S. Department of Homeland Security, Tech. Rep., 2019.