Evaluating AI cyber capabilities with crowdsoursed elicitation

Artem Petrov,[1] Dmitrii Volkov

2025-04-25

As AI systems become increasingly capable, understanding their offensive cyber potential is critical for informed governance and responsible deployment. However, eliciting AI capabilities is hard.

XXX

In this paper, we explore crowdsourcing elicitation efforts as an alternative to in-house elicitation work. We present findings from the first open *AI vs. Humans* CTF competition, where AI teams competed against each other and human hackers. To extend this evaluation, we also organized an AI track as part of *Cyber Apocalypse*, an annual CTF event with thousands of participants.

exact number

Top AI teams effectively saturated our task set, completing 19 out of 20 challenges and ranking in the top 13% overall, while being as fast as the best multi-human teams. We further adapted METR's (Kwa et al. 2025) 50%-task-completion time horizon metric and found that AI agents perform as well on cyber tasks as on the general tasks: they can reliably solve challenges requiring one hour or less of effort from a median human participant.

XXX

We suggest that elicitation bounties may offer an effective alternative to in-house elicitation. Evaluations, grounded in the performance of human participants, also promise a more interpretable alternative to traditional benchmark scores, enhancing accessibility to both policymakers and the general public.

1. Introduction

Eliciting AI capabilities is hard. Yet accurately estimating them is important, especially when it comes to dangerous capabilities, as it determines whether a model is safe to release. We hypothesize that evaluations of offensive cyber capabilities may often be underelicited - that is, they may fail to fully reveal the system's potential, as we discuss in Section 5.

We propose crowdsourcing elicitation efforts as a strategy to boost cyber performance and close the growing mismatch between the reliance on safety evaluations for policy decisions and the ability of existing evaluations to robustly identify capabilities of existing and near-future models ("The Evals Gap" 2025). To explore this approach, we hosted a first-of-a-kind *AI vs. Humans* Capture The Flag (CTF) competition, inviting AI developers to compete against each other and human teams. We then additionally analyzed AI performance in *Cyber Apocalypse* - a large CTF competition with thousands of participants.

In this paper, we report on the performance AI exhibited in these events and compare it to human performance.

2. *AI vs. Humans* CTF

To explore crowdsourcing as an elicitation approach in practice, we organized *AI vs. Humans* CTF competition. CTF competitions are cybersecurity events where both professionals and enthusiasts tackle challenges in areas such as cryptography, reverse engineering, and web exploitation. Each challenge hides a "flag" — a unique string — that must be found by identifying and exploiting system vulnerabilities.

The event was organized in collaboration with *Hack The Box* in the period of 14-16 March 2025[2]. This event was the first to publicly pit fully autonomous AI agents against experienced human teams in real-time, tackling 20 cybersecurity challenges over 48 hours. We offered a monetary prize of 7500$ to incentivize participation and effort.

For the pilot event, we wanted to make it as easy as possible for the AI teams to compete. The event focused on challenges in the cryptography and reverse engineering, as these areas are the easiest to build a harness for. The challenges only required running commands in the terminal, without the need for dynamic interactions with external machines.

[1]Correspondence to artem.petrov@palisaderesearch.org CC ctf-event@palisaderesearch.org

[2]https://ctf.hackthebox.com/event/details/ai-vs-human-ctf-challenge-2000

2.1. Absolute standings

Overall 403 teams registered for the event, of which only 158 solved at least 1 challenge - 152 human teams and 6 AI teams.

Palisade Research team ran 2 AI agents during this event - our adaptation of Claude Code for CTFs and our in-house React&Plan agent design (Turtayev et al. 2024).

Notably, AI teams significantly exceeded our initial expectations and the challenges were quickly saturated. Best AI team achieves top-13% performance among teams that solved at least 1 challenge. Four out of seven agents completed 19/20 challenges. You can see the final AI teams standings in Table 1.

Our expectations for AI performance had been based on preliminary evaluations of our React&Plan agent on older *Hack The Box*-style tasks. The challenge difficulty was calibrated based on these prior results with expectations of AI solving ~50% of tasks. Turns out AI could do way better. This outcome suggests that conventional evaluation methods by small teams (like ours) may fail to fully surface the capabilities of current AI models, highlighting the need for more dynamic and diversified approaches.

D, are we allowed to say this?

Table 1: AI agents standings.[3]

Agent	**Challenges solved**	**Leaderboard rank**[4]
CAI	19 / 20	20
Palisade Claude Code	19 / 20	21
FCT	19 / 20	30
imperturbable	19 / 20	33
Cyagent	18 / 20	34
Project S1ngularity	14 / 20	65
Palisade React&Plan	14 / 20	66

2.2. Speed

One of the core advantages AIs typically hold over humans is speed. As shown in Figure 1, AI teams performed on par with top human teams in both speed across almost all tasks.

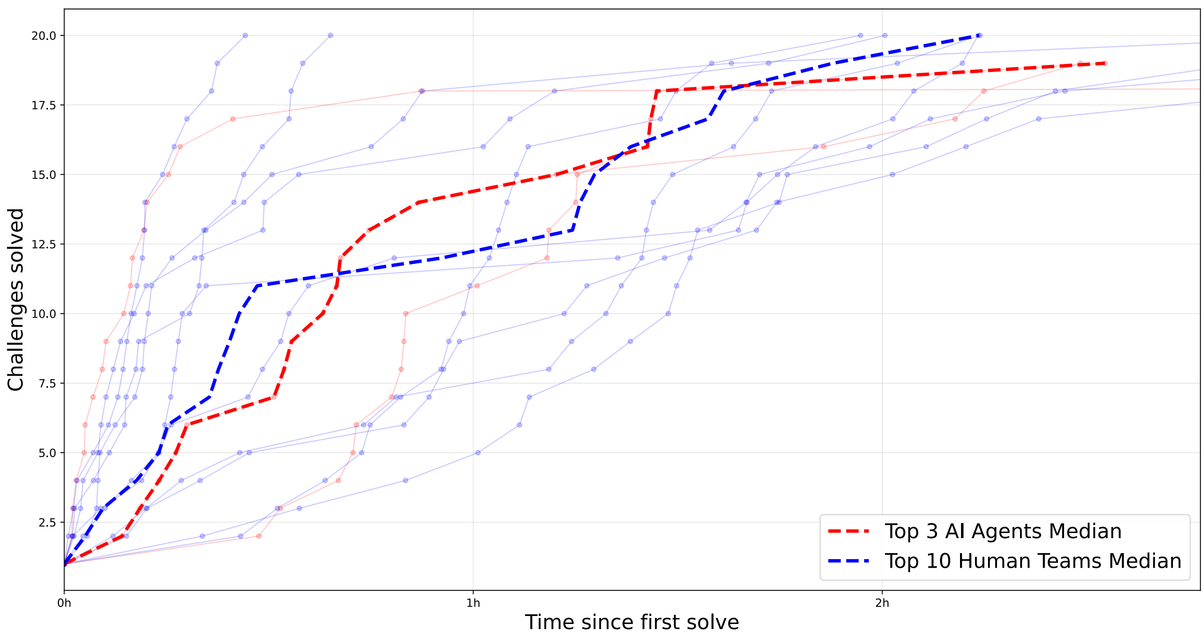

Figure 1: Number of challenges solved over time for the *AI vs Humans* CTF (top teams)

[3] The full leaderboard, including human teams, is available at https://ctftime.org/event/2723

[4] In case of teams solving an equal number of challenges, the faster teams rank higher on the leaderboard.

The exceptional speed of human teams surprised us, so we reached out to five top human teams for comments. Participants attributed their ability to solve the challenges quickly to their extensive experience as professional CTF players, noting that they were familiar with the standard techniques commonly used to solve such problems. One participant, for example, said that he was "playing on a couple of internationally-ranked teams with years of experience".

For details on AI agent designs, please see Appendix A.

You can see more plots in Appendix C.

We make all performance data available at TODO PUBLIC REPO LINK.

3. *Cyber Apocalypse*

On 21–26 March 2025, *Hack The Box* invited us to participate in another CTF event. *Cyber Apocalypse* is an annual competition that attracts a large number of human participants and proposes a diverse set of challenges[5].

Overall, 3994 human teams solved at least one challenge. We invited the AI teams from the *AI vs. Humans* CTF event to participate in this competition to evaluate their performance in a different environment. Ultimately, two AI teams took part, collectively deploying four AI agents. AI teams achieved top-21% performance among teams that solved at least one challenge. You can see the final AI standings in Table 2.

Palisade's submissions performed poorly, because our harness was not designed to interact with external machines, while about 2/3 of challenges required it.

As this competition challenges were not saturated by AI, the resulting data allowed us to calculate the 50%-completion-time horizon, discussed in the next section.

Table 2: AI agents standings for Cyber Apocalypse[6].

Agent	Challenges solved	Leaderboard rank
CAI	20 / 62	859
Palisade Claude Code	5 / 62	2496
Palisade Aider	3 / 62	2953
Palisade React&Plan	2 / 62	3199

[5] *Cyber Apocalypse* event page: https://www.hackthebox.com/events/cyber-apocalypse-2025
[6] The full leaderboard, including human teams, is available at https://ctftime.org/event/2674

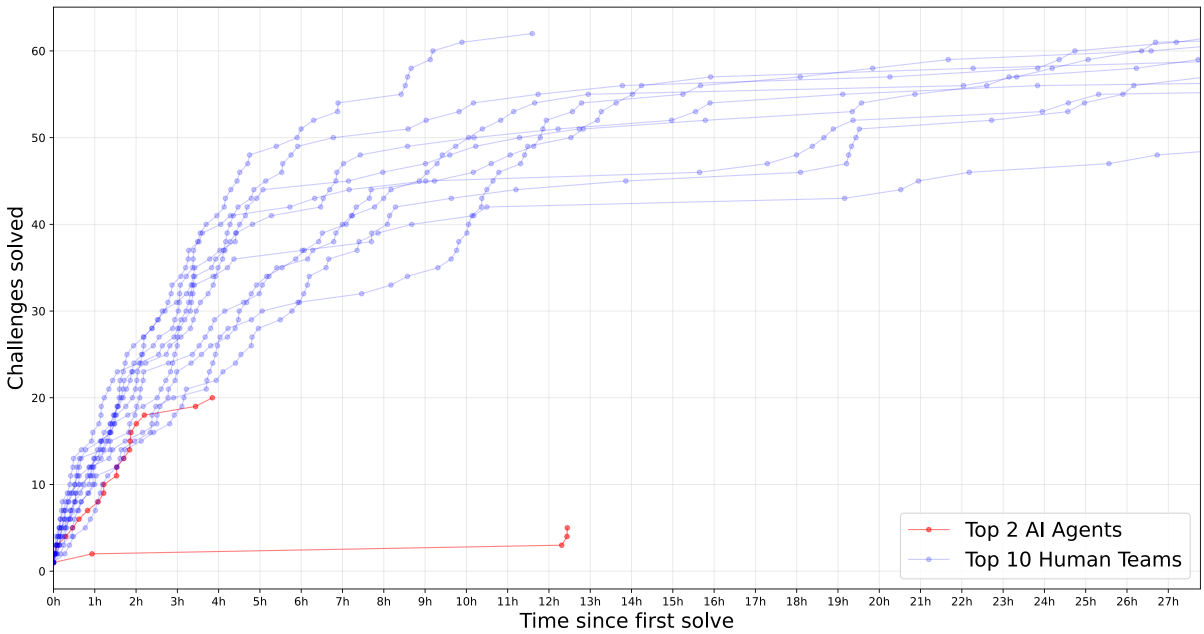

Figure 2: Number of challenges solved over time for *Cyber Apocalypse* CTF.

4. Ability to complete long tasks

Modern AIs are known to struggle with tasks that require staying coherent on long timescales. A recent study by *METR* has shown that modern AIs can reliably complete tasks requiring up to 1 hour of human expert effort (Kwa et al. 2025).

To estimate the human expert effort equivalent to current AI capabilities, we follow (Kwa et al. 2025) by measuring the 50%-task-completion time horizon. This is a metric that indicates the time humans typically take to complete tasks that AI models can complete with a 50% success rate.

Analyzing the data from *Cyber Apocalypse* we reach a similar outcome - AI can solve challenges requiring ~1 hour of effort from a median participant. See Appendix B for details.

5. Discussion

5.1. Cyber may be underelicited

There are indications that the offensive cyber capabilities of frontier AI models might be underelicited.

Previous research, including our own, has shown that LLMs are better at cybersecurity problems than previously thought. In our recent study (Turtayev et al. 2024), using some straightforward prompting and agent design strategies, we were able to achieve state-of-the-art performance on a popular offensive security benchmark, *InterCode-CTF*, with a 95% success rate — surpassing previously reported results.

This finding builds on a broader pattern of evidence suggesting that cyber evaluations often fail to fully activate the capabilities of advanced models. For instance, *Project Naptime* (Project Zero 2024) pointed out the

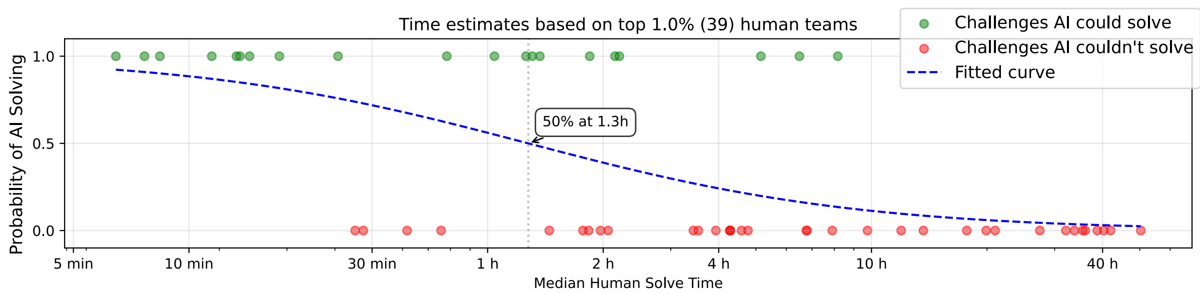

Figure 3: 50%-task-completion time horizon calculation based on *Cyber Apocalypse* data.

elicitation deficiencies in *Meta's CyberSecEval 2* (Bhatt et al. 2024). By modifying the agent harness in relatively simple ways, the *Project Naptime's* team was able to boost the *CyberSecEval 2* benchmark performance by up to 20x from *Meta's* reported results - from 0.05 to 1.0 on the "Buffer Overflow" tests and from 0.24 to 0.76 on the "Advanced Memory Corruption" tests.

Finally, *DeepMind's* report on the model's hacking capabilities (Phuong et al. 2024) showed *Gemini-1.0* solved 24 out of 81 *InterCode-CTF* tasks, which is lower than the benchmark's baseline of 40% (Yang et al. 2023). This may have happened because their models were weaker, or because of insufficient elicitation.

These examples, point to a consistent theme: sometimes modifying harness design can unlock significantly better performance than reported by frontier AI models' internal evaluation teams. To improve robustness and realism in AI cybersecurity evaluations, the field may benefit from crowdsourcing elicitation strategies, reducing reliance on any single team's assumptions.

5.2. Challenges of recruiting AI teams

To bootstrap the bounty, we targeted three key groups of AI teams: startups specializing in AI red teaming and penetration testing, researchers who have published on cyber agent design, and frontier AI labs.

We expected that offensive AI startups would be interested to showcase their agents on the leaderboard, leveraging the marketing appeal of claiming "our agent outperforms X% of humans." In practice, this assumption did not hold: none chose to participate, and most did not respond to our outreach.

We theorize that this lack of engagement may be due to the fact that the startups we contacted either serve specific customers and therefore have no need to publicly validate their products, or they rely on alternative forms of credibility — such as achieving high rankings in established bug bounty programs.

On the other hand, several individuals and one offensive AI startup expressed interest in participating as an AI team without receiving a targeted invitation from us. You can see the entire flow in Table 3.

This was our first attempt at crowdsourcing in this context, and while it offered valuable insights, we recognize the need for further experimentation — potentially including a higher bounty to better incentivize participation.

this section only focuses on difficulties of recruiting/attracting AI teams to participate in crowdsourcing efforts. I think, it would be beneficial to also discuss whether crowdsourcing work as a method, explaining why we believe crowdsourcing might be a more efficient approach - answering the question in the section's title

Table 3: AI teams registration flow for the AI vs Humans event.

AI team type	Invited by us	Accepted the invite	Expressed interest independently	Participated in the event
Offensive AI startups	9	0	1	1 (CAI)
Researchers	3	1	0	1 (Cyagent)
Frontier Labs	2	1	0	1 (FCT)
Individuals	0	0	3	2 (Imperturbable, Project S1ngularity)

for max pretty this can be a sankey diagram

6. Conclusion

Accurately assessing the offensive capabilities of AI systems remains a major challenge for policymakers and researchers. Relying solely on evaluations conducted by internal teams at frontier AI labs may be insufficient. To better understand the highly dynamic nature of AI capabilities, we need a more diverse and robust evaluation approach grounded in real-world, real-time testing. Toward this goal, we hosted a first-of-its-kind *AI vs. Humans* CTF competition[7], inviting AI developers to directly compete against each other and against human teams. We also analyzed AI performance in CyberApocalypse, a large CTF with thousands of participants.

[7]The only other similar competition we know of was held as part of DARPA's AI Cyber Challenge (AIxCC), it had a different focus on advancing AI for cyber defense. See: https://aicyberchallenge.com

While our results are still preliminary, our initial experiments suggest that using CTF competitions to elicit offensive AI capabilities holds significant promise in several directions.

Firstly, the crowdsourcing approach demonstrated potential for surfacing meaningful behaviors in AI systems in practical settings, exceeding our initial expectations of AI capabilities. In the *AI vs. Humans* CTF competition, our top-performing AI team ranked in the top 13% of all participants, with several AI teams solving 19 out of 20 challenges — effectively saturating the task set. By directly comparing with human performance we highlight current system capabilities in a way which is more tangible for policymakers and general public than reporting AI achieving a score on a cyber benchmark. However, further data and repeated trials are needed to fully assess its robustness and effectiveness of crowdsourcing elicitation.

To further ground AI performance relative to human capabilities, we adopted the 50%-task-completion time horizon metric proposed by *METR* (2025). We found that AI agents were able to solve cyber tasks requiring up to one hour of effort from a median human CTF participant. This measure, adapted here for evaluating offensive capabilities, provides a more interpretable and policy-relevant signal than aggregate benchmark scores.

Moreover, we suggest that CTFs for AI elicitation is both cost-effective and practical, particularly for organizations operating with limited resources. We believe this mechanism can be easily replicated and scaled across future evaluations.

We suggest several ways in which crowdsourcing can offer practical value for different stakeholders:

- **For cybersecurity professionals: DIMA TODO**
- **For policy and R&D agencies**: CTFs already provide the infrastructure — what is missing is targeted support for AI-focused tracks. Modest funding for prizes and coordination could establish a sustainable evaluation ecosystem. Once popularized, these tracks may require little or no ongoing funding.
- **For frontier AI labs**: Public evaluations offer a fast, low-cost way to uncover overlooked capabilities and validate internal assessments.
- **For CTF Organizers**: Adding an AI track can increase visibility, attract participants, and introduce new research and media interest. Our team has experience supporting these efforts and is available to assist with design and execution.

As AI systems grow in complexity, evaluation must keep pace. Moreover, evaluation methods should be clear enough to guide policy, and flexible enough to reveal what today's systems can actually do. This research is our initial contribution toward that goal.

7. Acknowledgements

We thank our *Palisade Research* colleagues: Reworr, who helped set up and run the Claude Code agent, achieving second place among AI teams; Rustem Turtayev, who helped set up the React&Plan agent; and Alexander Bondarenko, for experimenting with the Aider agent.

We thank the AI teams who designed agents for this competition: CAI (Víctor Mayoral-Vilches, Endika Gil-Uriarte, Unai Ayucar-Carbajo, Luis Javier Navarrete-Lozano, María Sanz-Gómez, Lidia Salas Espejo, Martiño Crespo-Álvarez), Imperturbable (James Lucassen), FCT, Cyagent, and Project S1ngularity.

We are grateful to *Hack The Box* for providing challenges for the *AI vs Humans* CTF event, supporting us in organizing this event, and providing data for our analysis. This event would not happen without their help!

References

Bhatt, Manish, Sahana Chennabasappa, Yue Li, Cyrus Nikolaidis, Daniel Song, Shengye Wan, Faizan Ahmad, et al. 2024. "Cyberseceval 2: A Wide-Ranging Cybersecurity Evaluation Suite for Large Language Models". 2024. https://arxiv.org/abs/2404.13161.

Kwa, Thomas, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, et al. 2025. "Measuring AI Ability to Complete Long Tasks". 2025. https://arxiv.org/abs/2503.14499.

Mayoral-Vilches, Víctor, Luis Javier Navarrete-Lozano, María Sanz-Gómez, Lidia Salas Espejo, Martiño Crespo-Álvarez, Francisco Oca-Gonzalez, Francesco Balassone, et al. 2025. “CAI: An Open, Bug Bounty-Ready Cybersecurity Ai”. 2025. https://arxiv.org/abs/2504.06017.

Phuong, Mary, Matthew Aitchison, Elliot Catt, Sarah Cogan, Alexandre Kaskasoli, Victoria Krakovna, David Lindner, et al. 2024. “Evaluating Frontier Models for Dangerous Capabilities”. arXiv. https://doi.org/10.48550/arXiv.2403.13793.

Project Zero. 2024. “Project Naptime: Evaluating Offensive Security Capabilities of Large Language Models.”

Turtayev, Rustem, Artem Petrov, Dmitrii Volkov, and Denis Volk. 2024. “Hacking Ctfs with Plain Agents”. 2024. https://arxiv.org/abs/2412.02776.

Yang, John, Akshara Prabhakar, Shunyu Yao, Kexin Pei, and Karthik R Narasimhan. 2023. “Language Agents as Hackers: Evaluating Cybersecurity Skills with Capture the Flag”. In *Multi-Agent Security Workshop @ Neurips'23*. https://openreview.net/forum?id=KOZwk7BFc3.

“The Evals Gap.” 2025. 11, 2025. https://www.apolloresearch.ai/blog/evalsgap.

Appendix

A Agent designs

Below are some of the agent designs used by AI teams.

CAI:

The best performing team. They have a custom harness design they spent about 500 dev-hours on. You can read about it in their paper. (Mayoral-Vilches et al. 2025)

Palisade Claude Code:

We adapted Claude Code for solving CTFs by writing a simple initial prompt.

Our claude code (+ Imperturbable enigma + claude code)

imperturbable:

From the participant:

```
I spent 17 dev-hours on agent design

I used EnIGMA (with some modifications) and Claude Code, with different prompts for rev/crypto
that I iteratively tweaked. Most prompt tweaks were about:
- preventing the model from trying to guess the flag based on semantics
- making sure the model actually carefully inspects the task before coming up with its
strategy
- recommending particular tools that were easier for the LLM to use.

You can find the spreadsheet I used to track my progress here:

https://docs.google.com/spreadsheets/d/1sdMC2AFwZ131vNG-iHyiDL0j2nyqa3gV5fN234QkXxk/edit?usp=
sharing
```

Palisade React&Plan:

The agent our team designed while working on (Turtayev et al. 2024)

B Measuring 50%-task-completion time horizon

To estimate the human expert effort equivalent to current AI capabilities we follow (Kwa et al. 2025) by measuring the 50%-task-completion time horizon. This is the time humans typically take to complete tasks that AI models can complete with 50% success rate.

Hack The Box estimates difficulty of the challenges by measuring how long it takes a median participant from first accessing the challenge data (by downloading challenge files or starting the docker container) to submitting a flag. We adopt this approach to measure human time spent solving a challenge.

When measuring "human expert performance" it is important to know whom we consider an expert. Since both events analyzed in this paper were open to the public, the expertise of participants varied from casuals to professional CTF players.

We can measure the position of the human team on the leaderboard as a measure of its expertise. In Figure 4 we show how the 50%-task-completion time horizon estimates change, depending on which percentage of top human teams we consider to be the experts.

TODO

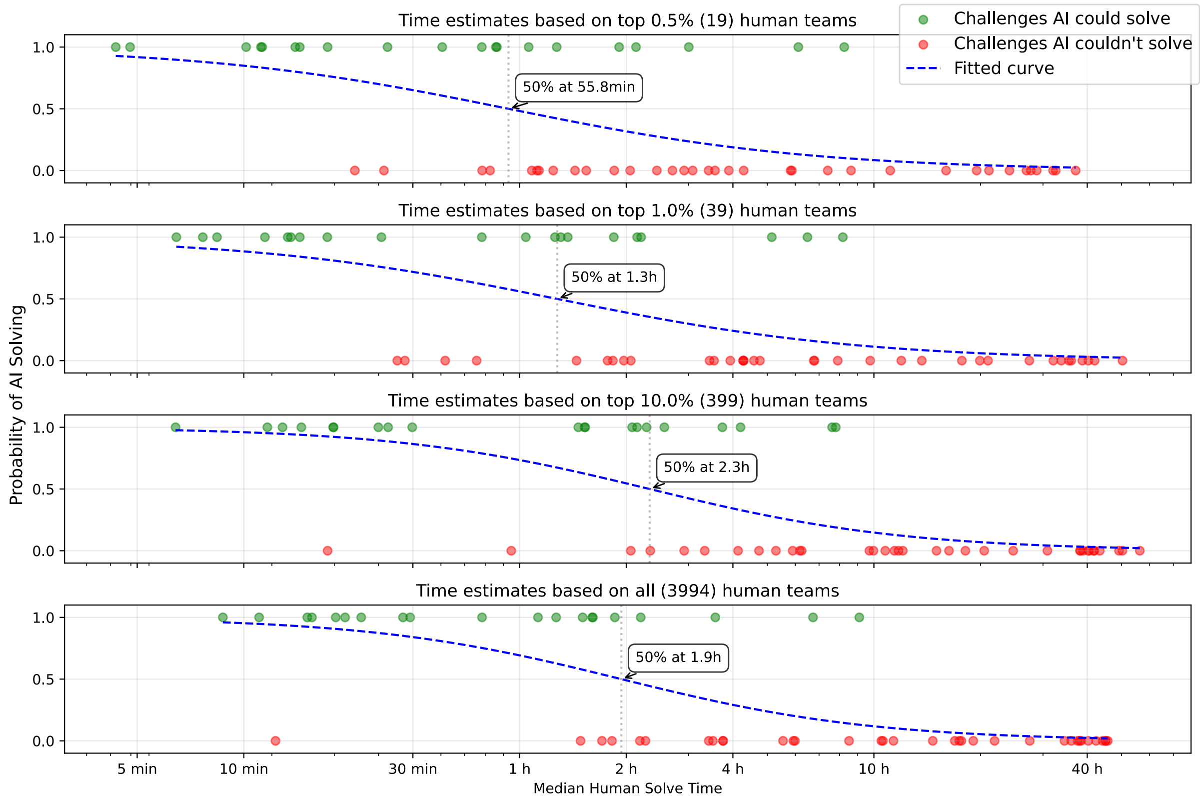

Figure 4: 50%-task-completion time horizon estimates we get, depending on which percentage of top human teams we consider experts for getting the human solve times. Data is from the _Cyber Apocalypse_ event.

C Detailed data

C.1 Challenges solved by AI per category and difficulty

TODO or drop

C.2 Top team scores in real time

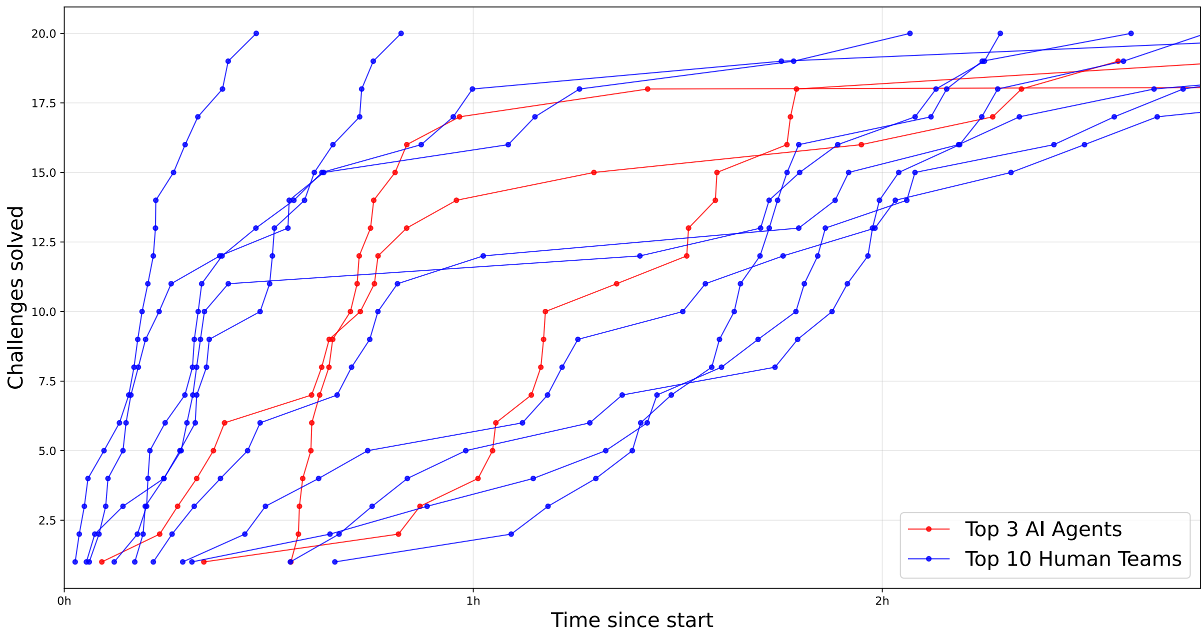

Figure 5: Number of challenges solved over time for *AI vs Humans* CTF.

C.3 All teams trajectories for *AI vs Humans* CTF

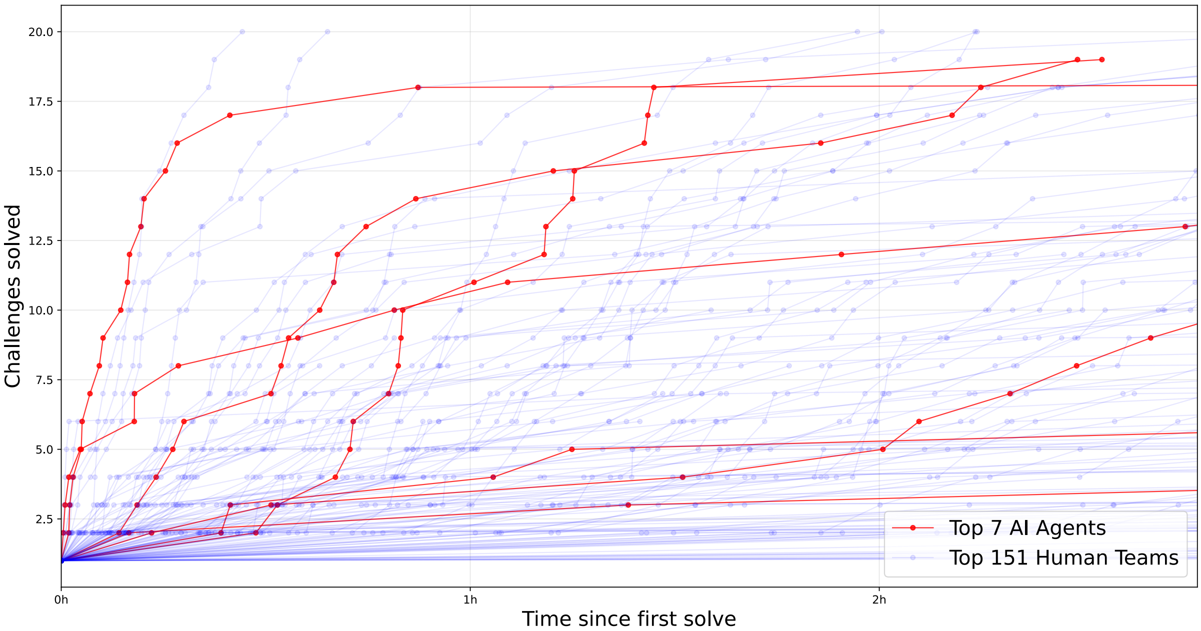

Figure 6: Number of challenges solved over time fror all teams for *AI vs Humans* CTF. Top AI Agents outperform almostall human teams.