

Evaluating AI cyber capabilities with crowdsourced elicitation

Artem Petrov,¹ Dmitrii Volkov

2025-04-25

Abstract

v1

AI capability evaluations often suffer from underelicitation. Evaluations of harmful capabilities, like cyber, exhibit this even more: frontier AI labs are not incentivized to put effort into proving their models dangerous, while safety organizations often lack the necessary resources to put much effort into agent design.

Additionally, classic benchmarks often lack the intuition of how its score translates to the equivalent human expertise.

To solve these issues we tried a crowdsourcing approach - we organized a Capture The Flag (CTF) event where AI developers and researchers competed against each other and human hackers. 6 AI teams joined in, one of which beat our best-effort approach.

AI teams ranked **top-13%** among teams that captured at least one flag, **while being as fast as top multi-person human teams**. AI teams solved Crypto and Reversing challenges of the “medium” level. Our setup allows us to report first ever 50%-task-completion time horizon for cyber. It is on the order of 1 hour, which is consistent with METR’s estimate (METR 2025) for general AI performance.

We also examine AI performance on another CTF event - CyberApocalypse, where it achieves top-21% score.

v2

We organized a first of a kind Capture The Flag (CTF) event where AI developers and researchers competed against each other and human hackers. 6 AI teams joined in, one of which beat our best-effort approach. This shows how crowdsourcing can help with underelicitation.

AI teams ranked **top-13%** among teams that captured at least one flag, **while being as fast as top multi-person human teams**. AI teams solved Crypto and Reversing challenges of the “medium” level. Our setup allows us to report first ever 50%-task-completion time horizon for cyber. It is on the order of 1 hour, which is consistent with METR’s estimate (METR 2025) for general AI performance.

We also examine AI performance on another CTF event - CyberApocalypse, where it achieves top-21% score.

1. Introduction

Eliciting AI capabilities is hard. Yet accurately estimating AI capabilities is important, especially when it comes to dangerous capabilities, as it determines whether a model is safe to release. Cyber is one of such capabilities.

Cyber evaluations may be underelicited, as we discuss in Section 5.

In this paper we explore crowdsourcing elicitation efforts as a strategy to boost cyber performance and close the evals gap. To do that, we host a CTF competition, inviting AI developers to compete against humans and each other for a 7500\$ prize pool.

We also report on the performance AI exhibited in this event and compare it to human performance. Finally, we analyze AI performance on CyberApocalypse - a large CTF competition with thousands of participants.

I think we should thank Reworr, Rustem, Alexander in Acknowledgements but not here

Ok. I think it would be cool to have some policy to manage expectations here

I feel confused reading this

cite apollo text

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

2. AI vs Humans CTF

In collaboration with Hack The Box we organized an “AI vs Humans CTF” event.

We focused on Crypto and Reversing challenge categories because these categories were the easiest to build a harness for - they only require running commands in the terminal, without the need for dynamic interactions with external machine.

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

You can see the final AI teams standings in Table 1.

Team	Challenges solved	Leaderboard rank ²
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

Table 1: AI teams standings³. Best AI team achieves top-13% performance among teams that solved at least 1 challenge.

2.2. Speed

One of the core advantages AIs hold over humans is speed.

If we look at Figure 1, we can see how quickly teams solved the challenges. AI teams seem to be on par with top human teams in speed.

²In case of teams solving an equal number of challenges, the faster teams rank higher on the leaderboard

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

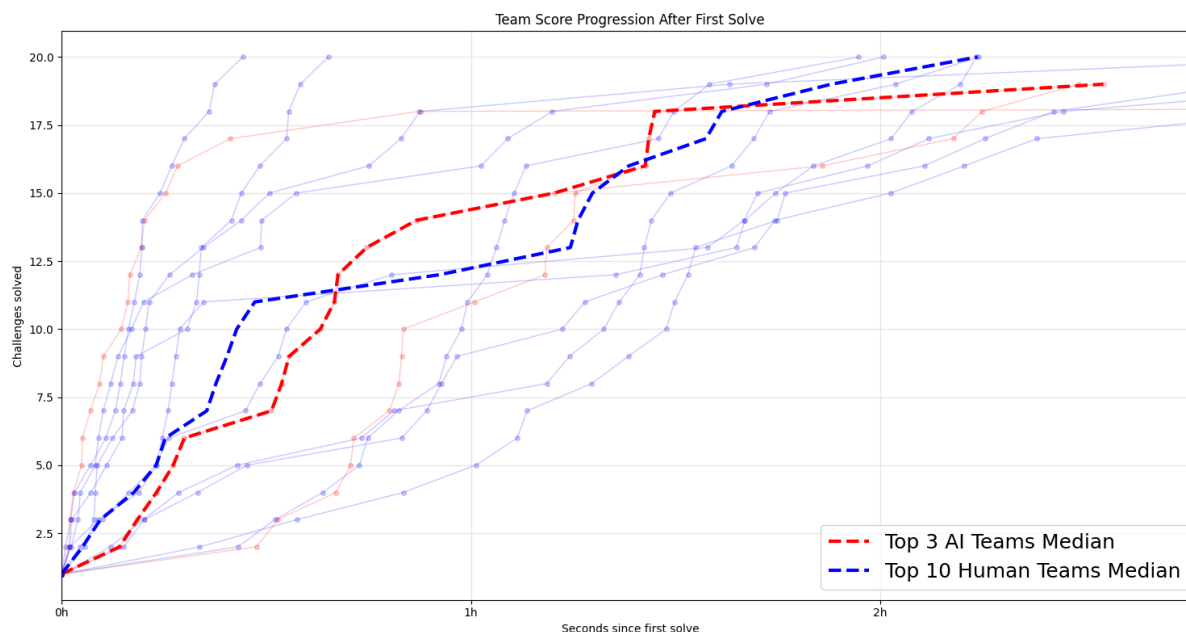


Figure 1: Number of challenges solved over time

The incredible speed of top human teams surprised us, so we reached out to the 5 top human teams for a comment. They say they are able to crack CTFs so fast, because they are professional CTF players and know all the common techniques to solve them. One participant told us that he was “playing on a couple internationally-ranked teams with years of experience”.

Do we
need the
median
line?

2.3. Agent designs

Here are some of the agent designs used by AI teams:

[AI] CAI:

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

Our claude code (+ Imperturbable enigma + claude code)

Our in-house

[AI] 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.

Here is the spreadsheet I used to track my progress

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

3. Cyber Apocalypse

Cyber Apocalypse is an annual CTF competition, organized by Hack The Box. Two AI teams participated in this event. Overall, 3994 human teams solved at least one challenge.

count
humans
rather
than
teams

You can see the final AI standings in table 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.

Team	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

Table 2: AI teams standings for Cyber Apocalypse⁴. AI achieves top-21% performance among teams that solved at least one challenge.

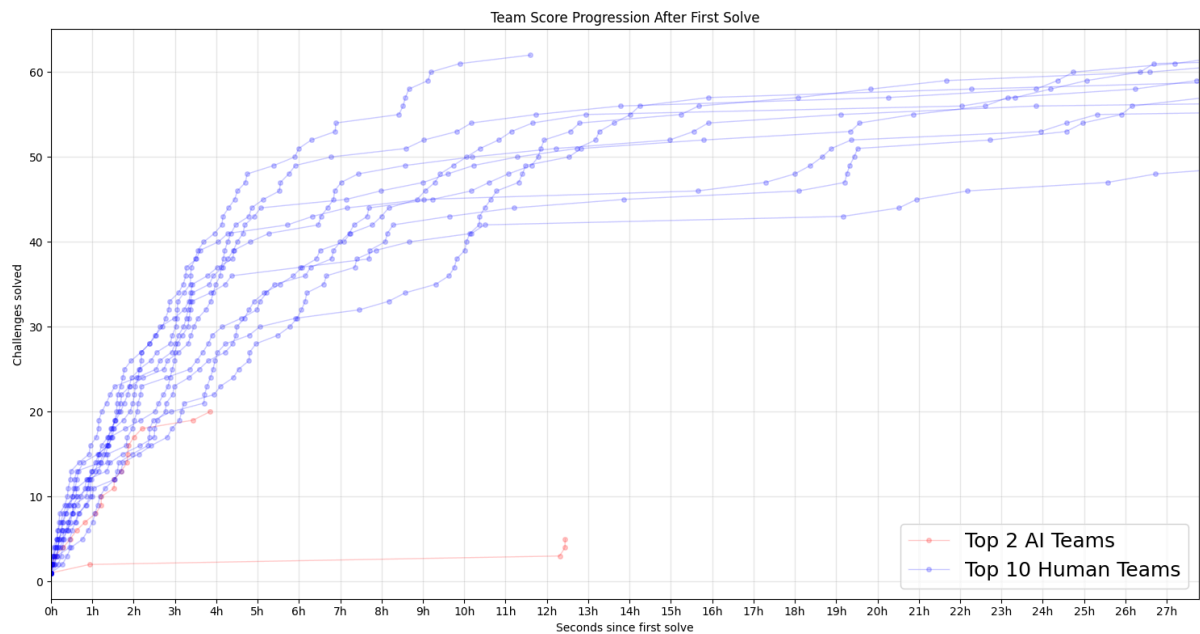


Figure 2: Number of challenges solved over time

4. Ability to complete long tasks

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

To estimate the human expert effort equivalent to current AI capabilities we follow (METR 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.

Analyzing the data from CyberApocalypse we reach a similar picture - AI can solve challenges requiring ~1 hour of effort from a median participant. See Appendix A for details.

⁴The full leaderboard, including human teams, is available at <https://ctftime.org/event/2674>

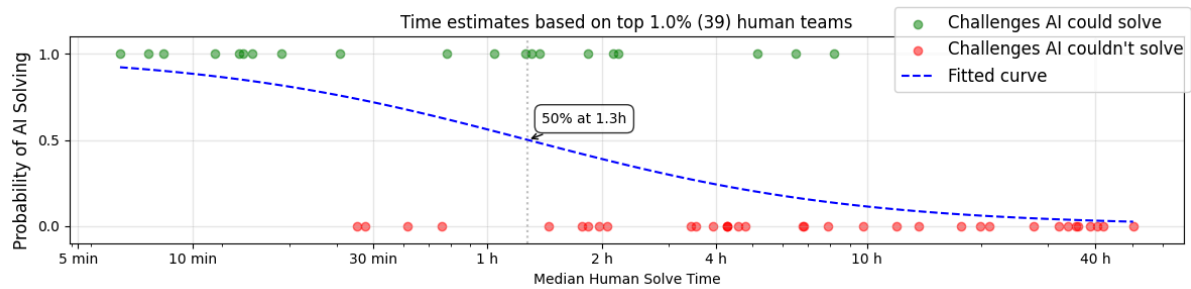


Figure 3:

5. Discussion

5.1. Cyber may be underelicited

Currently cyber evaluations of frontier models are usually done by a small internal evaluation team.

Cyber evaluations made by a single team sometimes may be underelicited.

Here are some anecdotal evidence for this:

- Deepmind reported (Phuong et al. 2024) 30% (24/81) score on the Intercode-CTF cyber benchmark (Yang et al. 2023), which is lower than its baseline of 40% (40/100). This may have happened because their models were weaker, or because of insufficient elicitation.
- Project Naptime (Project Zero 2024) pointed out the elicitation deficiencies in Meta's CyberSecEval 2 (Bhatt et al. 2024), bumping score from 0.05 to 1.0 on the "Buffer Overflow" tests and from 0.24 to 0.76 on the "Advanced Memory Corruption" tests. They achieve it by modifying the agent harness in simple ways (like, WAY 1 and WAY 2).
- Similarly, we managed to get SOTA on a Intercode-CTF by being careful with agent design (Turtayev et al. 2024), surpassing previously reported results.

We believe, crowdsourcing elicitation will increase evaluation robustness, because it reduces reliance on any specific team.

5.2. Does crowdsourcing work?

When inviting Alteams for this event, we focused on 3 types of AI teams- AI red teaming/pentester startups, researchers that published papers on cyber agents design, and Frontier AI labs.

Our hypothesis was that offensive AI startups would be interested to showcase their agent on the leaderboard, as it would be a good marketing to say "our agent beats x% of humans". This did not hold true.

AI team type	Participated/Invited
Offensive cyber AI startups	0 ⁵ /X
Researchers	1/3
Frontier Labs	1/2
Individuals	2/0

Table 3: AI teams standings for Cyber Apocalypse⁶. AI achieves top-21% performance among teams that solved at least one challenge.

Initially we reached out to startups building automated pentesters. However, none of them were interested. On the other hand, the CAI team, which looks like a startup, registered by themselves. We expected that showcasing their products on a leaderboard could be good for marketing, but it seems that is not the case.

Only researchers and enthusiasts expressed interest.

⁵TODO: check if CAI is here,

⁶The full leaderboard, including human teams, is available at <https://ctftime.org/event/2674>

improve
chart
design

Are there
external
evaluations
in any of
frontier
labs? How
much are
you
confident
about this
actually
being
true?

HTB invited Athropic's team.

CAI though did better than us.

- offensive vendors (who did not respond), (some may have a single customer and don't care about market opinion, others may have found a different kind of proof they заточились под)
- researchers,
- frontier labs

6. Conclusion

We want to have real time or better awareness of AI cyber capabilities

1. Policy-relevant grounding
2. Cost-efficiency + robustness

Bounties solve these 2.

Call to action for grantmakers: Moreover, CTFs are already happening - all we need is bounty money for AI track. If having an AI track that becomes popular default - no more grants needed.

Call to action for frontier labs: Way better to know in realtime than to not know. Faster decision making for better safety.

Call to action for CTFs: Consider AI track, for hype and attracting best experts (asks us for help - we have experience)

- AI achieved 1 hour time horizon on cyber (and having humans to measure this against is cool)
- Signs of crowdsourcing working, but results inconclusive
- AI outperformed our expectations based on initial evaluations. (saturated)

7. Results:

- Elicitation bounty suggested (seems like a good idea) (we tried, you should too) (we think it can be cost effective) (handwave it. 20k per month vs US salary to discussion + make a handwave table) (DARPA does similar AIXCC challenge, but with different purpose)
- time horizon (Metr suggested for R&D, we did for Cyber)
- AI top-13% scores (now this is **grounded** -> hence legible for policymakers (competition with humans is more legible than benchmarks in general. Hence doing it this way is better for policymaker awareness)(more comprehensive than cybench?))

8. Acknowledgements

AI teams, by team name and by people names for CAI and the team that filled the form.

HTB

Reworr, Rustem Turtayev, Alexander Bondarenko,

Bibliography

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>.

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>.

METR. 2025. "Measuring AI Ability to Complete Long Tasks". 2025. <https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/>.

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 Ctf's 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>.

Appendix

A Measuring 50%-task-completion time horizon

To estimate the human expert effort equivalent to current AI capabilities we follow (METR 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.

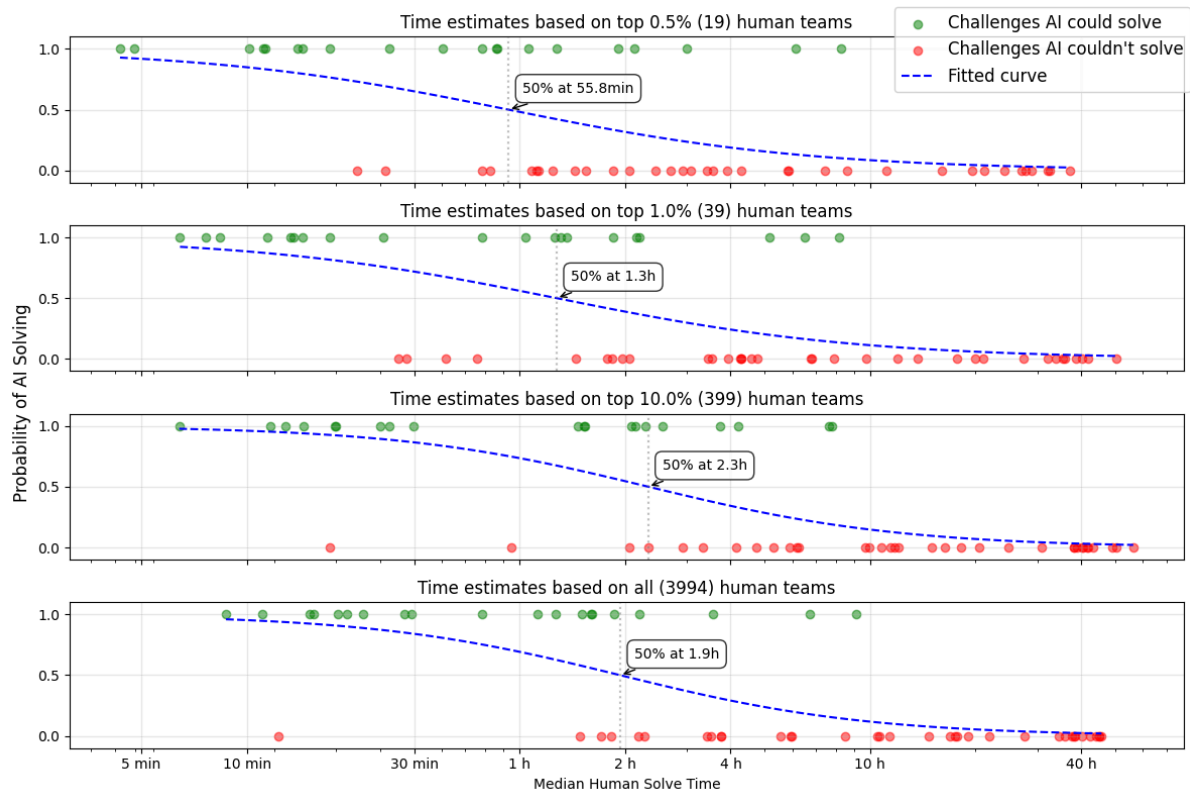


Figure 4: 50%-task-completion time horizon estimates we get, depending on which percentage of top human teams we consider experts for getting the human solve times. Data is from the Cyber Apocalypse event.

B Correcting for player number

If we naively normalize human team’s solve times by multiplying them by the number of players a team has, this effect is even more visible.

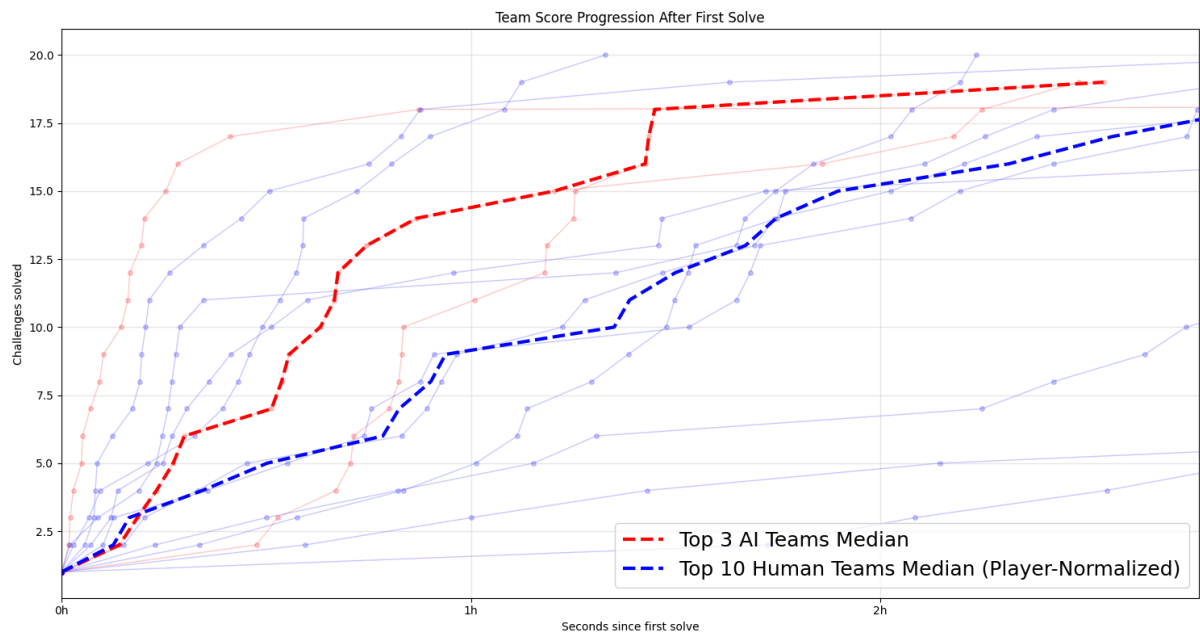


Figure 5: Number of challenges solved over time, with human teams' solve times normalized by the number of players

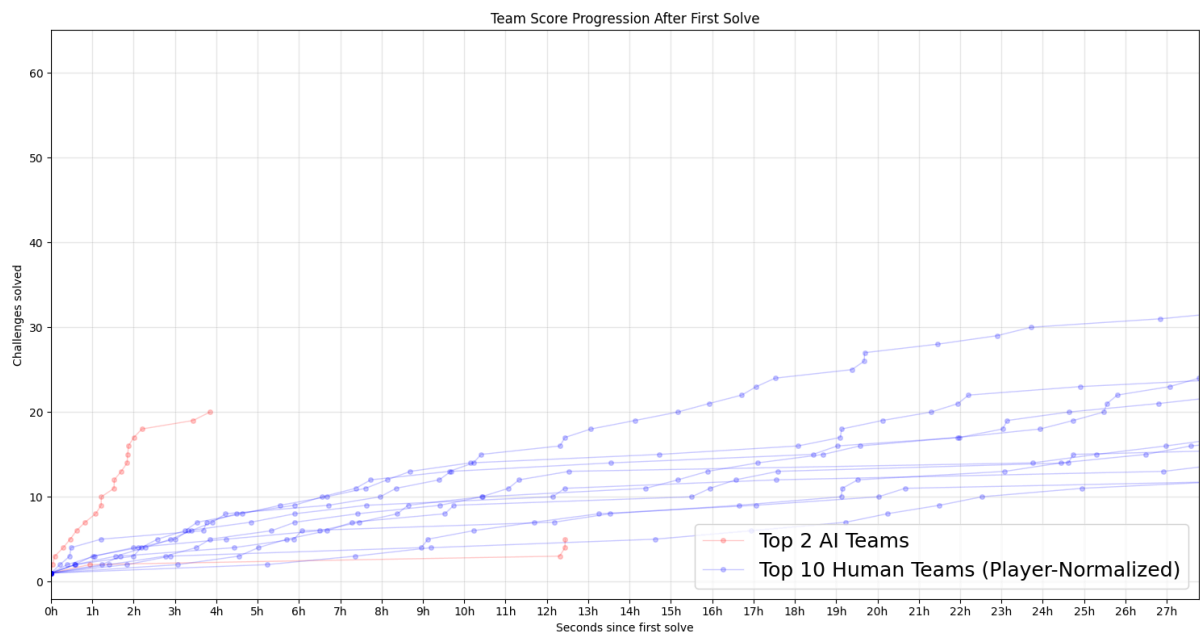


Figure 6: Number of challenges solved over time

C Links

<https://ctf.hackthebox.com/event/details/ai-vs-human-ctf-challenge-2000>

D TODO:

- [] We must comment on cyberapocalypse results
- [x] We must ask participants for their data on participants counts
- [] add acknowledgements by name where possible.

E Things to mention:

- [] Our event difficulty means human time spent, for CyberApocalypse it does not.