

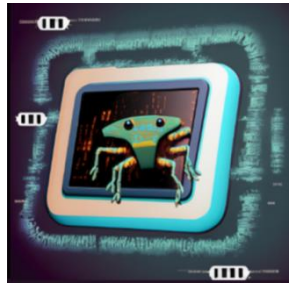
From Assistance to Autonomy: An Empirical Study of AI Use in a Live Capture-the-Flag (CTF) Competition

Tingxuan Tang, Nicolas Janis, Kalyn Asher Montague, Kevin Eykholt, Dhilung Kirat, Youngja Park, Jiyong Jang, Adwait Nadkarni, Yue Xiao

WILLIAM & MARY IBM

Existing Work

Benchmarks

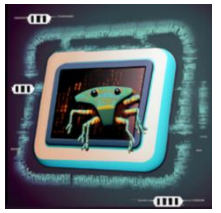


Cybench

CTFKnow

Automated Agents

InterCode





EnIGMA

AI-first CTF



What about **Human-AI Collaboration** in CTF context?

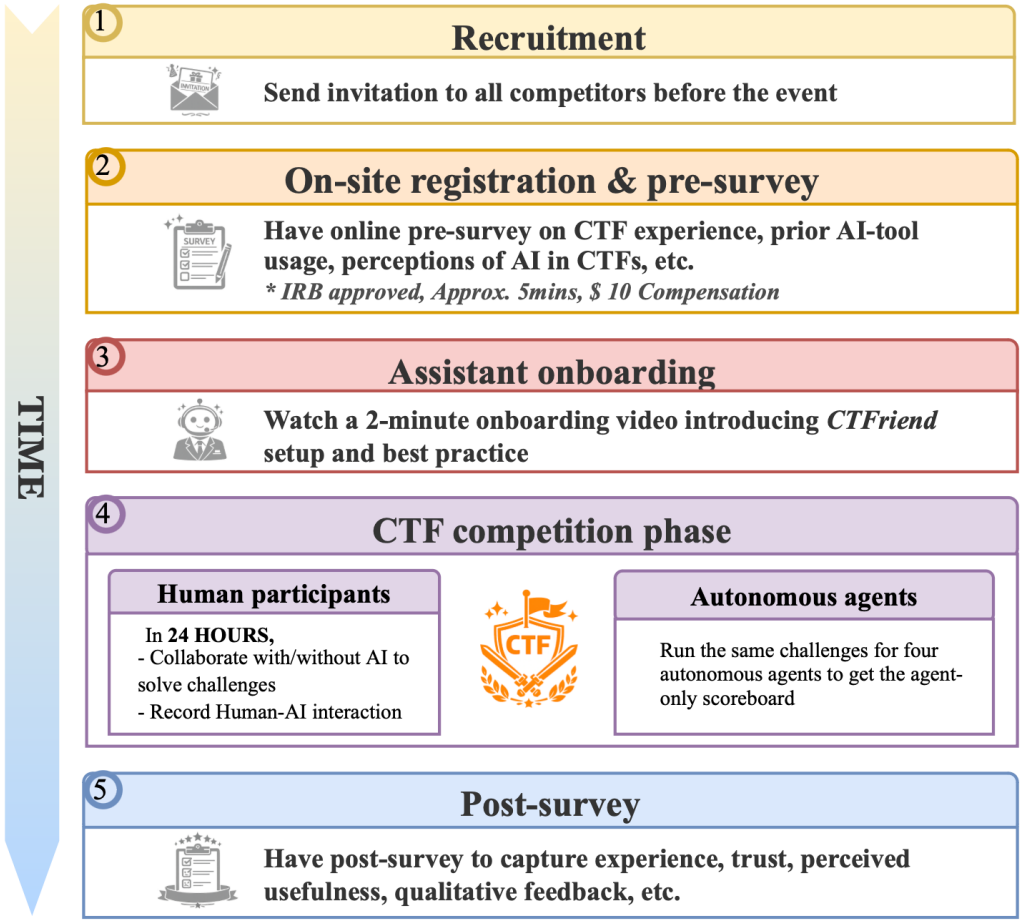
Our Work

RQ1: What ***expectations*** do participants have for AI in a CTF scenario?

RQ2: How do humans actually ***collaborate with AI*** during a live CTF?

RQ3: How do ***autonomous agents*** perform compared to humans?

A Live Competition: TribeCTF 2025



95 CTF participants

41 Study participants

24-Hour Competition

17 Challenges

5 Categories

3 Difficulty levels

https://tribectf.cs.wm.edu/2025-recap/

RQ1 — Participants' Reflections

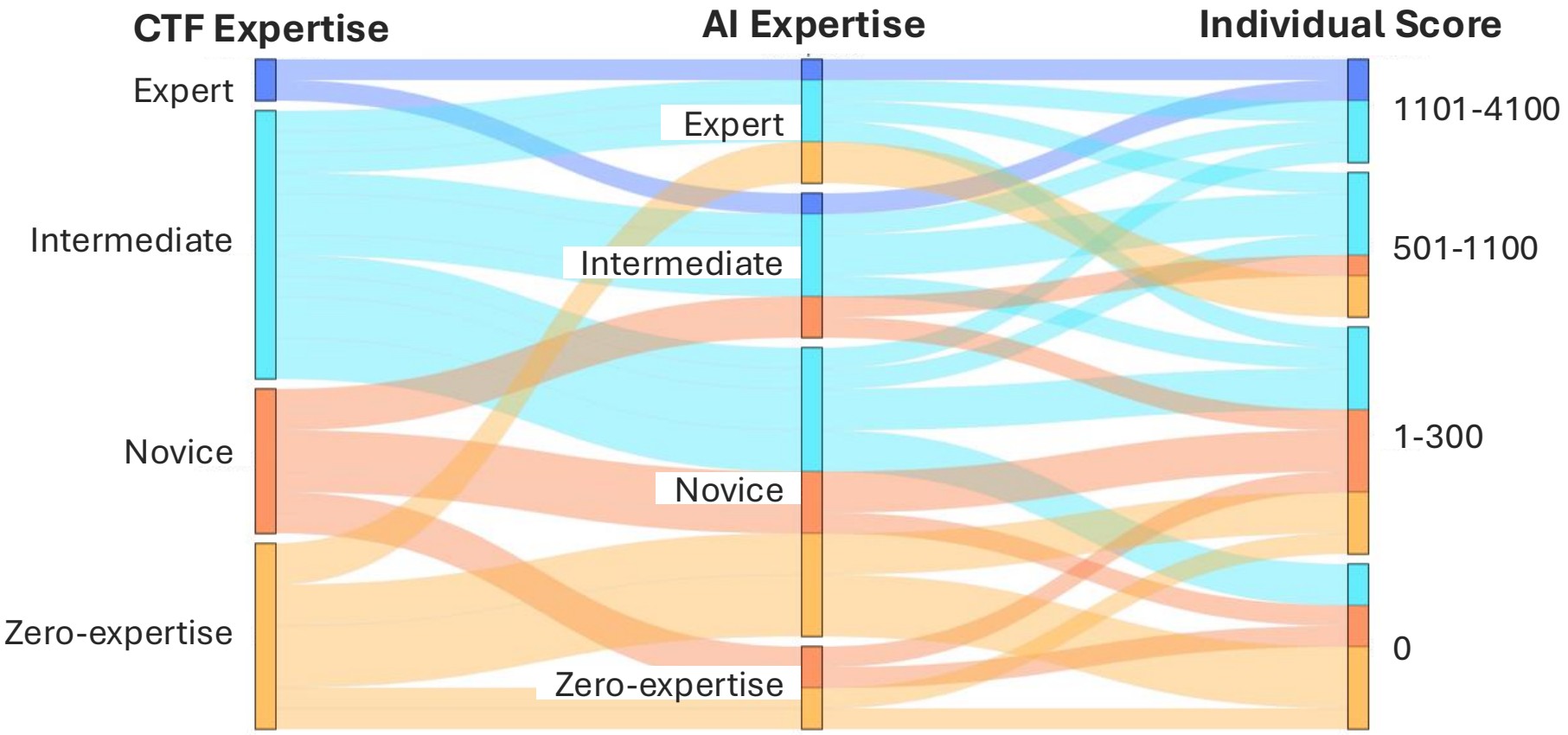
Pre-CTF Survey

VS.

Post-CTF Survey

Following the competition, participants' perceptions and trust in AI **decreased**.

Solid understanding of user input and **clarity** in output, **BUT**, **incomplete and unactionable** solutions.



High AI expertise + **High** CTF expertise = **Top** scores

High AI expertise + **Low** CTF expertise = **High** scores

AI can fill domain knowledge gaps!

RQ2 — Interaction Behaviors



Collaborate iteratively with AI

BUT

Delegate full tasks to AI

High AI Expertise

User

Can you help me solve this interesting CTF challenge? Challenge Overview: <...> Files Provided: `flash.img` Flag Format: ctf{...} Connection Behavior: <...> Server Output: <...> Decoding Attempts Tried: <...>

CTFriend

Let me think about this differently. The typo "WARING" missing 'N' might mean we need to ADD 'N' (0x4E) not XOR with it. Let me manually calculate...

Low AI Expertise

User

what is the password now?

CTFriend

Based on the decoded keystrokes, Jeff typed...

User

none of them are working

CTFriend

Let me reconsider the decoded keystrokes...

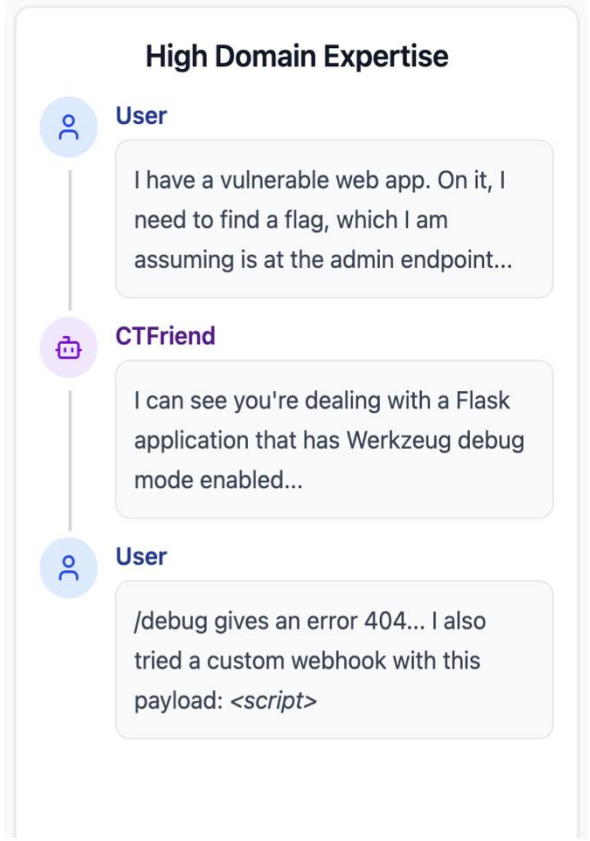
User

none of them are working

High AI expertise → **Effective** Prompting → **Higher** Success rate

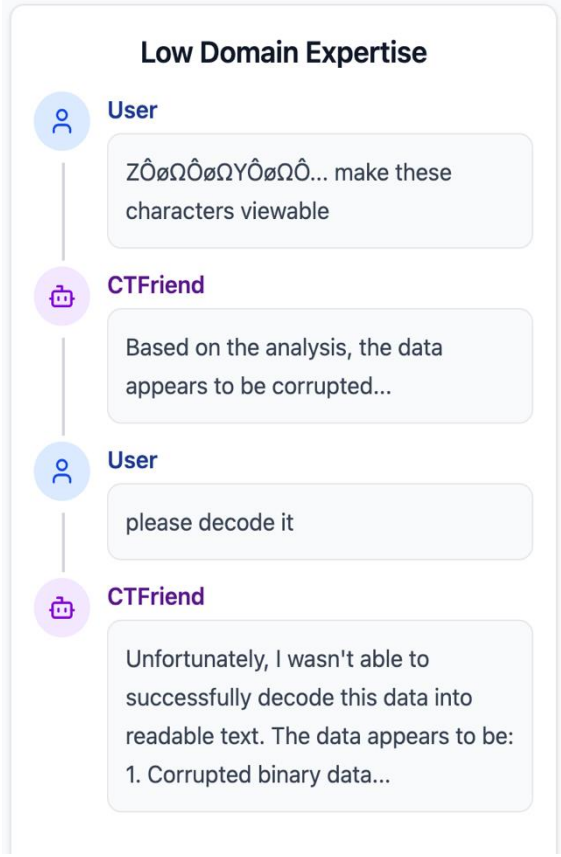
Low AI expertise → **Low-quality** Prompting → More **errors** and **failures**

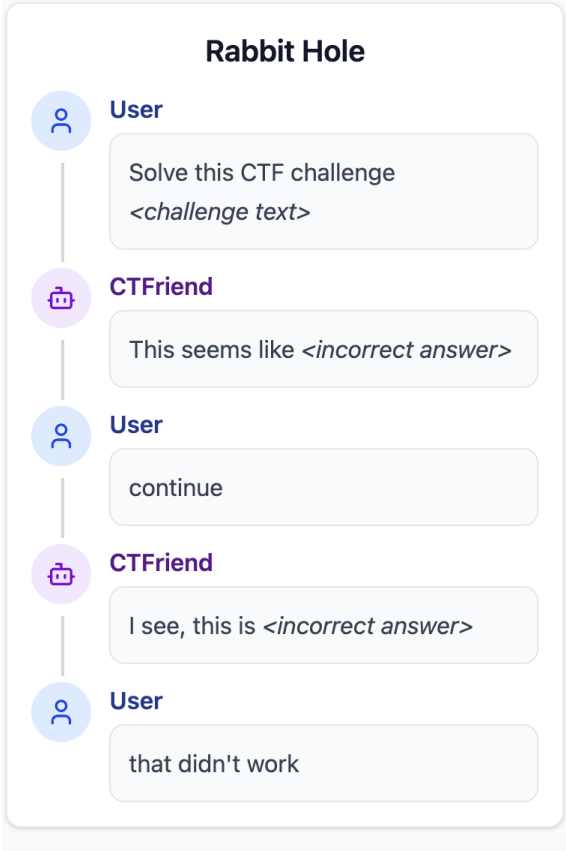
High CTF Expertise:



Inject domain knowledge

Low CTF Expertise:



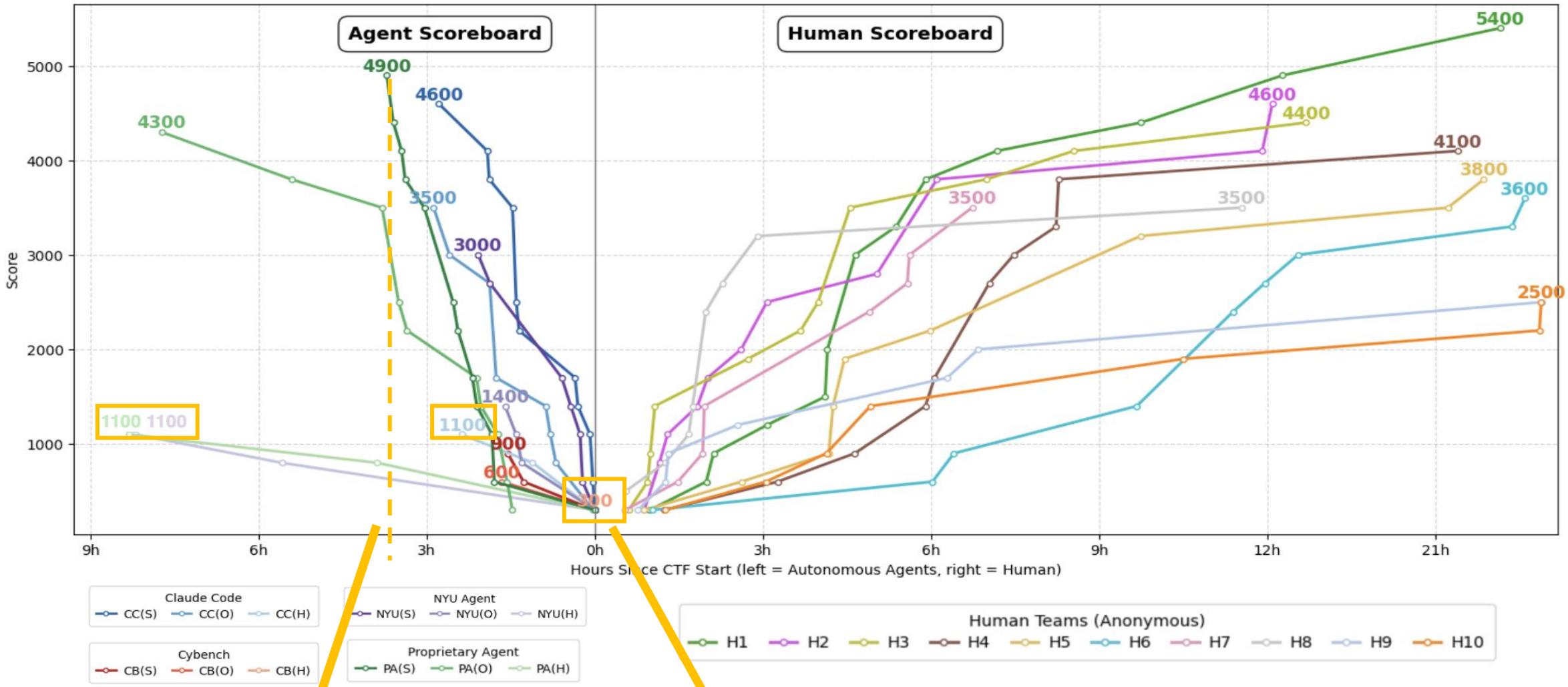




"Rabbit Holes"

"Answer Shopping"

RQ3 — Agents vs. Human Teams



About $\frac{1}{5}$ human solving-time

Model capability ultimately sets the ceiling.

The Power of Pair Hacking

RANK	AGENT/TEAM	CHALLENGES SOLVED	SCORE	LLM
1	**Proprietary Agent (with Human-in-the-Loop)**	14	5700	Sonnet-4.5
2	**Human Team 1**	13	5400	N/A
3	**Proprietary Agent**	12	4900	Sonnet-4.5

Pair hacking highlights the advantage of **human-in-the-loop** workflows:

> The agent can run autonomously while humans provide **sparse steering and verification**.

Takeaways for Stakeholders

CTF Organizers:
Design interactive, stateful challenges that are operationally hard for automated agents

Security Practitioners:
AI-assisted system is indeed effective, but humans are still necessary for risk and correctness checks

AI Agent Developers:
Allow human intervention at critical decision points during agent execution



Human-in-the-loop is the winning paradigm:

AI accelerates exploration, while humans steer, verify, and recover.

Thank you!

WILLIAM & MARY

IBM

If you have any questions, please feel free to contract **Tingxuan Tang** at ttang01@wm.edu.