Building Trust Infrastructure for Human-in-the-Loop Adversarial ML Research: Lessons from a Multi-Channel Recruitment Pilot for AI-Phishing Benchmarks

Camilla Balbis, Fred Heiding

camilla.balbis@gmail.com, fred@taici.org

Executive Summary

Adversarial machine learning benchmarks that evaluate large language model (LLM) capabilities, such as generating persuasive phishing emails, require ongoing human-in-the-loop evaluation to generate valid data. Yet, no published methodology exists for recruiting voluntary human subjects for deception-based adversarial research. This report documents a three month no-budget pilot across sixteen Canadian universities that tested ted messaging strategies, four outreach channels, and peer-network incentive model to recruit participants for ScamBench, and adversarial phishing benchmark developed by The AI and Cybersecurity Institute (TAICI).

Key findings challenge three common assumptions in the current phishing literature: 1) the most technically relevant audiences are not the most responsive, namely cybersecurity and computer science clubs converted at 0%, while general STEM and women-in-STEM groups converted at a 8-11% rate; 2) institutional newsletters were more willing to support recruitment efforts than student clubs (9.1% vs 2.7%), showing that formal broadcast channels can beat targeted technical communities for visibility potential; 3) a Student Ambassador model which delegates trust-building to existing and in-person peer networks, projects 500-1000 signups for a single partnership, offering a scalable alternative to labor-intensive cold outreach. The pilot secured a total of twelve student clubs partnerships, fifteen newsletter mentions, and one high-impact Student Ambassador agreement, with the potential for another.

This document provides nationalable recommendations for researchers, funders and institutions, first and foremost: deception-based research requires institutional trust-building that exceeds short-term timelines.

Introduction & Problem Statement

Phishing remains one of the most consequential cyber threats, contributing to 90% of cybersecurity attacks and costing organizations billions annually (Cybersecurity and Infrastructure Security Agency, n.d). The emergence of large language models (LLMs) has dramatically lowered the barrier for generating persuasive, context-aware phishing content, enabling adversaries to craft attacks that evade both technical filters and human detection (Heiding et al., 2024). Understanding how these AI-generated phishing emails perform against human judgment is therefore critical for developing effective defenses.

ScamBench, developed by TAICI, an independent institution born from a Harvard/MIT-affiliated research collaboration, addresses this need by creating a human-validated benchmark for LLM-generated phishing detection (ScamBench, n.d.). ScamBench requires ongoing participation from human volunteers who receive simulated phishing emails in controlled, safe environments and provide ground-truth labels on their detectability. This human-in-the-loop design is methodologically essential but creates a recruitment challenge: how do you persuade people to voluntarily receive fake phishing emails?

This problem sits at the intersection of three governance gaps. First, the adversarial ML literature has focused overwhelmingly on technical benchmark development while treating human-subject recruitment as a consequence (Das et al., 2019). Second, the phishing user-study literature reveals that recruitment is frequently the weakest methodological link, with studies relying on convenience samples, captive populations, or underreported participant pools (Parsons et al., 2015). Third, as mentioned above, no framework exists for the specific case of voluntary, ongoing participation in deception-based research- where the research topic itself may trigger defensive reactions that undermine recruitment.

This report documents a pilot designed to address this gap. Over three months (April–May 2026), I tested multiple outreach channels, messaging strategies, and incentive structures across sixteen Canadian universities to build a sustainable recruitment pipeline for TAICI's ScamBench. The pilot was not designed to maximize immediate signups but to understand the

dynamics of recruitment for sensitive adversarial research, and to build infrastructure that would continue well beyond the pilot period.

Background & Context

The Recruitment Problem in Phishing User Studies

The literature on phishing user studies reveals a persistent recruitment challenge. Das et al. (2019) conducted a systematic literature review of 367 phishing papers in the ACM Digital Library and found that only 13.9% (n = 51) included any user-focused study methodology. The authors conclude that this underreporting likely reflects "*a lack of intentional, strategic recruitment strategy*" and call for future researchers to "make every effort to recruit more diverse and representative samples of participants" (Das et al., 2019).

This recruitment opacity has methodological consequences. Parsons et al. (2015) examined the design of phishing studies and found that participant priming (informing subjects they are in a phishing study) significantly alters behavior, with "alerted" participants showing better discrimination between phishing and genuine emails than uninformed controls. The authors caution that "studies where participants are directly asked to identify phishing emails may not represent the performance of real world users, because people are rarely reminded about the risks of phishing emails in real life" (Parsons et al., 2015).

Existing Recruitment Approaches and Their Limitations

- Captive populations, corporate employees, students in required courses: This approach leverages the power of social context, especially in frames where participants cannot opt out, but also raises ethical concerns and produces data that may not generalize to other voluntary settings.
- Laboratory studies with paid participants: These approaches provide controlled conditions and informed consent but are limited by: (a) one-time participation (no learning or adaptation over time), (b) artificial task framing that primes participants, and (c) demographic homogeneity.
- Synthetic data and automated metrics: Many adversarial ML benchmarks bypass human subjects entirely, using automated detection systems or synthetic user models. While scalable, these approaches lack the ecological validity of human judgment.
- Peer Networks and Trust-Based Participation: peer-mediated recruitment. The mechanism is twofold: trust transfer (the intermediary's credibility reduces perceived risk) and social proof (visible peer participation normalizes engagement).

The Student Ambassador model tested in this pilot, rather than recruiting directly, recruits peer recruiters who leverage existing social networks within student clubs and academic programs, shifting the trust-building burden and potentially solving the scalability limits of manual outreach.

Analysis

3.1 Dataset and Methodology

The pilot generated 449 unique contacts across sixteen Canadian institutions, segmented by outreach channel:

Channel	Contacts	Secured Partnerships	Conversion Rate
Student clubs	372	10	2.7%
Newsletters / general media	22	2	9.1%
Faculty / professors	53	0	0.0%
Programs / classes	2	0	0.0%
Total	**449**	**12**	**2.7%**

Contacts were categorized by institution, channel type, thematic focus (for clubs), and whether a named individual or generic institutional email was used. Ten distinct email variants were sent over the three- month period, with content and framing systematically varied to test messaging effects. Two custom Python tools were developed: a contact scraper for automated extraction from university directory pages, and a mass outreach pipeline for personalized batch email generation and delivery.

3.2 Channel Efficacy: Newsletters and Named Contacts Outperform Direct Outreach

The data revealed a clear hierarchy of channel effectiveness that challenges conventional assumptions about targeted outreach.

- Newsletters convert at 3.4× the rate of direct club outreach. Despite representing only 4.9% of total contacts (22/449), newsletters accounted for 17% of secured partnerships (2/12) and showed the highest conversion rate (9.1%). This finding is consistent with the broader literature on institutional credibility: newsletters provide editorial vetting, passive reach (readers encounter content without active solicitation), and lower perceived pressure compared to direct email asks.

- Named individual contacts convert at 5.5× the rate of generic institutional emails. Contacts with a specific named person (38 total) secured 4 partnerships (10.5%), while generic/institutional contacts (411 total) secured 8 (1.9%). This validates the labor-intensive manual research approach used, however, at fifteen hours per week, this approach is not scalable beyond approximately 150 contacts per quarter.
- Faculty contacts are a resource trap. Fifty-three faculty contacts across three UofT campuses and TMU yielded zero conversions. This finding challenges the common assumption that authority figures can serve as effective recruitment gateways. Possible explanations include: saturated inboxes, timing (pilot occurred at the cusp of spring break), lack of direct student distribution channels, absence of incentive structures for faculty, and delegation to teaching assistants or club leaders.

3.3 The Messaging Experiment: Ten Variants and Their Performance

Ten email variants were tested, with variation in framing, tone, length, and persuasive features.

Key messaging findings:

- The credibility saturation effect: Variants with excessive credibility markers (Harvard/MIT, Reuters, arXiv, etc.) performed worse than those with moderate credibility. Variant E, with 17 credibility references and 347 words, converted at 0% while Variant B, with 6 references and 105 words, converted at 13.1%. Beyond a threshold, additional credibility signals may trigger skepticism or imposter syndrome in student recipients.

- The fear-benefit asymmetry: Variants that aroused fear without offering clear benefits created paralysis rather than action. Variant E (7 fear words, 3 benefit words, 347 words) failed completely. The optimal formula combined moderate fear (7 words) with career benefits (2 words), urgency (3 words), and community framing (12 words), as in Variant I, which achieved 14.7% conversion.

- The follow-up multiplier: Emails with value-adding follow-ups (not just reminders) converted at 5.9% average; those without follow-ups converted at 0.2%, a full 29.5× difference. Specifically, variant H's follow-up added new information (career benefits, specificity about attack types, urgency about closing cohorts), transforming a 0% initial email into 5.6% conversion.

- The generational responsibility frame. Variant I's most distinctive element was its generational framing: "*You're the first generation that needs to think like this. Your parents and grandparents are the primary targets of AI scams right now. You're the ones they'll call when*

something looks off." This shifted the beneficiary from the researcher to the participant's social network, achieving 10 positive replies with zero rejections, making it the only variant with 100% positive reply rate.

3.4 The Relevance Paradox: Technical Audiences Are the Least Receptive

The most counterintuitive finding concerns audience selection. All 372 student clubs were categorized by thematic focus, revealing an inverse relationship between technical relevance and conversion:

Thematic Focus	Clubs	Secured	Conversion Rate
General STEM/ Science	18	2	11.1%
Women/Gender/ Diversity in STEM	23	2	8.7%
AI/ML/Data Science	31	2	6.5%
Arts/Creative/Media	20	1	5.0%
Cybersecurity/InfoSec	4	0	0.0%
CS/Software Engineering	7	0	0.0%
Business/Finance/ Fintech	9	0	0.0%
Healthcare/Medical/ Bio	13	0	0.0%

Cybersecurity and computer science clubs, the ostensibly most relevant audience, converted at 0%. This challenges the common assumption that domain expertise increases willingness to participate in related research.

Possible explanations include: (1) threat modeling literacy makes security-aware populations more risk-averse to "scam" research; (2) the name "ScamBench" triggers defensive skepticism in those who understand phishing mechanics; (3) these clubs are saturated with corporate and research solicitations; and (4) competitive event-focused cultures (CTFs, hackathons) prioritize skill demonstration over research participation.

Conversely, general STEM and especially women-in-STEM groups converted at 8- 11%. These groups have broader, less specialized memberships (lower gatekeeping), value research participation for resume and portfolio building, are actively seeking partnerships to increase engagement, and are less likely to be saturated with similar solicitations.

This finding has direct methodological implications for the phishing literature. Das et al. (2019) found that most user-focused phishing studies rely on convenience samples of university students but fail to report demographic diversity. This pilot suggests that which students are recruited matters as much as how many and that the most "technically" relevant populations may be systematically underrepresented in voluntary research due to self-selection bias.

3.5 The Trust Gradient and Temporal Mismatch

The pilot data reveal a trust-building progression that cannot be shortcut:

Mass Email ("Scary" Name) → 2.7% conversion
↓
Institutional Endorsement (Newsletter) → 9.1% conversion
↓
Named Personal Contact + Follow-up → 10.5% conversion
↓
Peer Network (Student Ambassador) → Projected 500–1,000 signups

Each step requires time and relationship investment. The largest partnership indeed during the pilot, the one with a Queen's University AI student club, was secured in April but projects impact for Autumn, a five-month lag between contact and signup. Similarly, the 15 newsletter mentions with an estimated 3,000 combined readership function as persistent recruitment assets, though one that given the lack of referral links is impossible to exactly estimate. This is to say, a three-month pilot cannot capture the true impact of institutional trust-building.

3.6 The "ScamBench Naming Effect"

Qualitative feedback confirmed a significant naming problem:

- Four personal contacts were "stressed/convinced I was hacked" when seeing "ScamBench"
- When approached in person, anecdotally, people "stopped listening/got turned off when the name was mentioned"
- A "good chunk" of initial non-respondents, after a successful follow-up admitted thinking the email itself was a scam
- Two contacts reached out to me on LinkedIn to verify the sender's identity, fearing a scam

The word "Scam" appears in every email variant, since it's the name of the benchmark being tested, triggering semantic priming of threat schemas and source confusion (conflating research topic with email intent). This naming effect is consistent with the phishing literature's findings on defensive processing: when threat cues are activated, recipients scrutinize rather than engage (Vishwanath et al., 2016).

The most successful variants (B and I) buried or avoided “ScamBench” in subject lines, leading with “AI Safety” or institutional credibility instead. The Ambassador variant (L) barely mentioned the project name, focusing entirely on career benefits. This suggests that project naming is a governance variable that should be systematically tested against alternatives.

3.7 Tooling and Scalability Tradeoffs

Two Python-based tools were developed to address resource constraints:

1. Contact scraper: Automated extraction of club/faculty emails from university directory pages, with deduplication and structured output to Google Sheets.
2. Mass outreach pipeline: Personalized batch email generation with Gmail SMTP integration, delay management, and sheet-based recipient lists.

These tools reduced per-contact research time from approximately eight minutes (manual) to approximately thirty seconds (automated). However, conversion dropped from 4.7% (manual) to 1.7% (automate) a 2.8× decrease. This documents a scalability-quality tradeoff that has direct implications for resource allocation in research recruitment: automated tooling enables reach but sacrifices the personalization that drives conversion.

The Student Ambassador model offers a potential resolution. By delegating trust-building to peer, often in-person, networks, the model maintains the personal relationship quality of Phase 1 outreach while achieving the scale of Phase 3 automation. The 90% conversion rate with warm audiences (those who already expressed interest) validates this two-step funnel: initial low-commitment partnership → Ambassador conversion → peer-network recruitment.

Recommendations

For Researchers Conducting Adversarial ML Studies Requiring Human Subjects

1. Lead with “AI Safety” rather than project-specific names in cold outreach. Certain project names can trigger defensive processing that undermines otherwise effective messaging. Subject lines and opening sentences should frame the research in terms of protective or altruistic outcomes (“AI safety research,” “phishing detection study”) rather than adversarial mechanics (“ScamBench,”or “phishing benchmark”).

2. Deprioritize faculty channels, invest instead in student-facing infrastructure. The 0/53 faculty conversion rate, despite admittedly limited, is for the purpose of this study unambiguous. Redirect effort to student clubs, course unions, program newsletters, and peer ambassadors.

3. Budget 6-12 months for trust-building, playing a longer game. The Queen's University partnership case (5-month lag) and newsletter persistence (ongoing passive reach) demonstrate that institutional trust-building requires time.

4. Use value-adding follow-ups, not empty reminder bumps. The 29.5× follow-up multiplier applies only when follow-ups contain new information (career benefits, specificity, urgency). Simple reminders perform marginally better than no follow-up but dramatically worse than value-adding follow-ups.

For Funders and Institutions Supporting Adversarial ML Research

5. Fund named-contact outreach, not just automated scraping. The 5.5× conversion advantage of named contacts justifies the labor investment.

6. Support newsletter partnerships as force multipliers. Newsletters' 9.1% conversion rate and persistent reach make them high-ROI channels. Institutions should develop relationships with university communications offices and student media as standard practice for research dissemination. Also because, despite no direct signup connection proof, a successful newsletter appearance establishes a precedent of trust which can be re-leveraged in the future for another similar ask or warm intro.

7. Require longitudinal tracking (6–12 month follow-up) in recruitment evaluations. Current evaluation practices that measure success at project conclusion systematically undercount the value of institutional partnerships. It was not available for this pilot, but the use of referral links is one way to facilitate evaluation, and based on budget constraints, other strategies can be deployed.

For the AI Safety Field

8. Test alternative project names with A/B experiments. The ScamBench naming effect should be systematically validated. A controlled experiment comparing "ScamBench" vs. neutral alternatives on identical audiences would provide generalizable guidance for research branding.

9. Build incentive structures that offer tangible benefits for the targeted demographic. The Ambassador model's 90% warm conversion rate demonstrates that for the 18-30 student

demographic, career-oriented incentives (resume lines, references, early access, hiring pipeline) outperform purely altruistic framing. The field should incorporate standardized recognition frameworks for student research participants.

Conclusion

This pilot contributes to a gap in the adversarial ML and phishing literatures: the absence of methodology for recruiting voluntary human subjects for ongoing deception-based research. The findings challenge three background expectations from the literature: that technical audiences are the most responsive recruitment targets, that direct outreach outperforms institutional channels, and that recruitment scales through domain-fit rather than social-network effects.

The evidence suggests the opposite. The best-performing channels were broad institutional newsletters and peer-mediated Ambassador pathways, not the most technically aligned communities. Cybersecurity and computer science clubs converted at 0%, while general STEM and women-in-STEM groups achieved 8-11% conversion. Faculty, despite their subject matter authority, yielded zero partnerships across 53 contacts.

These findings align with the broader phishing literature's emphasis on recruitment as a methodological variable offering a concrete example of how ambassador-led recruitment can outperform cold outreach in hard-to-reach populations.

Future work should include: longitudinal tracking to measure actual signup attribution from each partnership, A/B testing of project names, cross-institutional replication of the manual reachout approach, and incentive optimization comparing certificate-based, credit-based, and monetary structures for Ambassador recruitment.

References

Cybersecurity and Infrastructure Security Agency. (n.d.). *Shields Up: Guidance for families.* https://www.cisa.gov/shields-guidance-families

Das, S., Kim, A., Tingle, Z., & Nippert-Eng, C. (2019). *All about phishing: Exploring user research*

through a systematic literature review. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13. https://arxiv.org/pdf/1908.05897

Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J., & Park, P. S. (2024). *Devising and detecting phishing emails using large language models*. *IEEE Access, 12*, 42131–42146. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10466545

Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., & Jerram, C. (2015). *The design of phishing studies: Challenges for researchers*. Computers & Security, 52, 194–206. https://www.sciencedirect.com/science/article/abs/pii/S0167404815000231

ScamBench. (n.d.). *ScamBench: Evaluating AI-powered phishing*. https://scambench.com/research

Vishwanath, A., Harrison, B., Ng, Y. J., & Rao, H. R. (2016). Suspicion, cognition, and automaticity model of phishing susceptibility. Communication Research, 43(8), 1073–1095. https://www.researchgate.net/publication/278676335_Suspicion_Cognition_and_Automaticity_Model_of_Phishing_Susceptibility