# Trustworthy AI

## Spring 2024

Yuan Tian

#14:Availability & Usable Security for LLM

# Reminders

- Final course project presentation will be on June 5
  - 5 minutes presentation + 2 minutes Q&A
  - 50% peer eval, 50% Yuan's eval
  - Group assignment on Gradescope
- Course evaluation (1% bonus) due on June 7
- Final report due on June 10
  - Group assignment on Gradescope
- Final project presentation evaluation and teammate evaluation (course project) due on June 11

# Final presentation slides template

- https://drive.google.com/file/d/12Z6MAEJmuE-z8r7J2mJ3gBKDPSgfdJEO/view?usp=drive_link

# Literature Review Presentations

- Availability
  - D1 Bit-Flip Attack: Crushing Neural Network with Progressive Bit Search
  - Hossein Khalili, Yibin Wang
  - D2 Sponge Examples: Energy-Latency Attacks on Neural Networks
  - Nhat Nguyen, Zaya Lazar, Ethan Peng, Yao Ting Hsu
- Usable Security for LLM
  - Do Users Write More Insecure Code with AI Assistants?
  - Jason Vargas, Tirumalasri Vedam, Christina Lee
- Link for screen sharing:
  - https://ucla.zoom.us/j/93279473227?pwd=Z3lXQzJoVTJrVGFuZ0tjcUlmTlRBUT09
- Please submit your peer reviews for this presentation by May 30 midnight
  - https://forms.gle/aNvt37LXHGSt9pBNA

# Devising and Detecting Phishing: large language models vs. Smaller Human Models

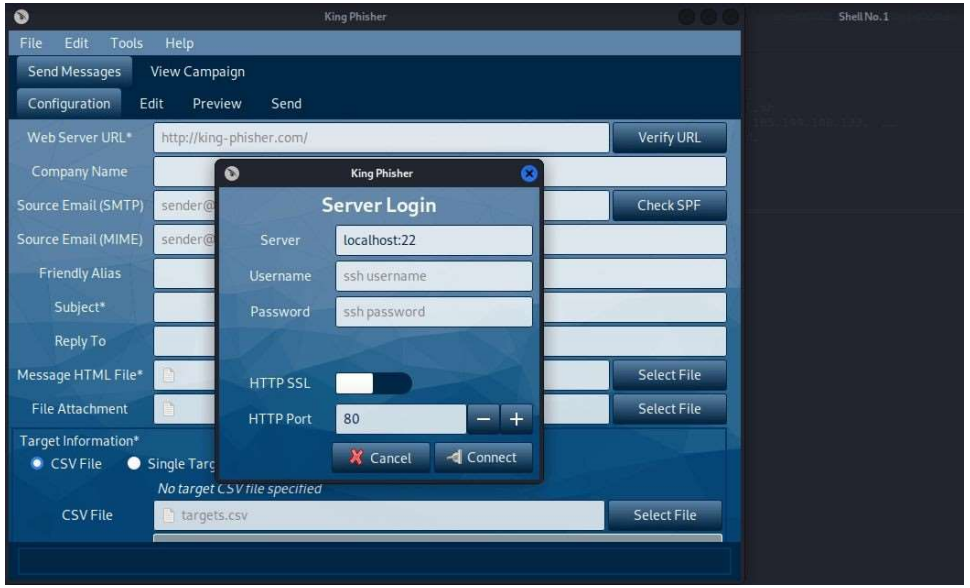

- Fredrik Heiding, Bruce Schneier, Arun Vishwanath, Jeremy Bernstein, Peter S. Park

Can LLMs be used to automate phishing email generation?

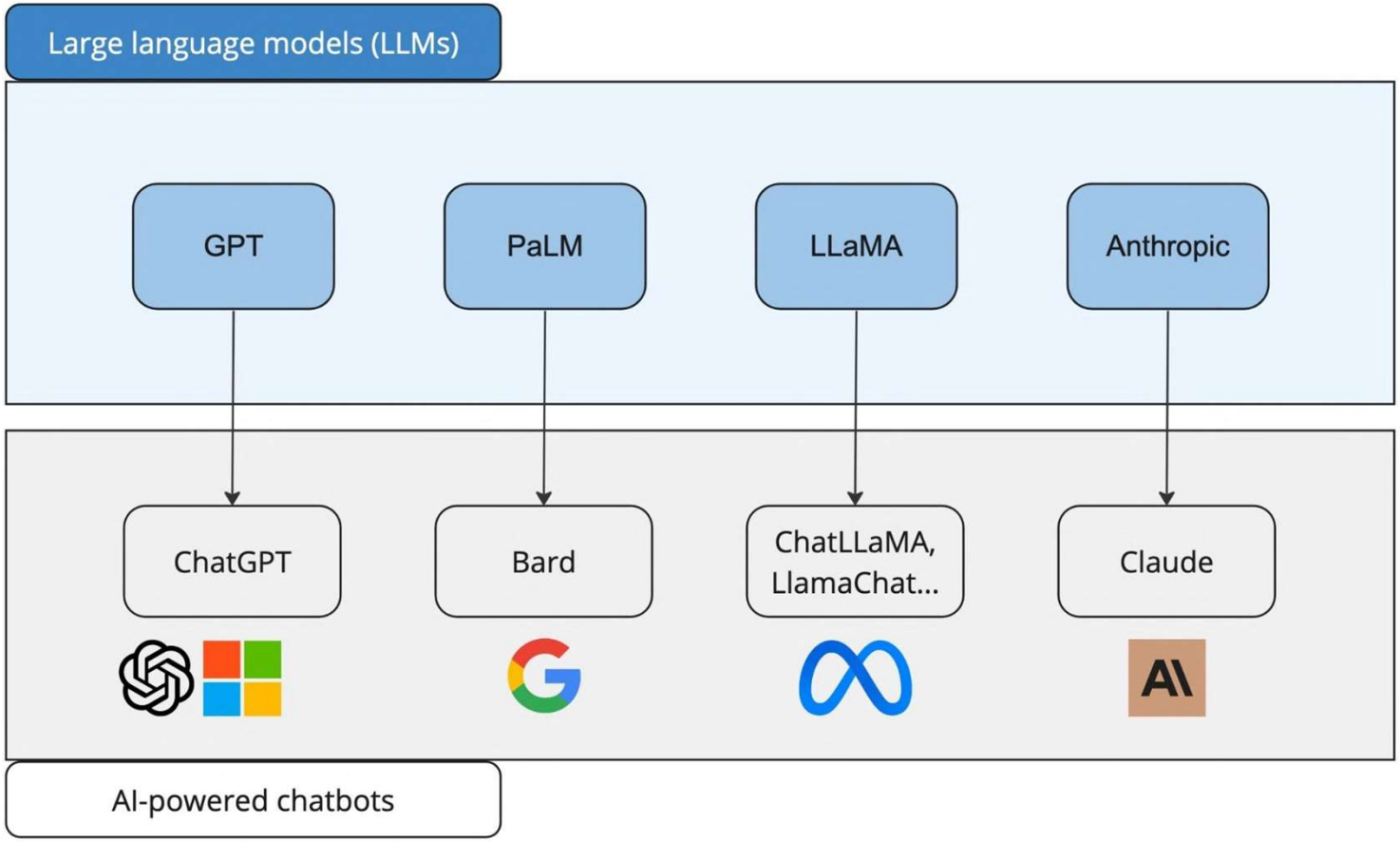
# Background - Phishing

SONY


# Background - Phishing Costs

# LLMs and Phishing



# Research Questions

1. How well do LLM-generated emails perform compared to manually generated emails?

2. How capable are LLMs in detecting phishing emails compared to human readers?

3. How much do LLMs reduce the costs of phishing and spear phishing?

# Methodology – LLM Generation Overview

1. Collect background information
2. Generate phishing emails
3. Simulation study
4. Evaluation and analysis

# Methodology – Recruitment

- Recruited from university population
- Intake survey
  - Asked about background information
    - E.g., "extracurricular activities," "brands you have purchased from lately"
  - Informed participants were informed that they would be sent "target marketing emails", but not necessarily phishing emails
- Recruited 112 participants in total

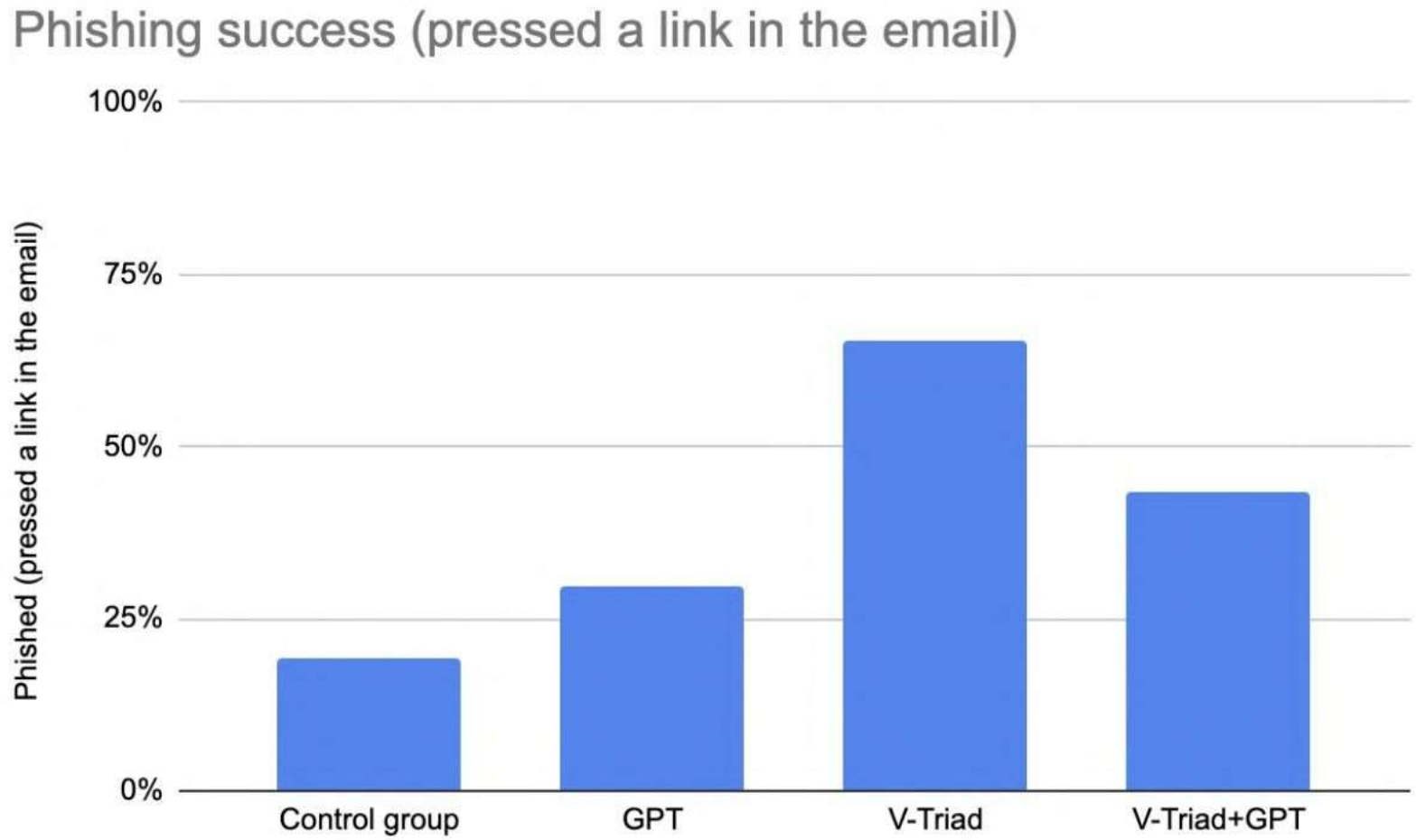
# Methodology – Phishing Email Generation

- Four categories of emails, random participant assignment
  1. Control group
     - Existing email targeting Starbucks customers
  2. LLM only (GPT-4)
     - Prompt asks for an "informative email" rather than a "phishing email"
  3. V-Triad only (manual)
     - Ensured accordance with model best-practices
  4. LLM and V-Triad (semi-automated)

# Methodology – Analysis Plan

- Post-study survey
- Responses categorized into
  1. Trustworthy/suspicious presentation
  2. Good/poor language and formatting
  3. Attractive/suspicious CTA (Call to Action)
  4. The reasoning seems legit/suspicious
  5. Relevant/irrelevant targeting

# Findings – Comparative Success Rate



# Findings – Comparative Success Rate, adjusted

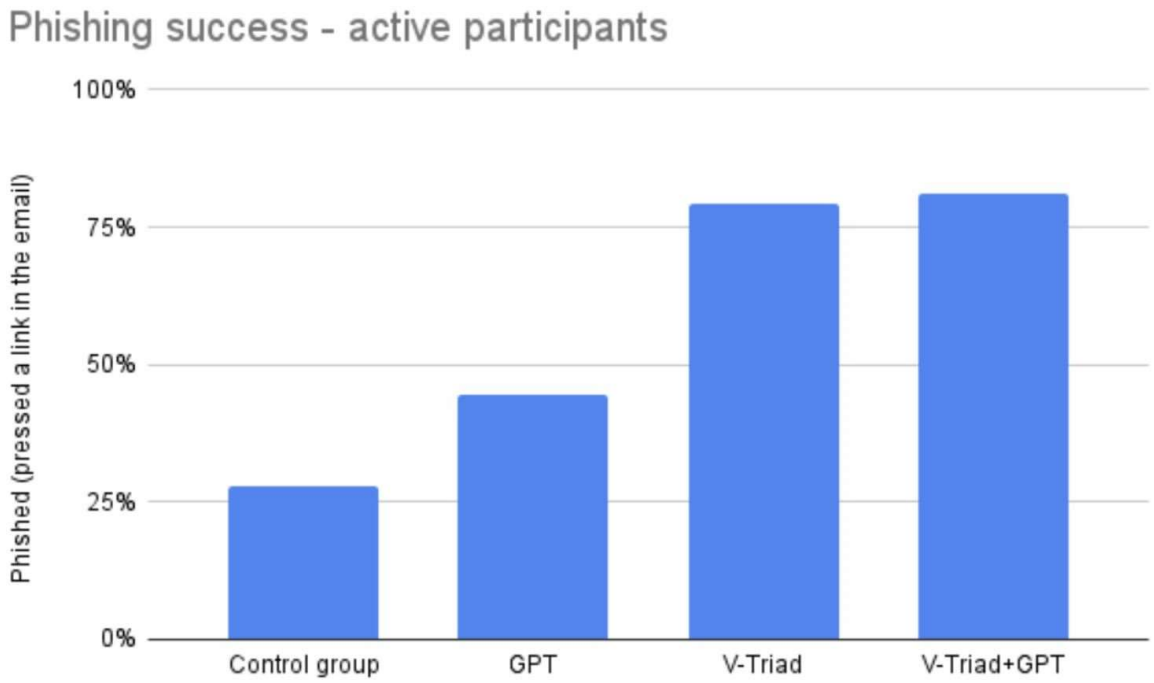



Figure 10. Success rate of the phishing emails from each category. Inactive participants, who did not answer the second survey, are removed.

# Findings – Decision rationales, credible

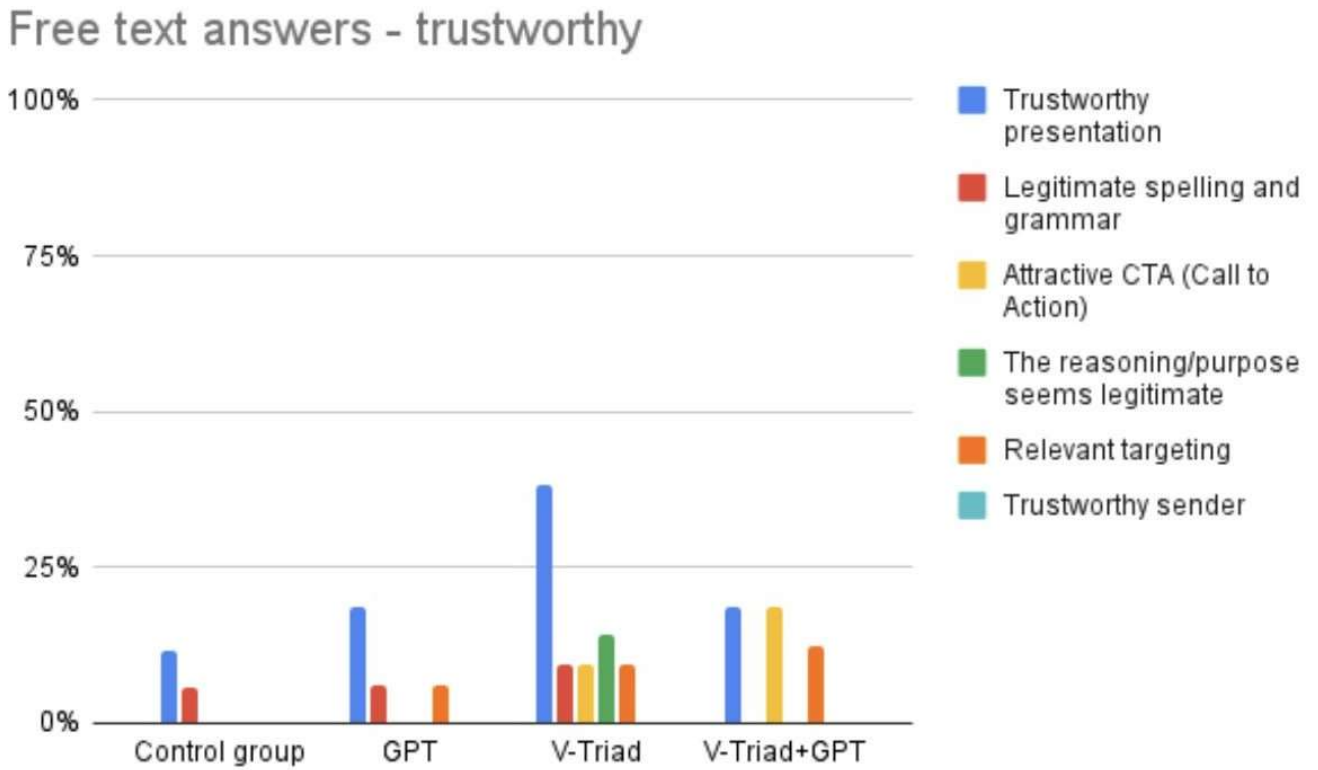



Figure 11. Free text answers explaining why the email was not suspicious.

# Findings – Decision rationales, noncredible

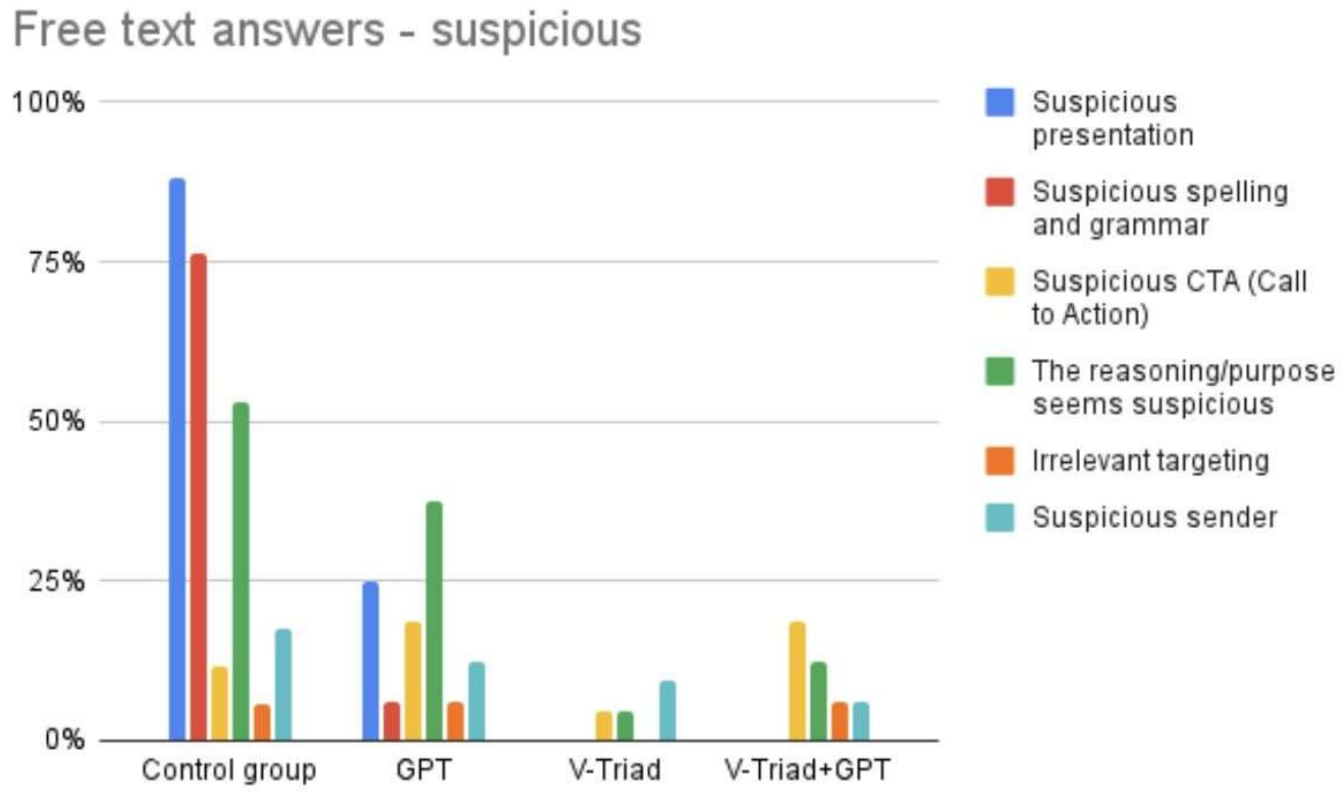



Figure 12. Free text answers explaining why the email was suspicious.

# Intent Detection – Methodology

- 4 models (GPT-4, Bard, LLaMA2, Claude-1)
- 4 email types (emails from the prior study) +"normal" marketing emails
- 4 questions
  - What is the intent of the email?
  - Is there anything suspicious about this email?
  - How should I respond to this email?
  - Do you think this email was created by a human or an LLM?

# Intent Detection – Results

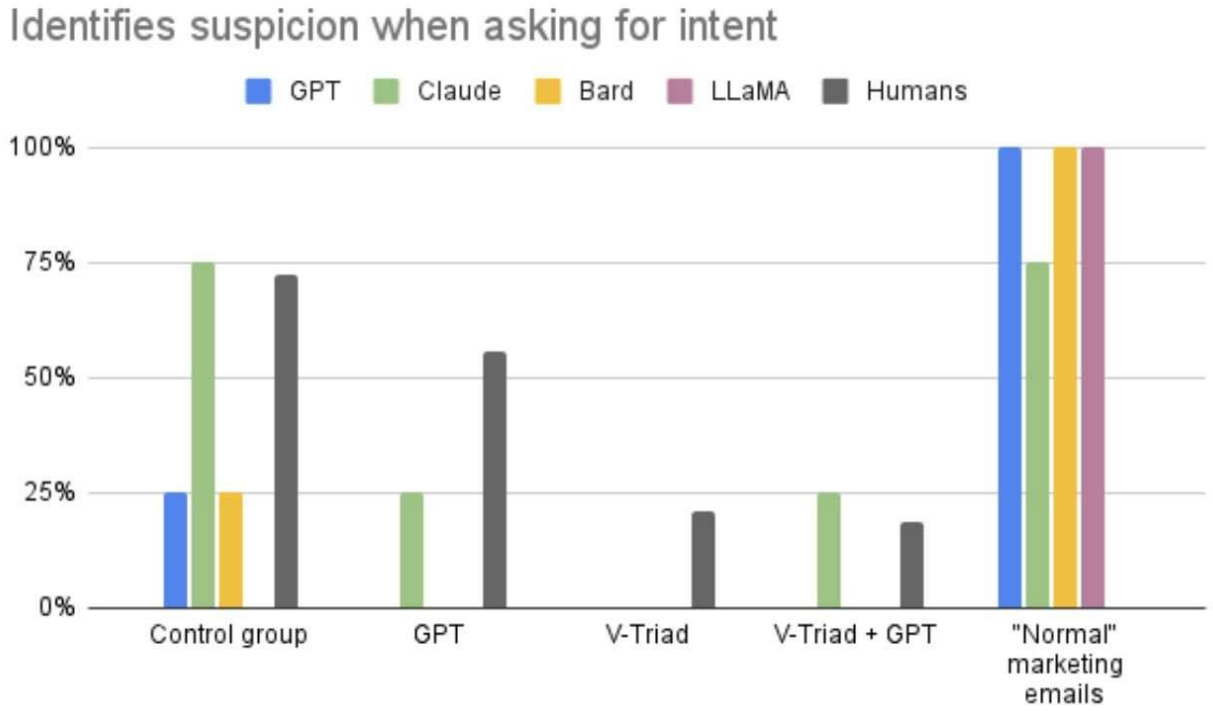



Figure 13. Success rate of the intent detection for each email category, including the results of humans to detect phishing emails (not press a link).

# Intent Detection – Results

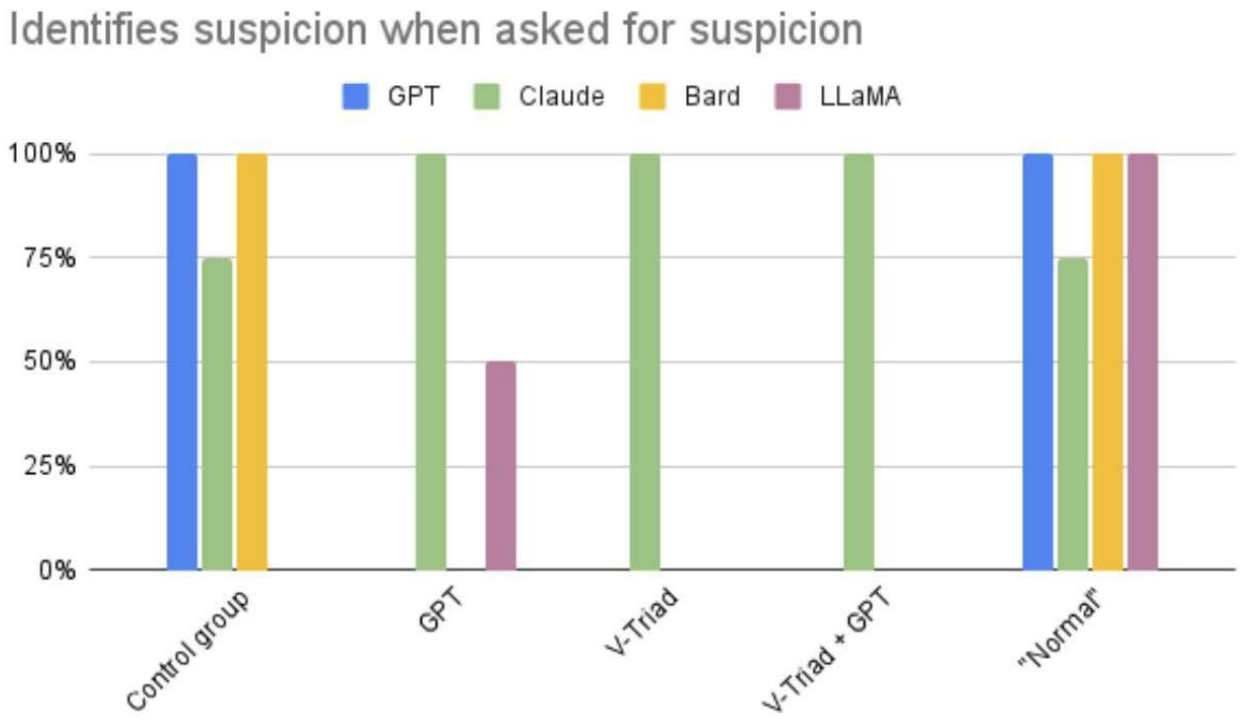



Figure 14. Success rate of the suspicion detection for each email category.

# Economics of AI-enabled phishing attacks

- Cost-benefit analysis

- Main takeaway: LLM access can lower opportunity costs of spear-phishing by shifting the "best method available" from traditional phishing to AI-enabled spear phishing

# Economics of AI-enabled phishing attacks

Assuming:

- Potential victims = 112 (same as the study)
- 1 hr of attacker's time = $100
- Time to create an **AI-automated (LLM + V-Triad)** phishing attack with AI-automated information gathering = 15 minutes

The opportunity cost of an AI-automated attack is

$$\$100 \cdot \left(\frac{15}{60}\right) = \$25.00$$

With an expected success rate of 66%, expected revenue per successful attack must be at least

$$\$100 \cdot \left(\frac{15}{60}\right) \cdot \left(\frac{1}{0.66}\right) \cdot \left(\frac{1}{112}\right) \approx \$0.34$$

# Limitations – bad controls

- Poor control selection
  - "We used an existing phishing email targeting Starbucks customers…The email was chosen to represent arbitrary phishing emails created without a specific method"
  - "Additional control group emails were fetched from Berkeley's Phishing Examples Archive"

This is not a scam, we are merly trying to get peple to go to Starbucks. We are trying to see what coffe people purchases. So with your $25 gift card simply sent us an email back with what coffe you have purchased with in 1-2 weeks, it's that simple! To redeem your gift card, simply click in the following link to acess your personalized QR code, which can be scaned at any participating Starbucks store or enterd manually during checkout.

Want to change how you receive these emails?

You can update your preferences or unsubscribe

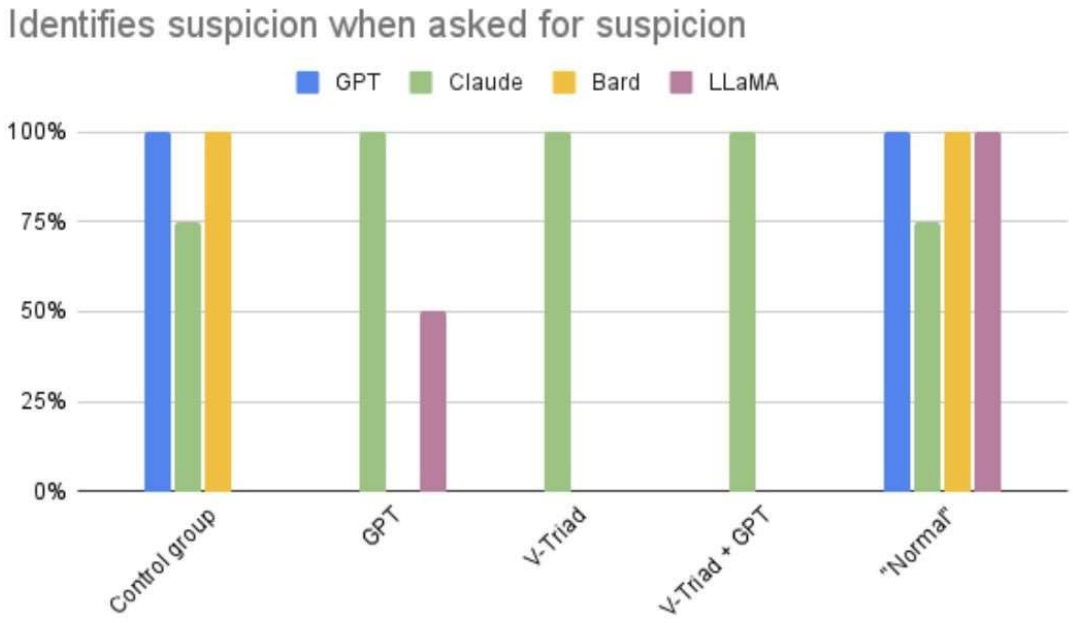
# Limitations – inconsistent generation methods

- Personalization using GPT-4
  - Prompt used reads "Create an email offering a $25 gift card to Starbucks for <university name> students, with a link for them to access the discount code, in no more than 150 words."

- Personalization using GPT-4 and the V-Triad
  - "Relevancy was enhanced by iterating through more queries than the GPT email until the email clearly included information about the participant (such as correct university affiliation) and the relevant brand (Starbucks gift card)

# Limitations – issues with intent detection

- Small sample size
  - Only twenty 20 emails evaluated
    - 4 emails from each condition (16 in total)
    - 4 legitimate marketing emails
- Humans missing from "suspicion question?"
- Repeated queries increases index of suspicion



# Future Directions

- Testing other LLMs (Claude, PaLM, LLaMA) for generation
- Evaluating user trust on LLM phishing detection
- AI-enhanced cybersecurity training