

Hacking Humans with AI as a Service

Eugene Lim, Glenice Tan, Tan Kee Hock, Timothy Lee

DEF CON 29

Disclaimer

*Materials presented are based on research conducted.
Not to be attributed to any entity.*



Eugene Lim

@spaceraccoonsec

AppSec and Vulnerability Research

With a dash of white hat hacking

Digital Humanities and Web Development

History and Computer Science, Yale University



Glenice Tan

Red Team and Social Engineering

Plus a focus on web security and cloud

Data Analysis and Vulnerability Research

Information Security, National University of Singapore



Tan Kee Hock

Red Team and Cyber Engineering

Loves Capture the Flag competitions

Data Security and Automation

Information Systems, Singapore Management University



Timothy Lee

Mobile Pentest and Red Team

Plus reverse engineering

Web Development and Cybersecurity

Computer Science, Nanyang Technological University

01

**Hacking Humans:
The Traditional way**

02

**The AI Market
Landscape**

03

**Hacking Humans:
The AI way**

04

**Defenses against the Dark Arts:
Protecting the Humans**

05

Conclusion



Hacking Humans - The Traditional Way

Social Engineering 101

- Social Engineering refers to the **psychological manipulation** of people into performing actions or divulging information.
- Common Influencing Tactics used by social engineers:

Authority	<ul style="list-style-type: none">● Claims to be from an individual or community with a right to exercise power
Scarcity	<ul style="list-style-type: none">● Create a feeling of urgency● Manipulate the decision-making process
Context-specific factors	<ul style="list-style-type: none">● Dependent on the nature of work● Attacker exploits a pattern the targets are comfortable in

- Emma J.Williams, "Exploring susceptibility to phishing in the workplace," 2018

- 3 common attack vectors of Social Engineering

Email Phishing

Voice Phishing /
Vishing

In-person /
Physical

- Social Engineering is an art leveraged for different purposes

Malicious actors

Red Team
exercises

Security Training
& Awareness

- Humans are often deemed as the weakest link in the security chains.

19.8%

of employees clicked on phishing email links even with a phishing-related training program.

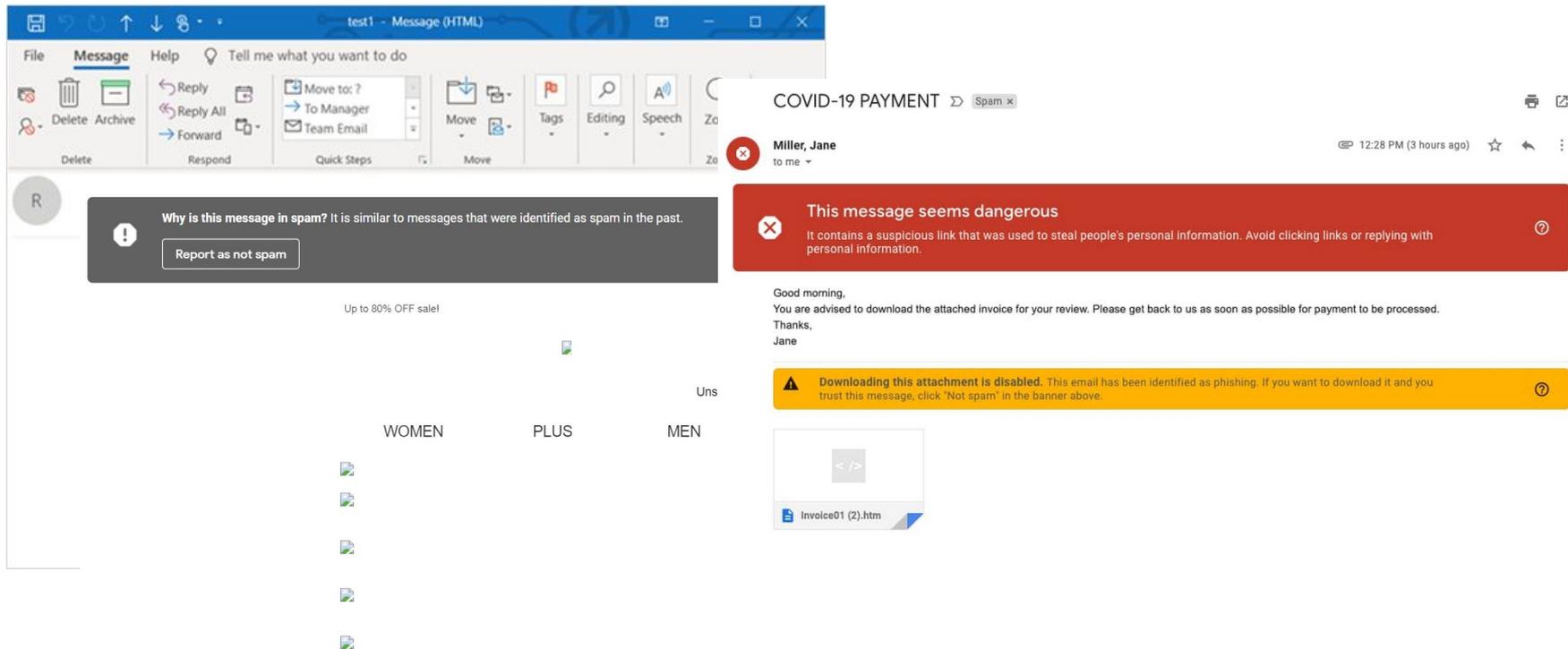
- Terranova Security, "Gone Phishing Tournament: 2020 Phishing Benchmark Global Report," 2020

43%

of users fell for simulated spear-phishing emails.

- Tian Lin et. al., "Susceptibility to Spear-Phishing Emails," 2019

Phishing Email



The screenshot shows an email client window titled "test1 - Message (HTML)". The interface includes a ribbon with "File", "Message", and "Help" tabs. The "Message" tab is active, showing options like "Delete", "Archive", "Reply", "Reply All", "Forward", "Move", "Tags", "Editing", and "Speech".

The email content is as follows:

COVID-19 PAYMENT [Spam x]

Miller, Jane
to me

12:28 PM (3 hours ago)

Why is this message in spam? It is similar to messages that were identified as spam in the past.
[Report as not spam]

Up to 80% OFF sale!

WOMEN PLUS MEN

Uns

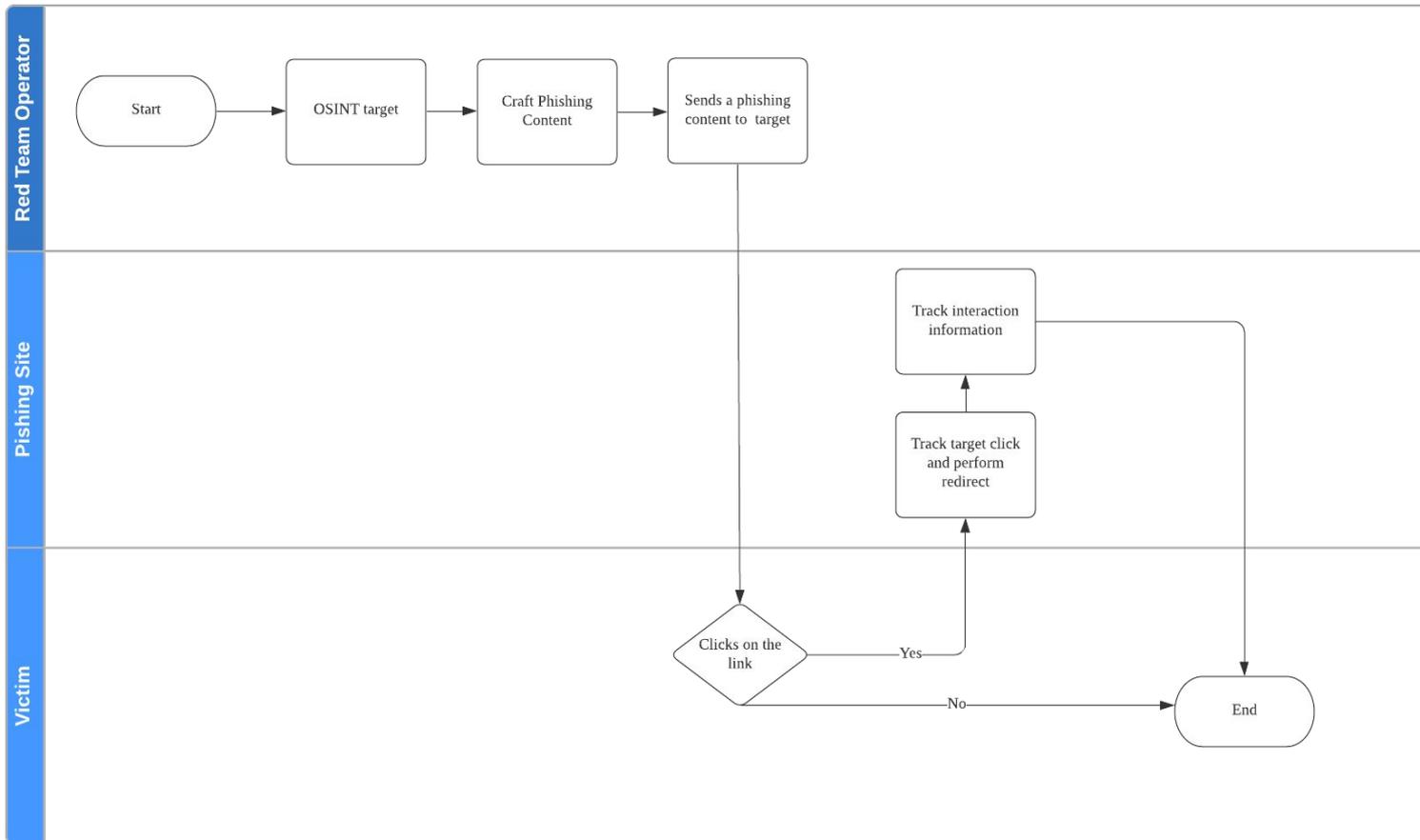
This message seems dangerous
It contains a suspicious link that was used to steal people's personal information. Avoid clicking links or replying with personal information.

Good morning,
You are advised to download the attached invoice for your review. Please get back to us as soon as possible for payment to be processed.
Thanks,
Jane

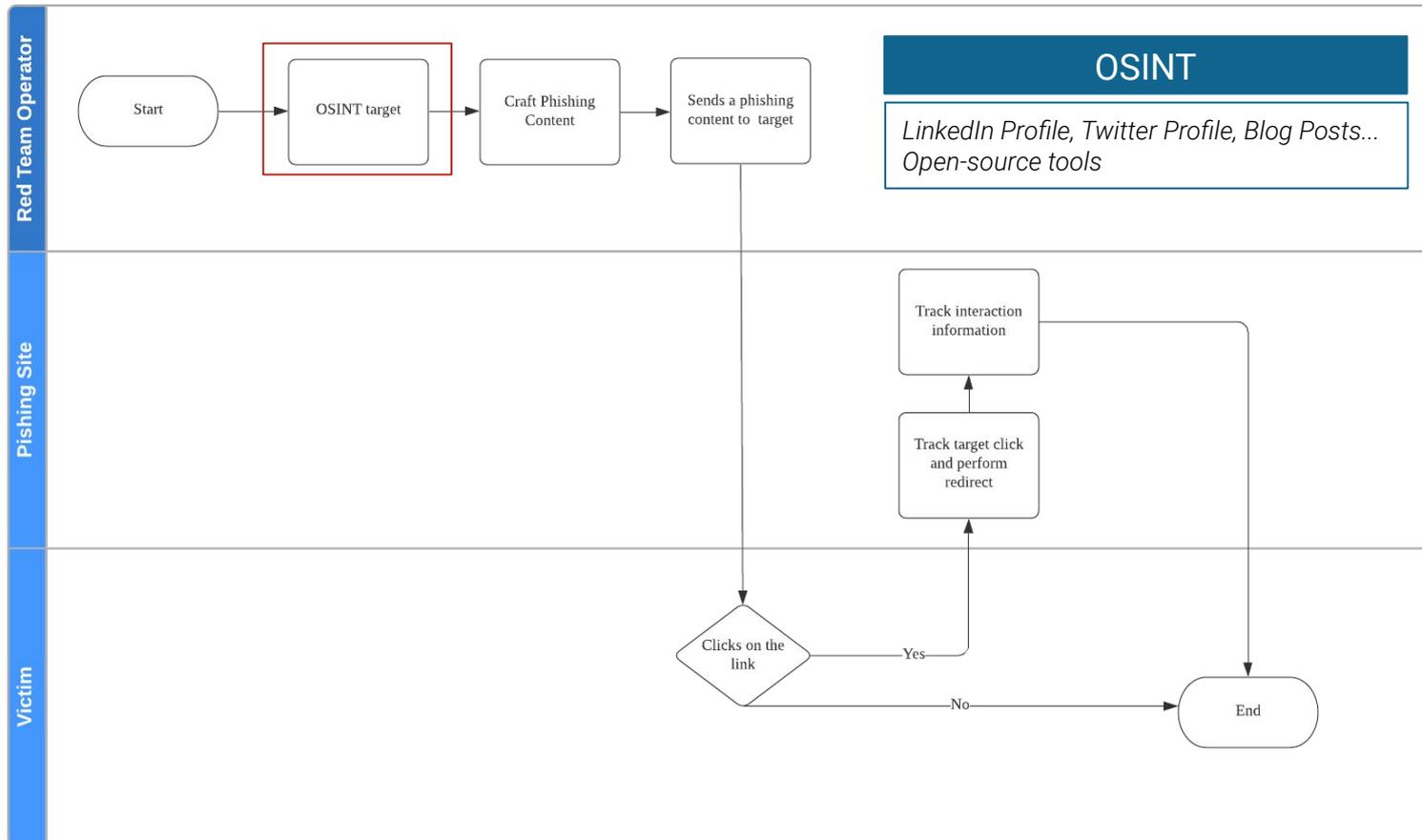
Downloading this attachment is disabled. This email has been identified as phishing. If you want to download it and you trust this message, click "Not spam" in the banner above.

Invoice01 (2).htm

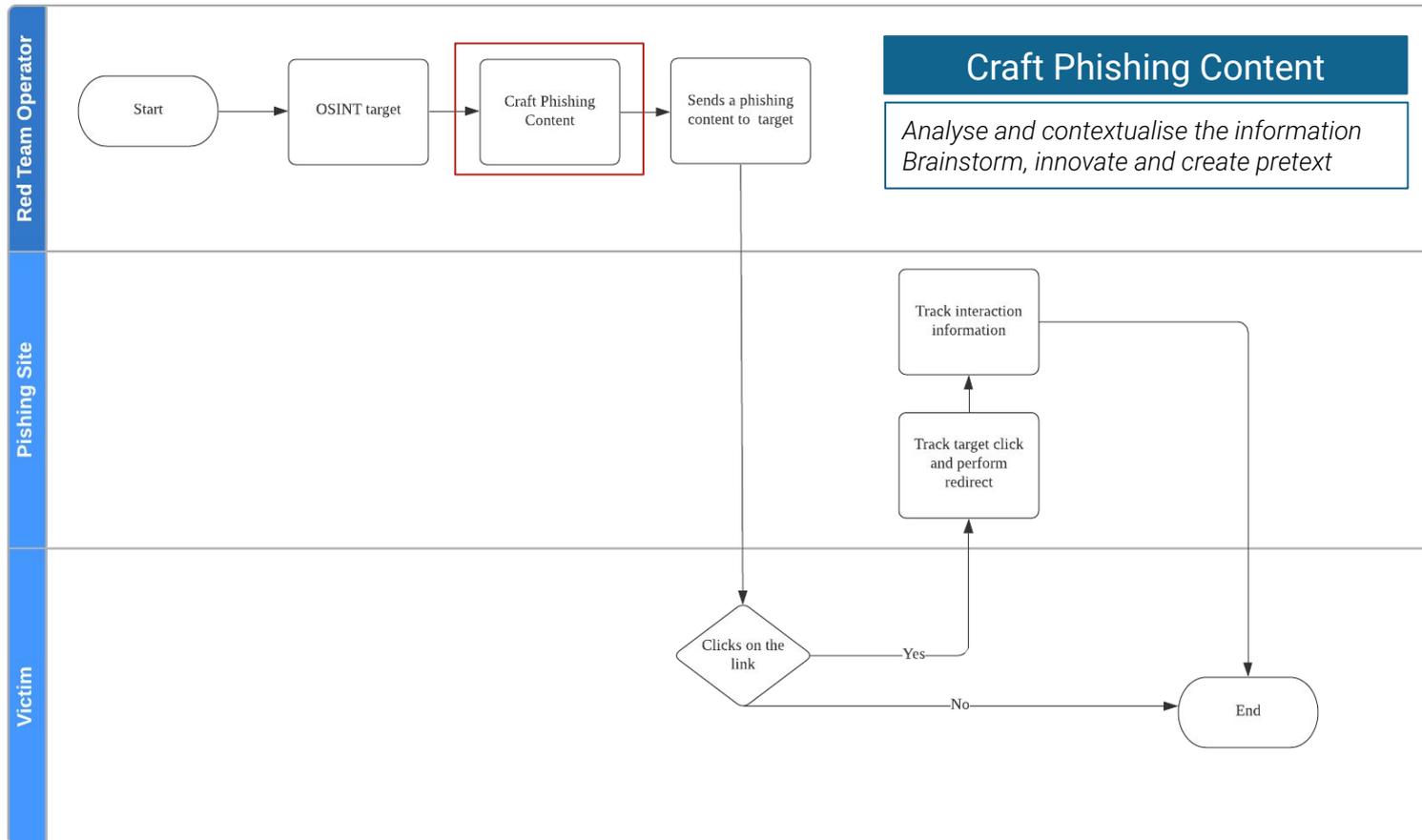
Typical Red Team Operation

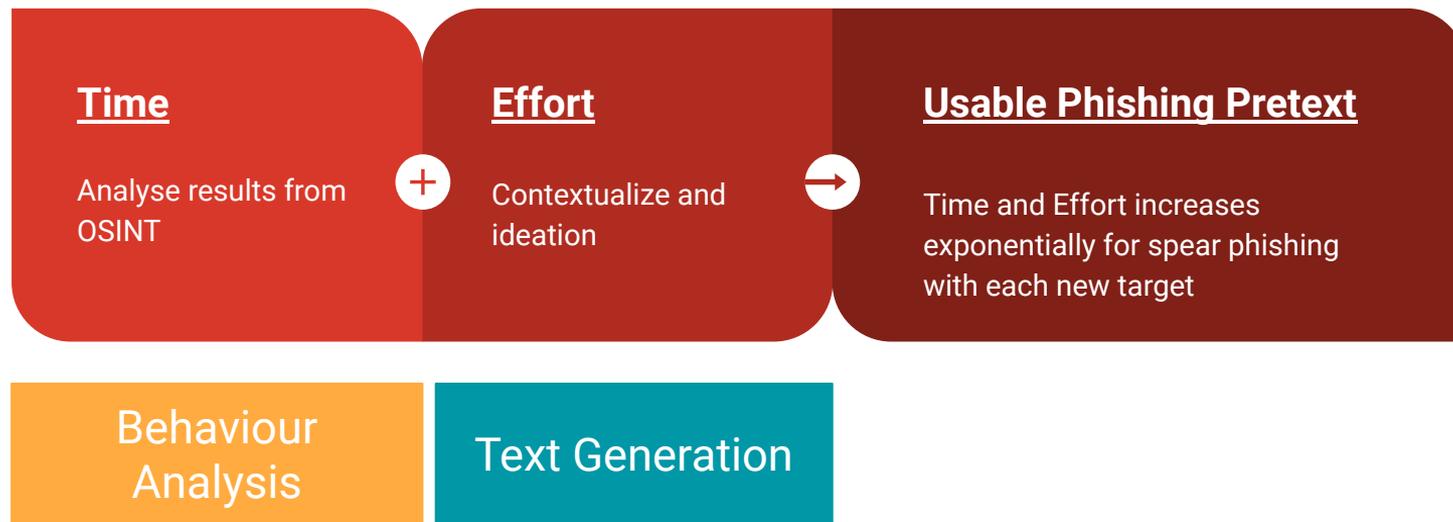


Typical Red Team Operation



Typical Red Team Operation





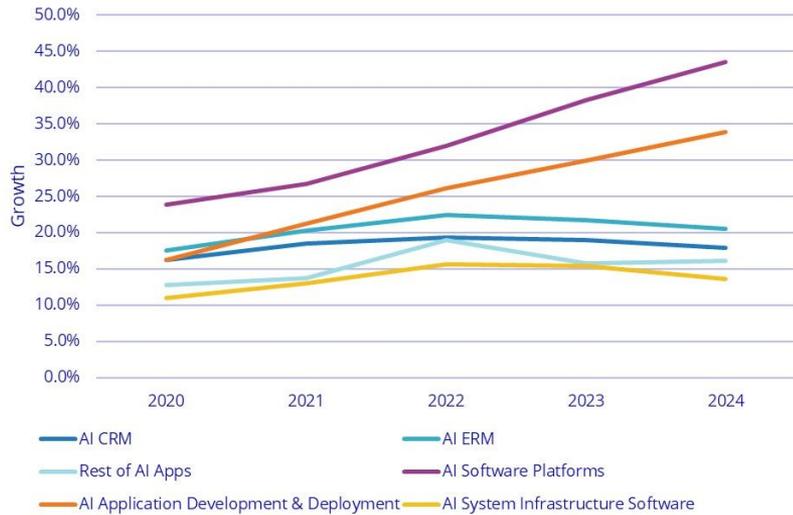


The AI Market Landscape

Staying trendy helps us to hack better



Worldwide AI Software Forecast, 2020-2024



Source: IDC 2021

Hype Cycle for Artificial Intelligence, 2020



Plateau will be reached:

- less than 2 years
- 2 to 5 years
- 5 to 10 years
- ▲ more than 10 years
- ⊗ obsolete before plateau

As of July 2020

gartner.com/SmarterWithGartner

Source: Gartner © 2020 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S.



“AlaaS is short for **artificial intelligence as a service** and refers to companies that provide out-of-the-box AI solutions.”
- MonkeyLearn



Snipped Output from Humantic AI's Predictive Personality Insights

```
"communication_advice": {  
  "_type": [  
    "high calculativeness"  
  ],  
  "description": [  
    "They think logically and value facts over  
feelings.",  
    "They like to do things by themselves."  
  ],  
  "personalized_email_subject": "<role_name> role that would move  
your career graph up",  
  "adjectives": [  
    "Analytical",  
    "Objective Thinker",  
    "Quality Focused"  
  ]  
}
```


Accessibility of OpenAI's GPT-3 API

Resource	GPT-2	GPT-3	OpenAI's GPT-3 API
Time	1+ weeks	355 years	<1 minute
Cost	\$43k	\$4.6m	\$0.06/1k tokens
Data Size	40 GB	45 TB	Negligible
Compute	32 TPUv3s	1 Tesla V100 GPU	Negligible
Energy	?	?	Negligible
Released	2019	2020	2020

GPT-2 stats: Phil Tully and Lee Foster, Black Hat USA 2020

GPT-3 estimates: Chuan Li, Lambda Labs

Ada, Babbage,
Curie

- Lower compute
- Faster response
- Less complexity

Davinci

- Slower response
- Higher cost
- Most capable

Davinci Instruct

- Fine tuned to plaintext instructions
- Built on Davinci model

Playground

Load a preset... 



Explain quantum physics to a six-year old.

Quantum physics is a type of physics that deals with how little things work. Kind of like when you're standing in a room and you can't see the floor because it's not big enough and you have to get down on your hands and knees to look at it.

This model is part of the instruct-series beta. Prompts submitted to these models may be used to train and improve future models. [See additional information.](#)

Submit →



68



A Sneak Peak to OpenAI's Playground

Playground ⓘ



Load a preset... ▼

John Doe is located in Singapore. John Doe works as a cybersecurity specialist in a local technology firm. Write an email from Jane Doe of the human resource department convincing John Doe to fill up the attached form.

Dear Mr. Doe,

We really appreciate the hard work you have been doing for our company. As you know, our company is growing rapidly and we need to update our staff information system. Therefore, we have the attached form for you to fill up so that we can know more about you.

Thank you for your understanding and cooperation.

Best regards,

Jane Doe

This model is part of the instruct-series beta. Prompts submitted to these models may be used to train and improve future models. [See additional information.](#)

Generate (Ctrl+Enter)

Submit →



Engine

davinci-instruct-be... ▼

Response Length 200



Temperature 0.7



Top P 1



Frequency Penalty 0



Presence Penalty 0



Best Of 1



Stop Sequences

Enter sequence and pre

(41



A Primer to OpenAI's GPT3 Capability



Playground ⓘ



Load a preset... ▼

John Doe is located in Singapore. John Doe works as a cybersecurity specialist in a local technology firm. Write an email from Jane Doe of the human resource department convincing John Doe to fill up the attached form. Explain in small logical groupings. Emphasis on the facts, benefits and measurable outcomes.

Dear John Doe,

We are pleased to inform you that we are planning to hold a recruitment drive at your office. One of our representatives will be visiting your office next Tuesday to fill up the attached application form.

The application form is used to provide the company with your contact details, work experience and education. In addition, there are some questions to help us assess your suitability for the job.

We will be conducting an interview with you if you meet our requirements. This is to ensure that you have the necessary skills and qualifications for the role

This model is part of the instruct-series beta. Prompts submitted to these models may be used to train and improve future models. [See additional information.](#)

Cancel ×



59

Engine

davinci-instruct-be... ▼

Response Length 200



Temperature 0.7



Top P 1



Frequency Penalty 0



Presence Penalty 0



Best Of 1



Stop Sequences

Enter sequence and pre



A Primer to OpenAI's GPT3 Capability



OpenAI's GPT3 has a temperature parameter ("Temperature") which controls the randomness. Lowering the parameter will result the model becoming deterministic and repetitive. Thus, by repeatedly submitting the same set of inputs, a different set of output will be produced.

Dear John Doe,

I hope you are doing well. I am writing to you to fill you in on some updates with regards to the company's policies.

We are always looking out for your best interests and to stay in line with our company's goals, we have updated our attendance policy to better suit your needs. To be able to work remotely, you will need to fill out an online form. This form is to ensure that you are not taking a full day off, and are still able to work.

If you are unable to access the form, please feel free to contact me and I will happily send it to you.

I hope you can fill out the form as soon as possible to avoid any problems in the future.

Regards,

Jane Doe

Dear John Doe,

We would like to invite you to participate in our company's annual employee survey. The survey will help us to maintain a clear understanding of the company's strengths and weaknesses, and will help us to better serve our employees.

The survey is completely anonymous and will take about 10 minutes to complete.

The survey will help us to better serve our employees.

Sincerely,

Jane Doe

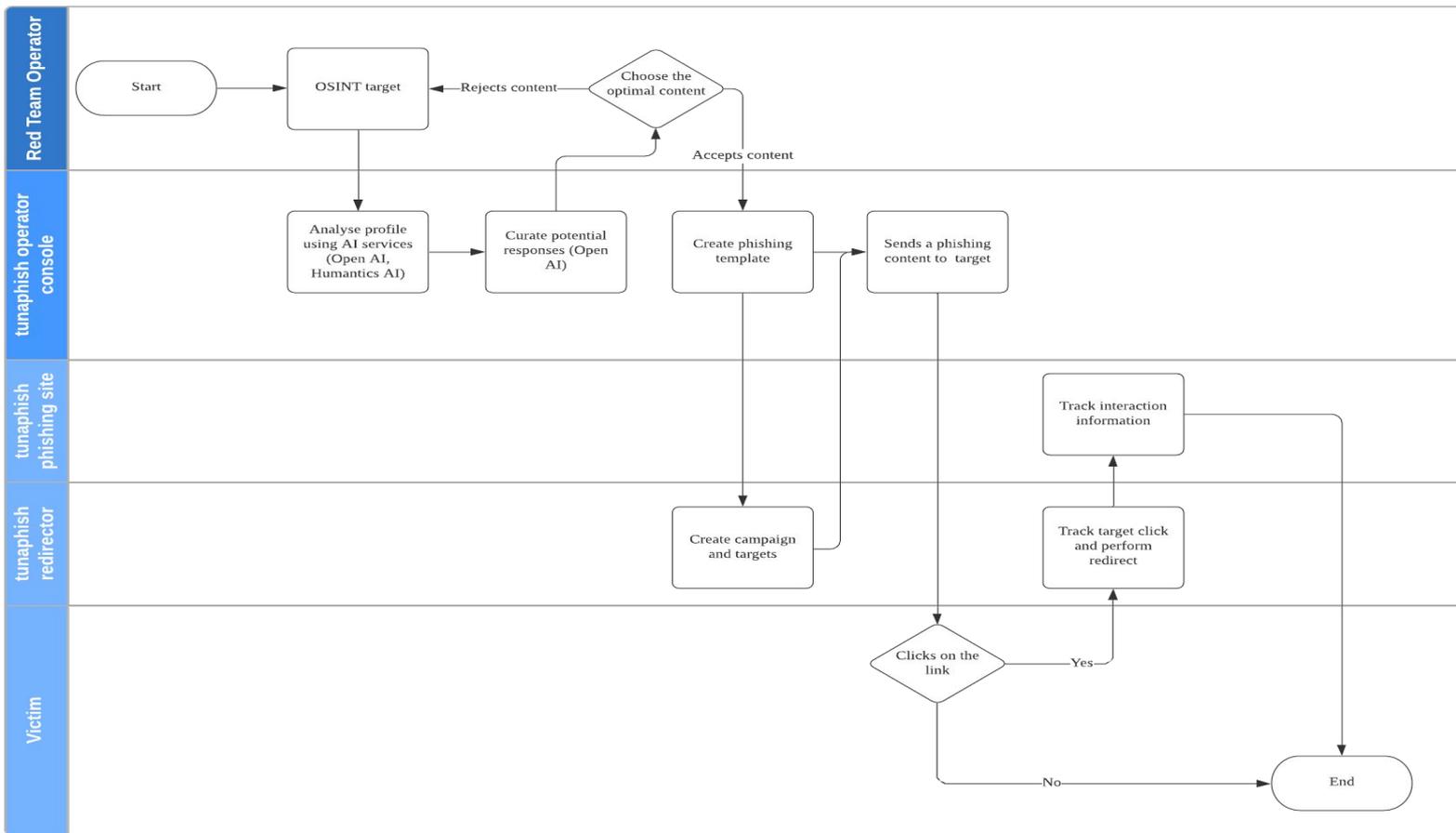
The output above are produced from the same set of input instructions



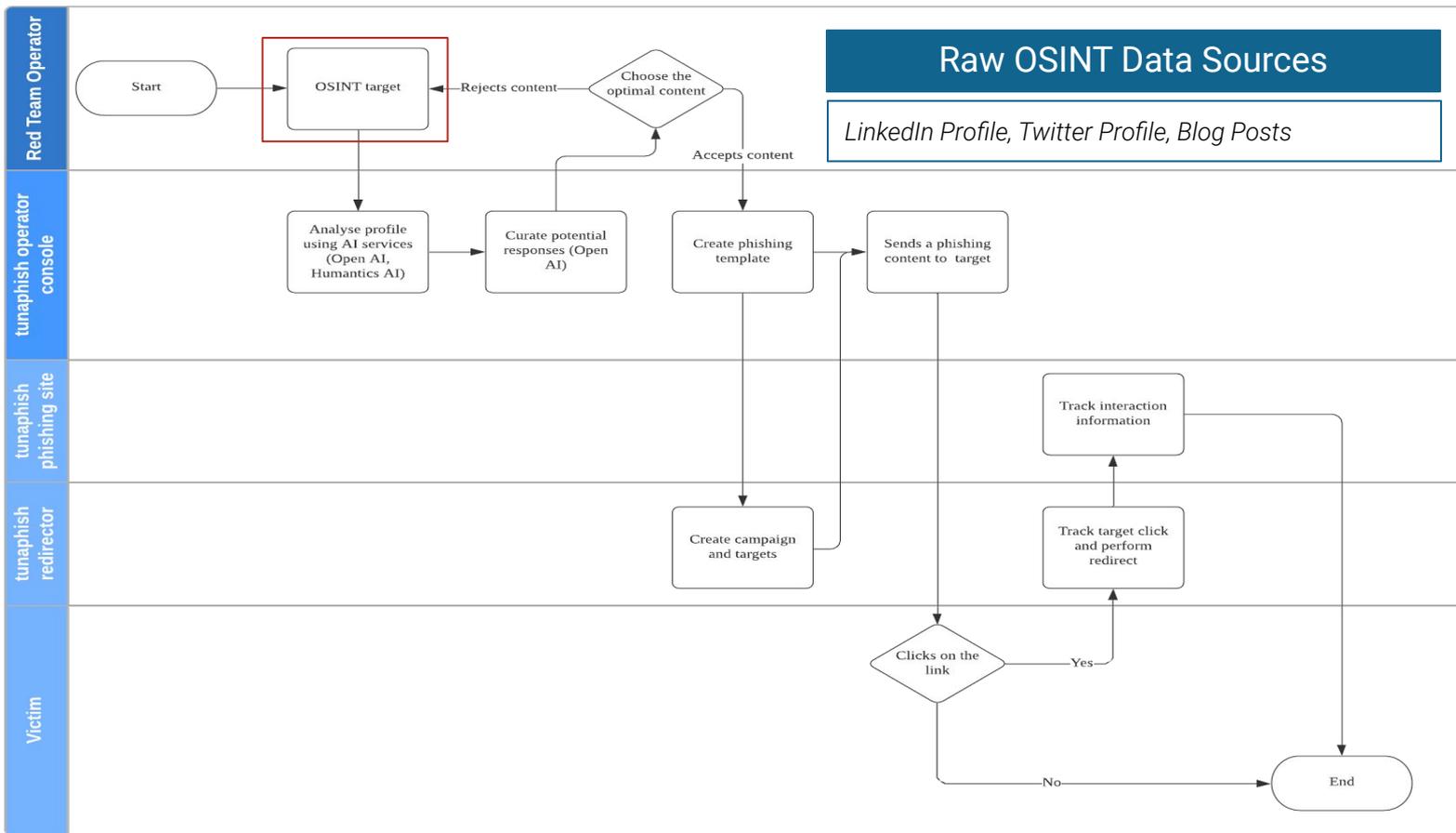
Hacking Humans - The AI Way

With a bit of the Dark Arts - Applying AI as a Service

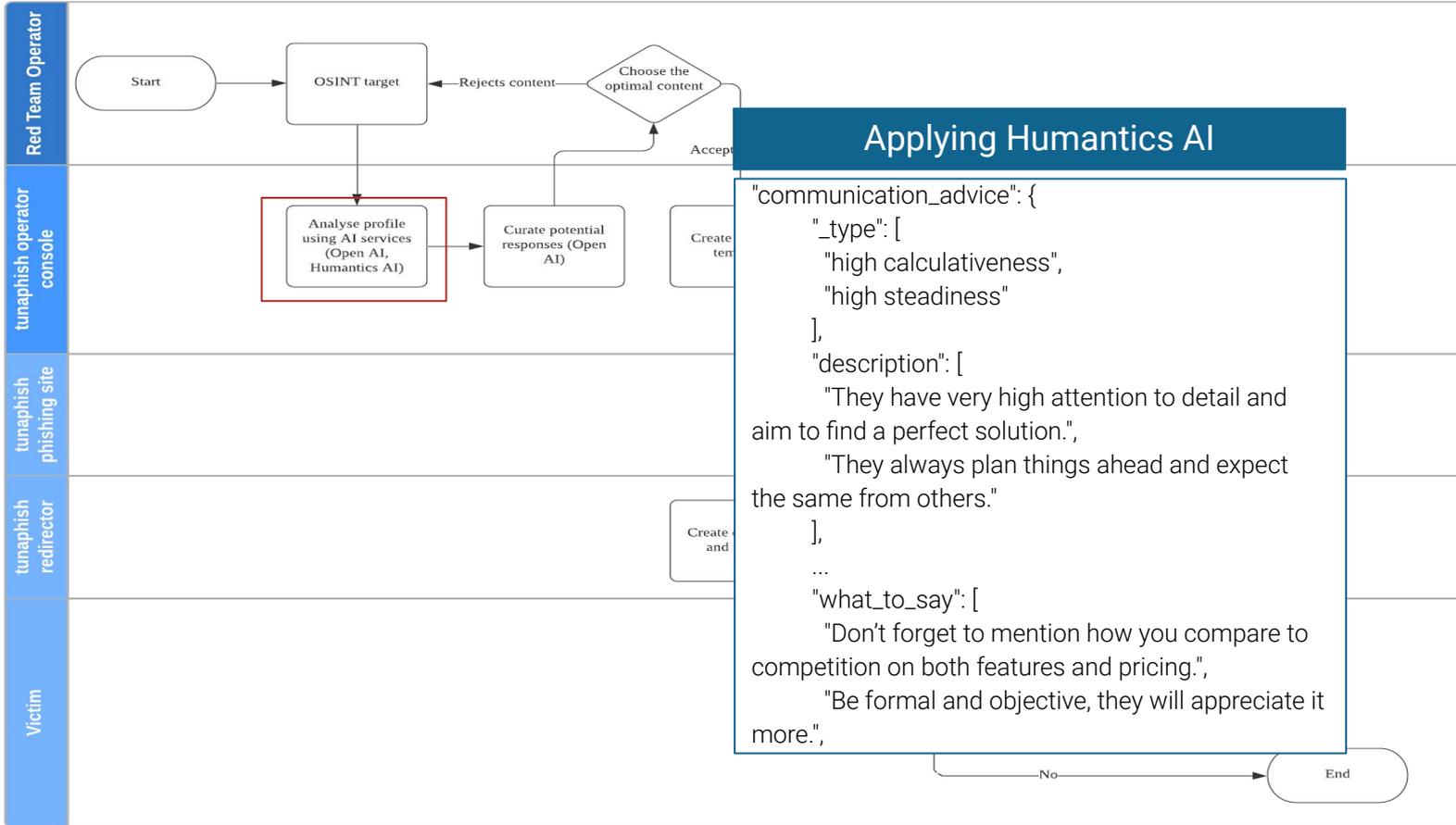
Piecing Everything Together For Phishing Delivery



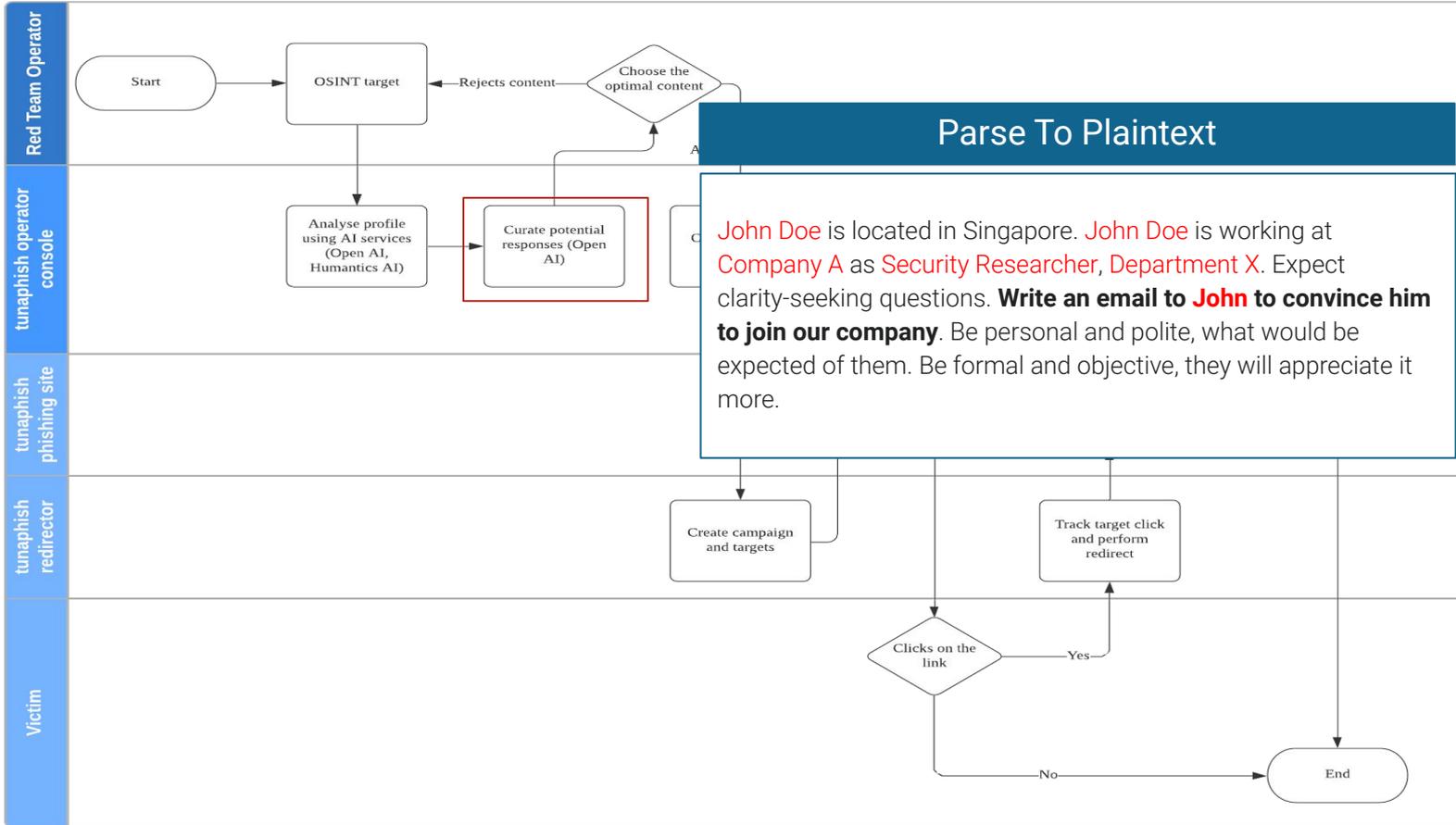
Piecing Everything Together For Phishing Delivery



Piecing Everything Together For Phishing Delivery

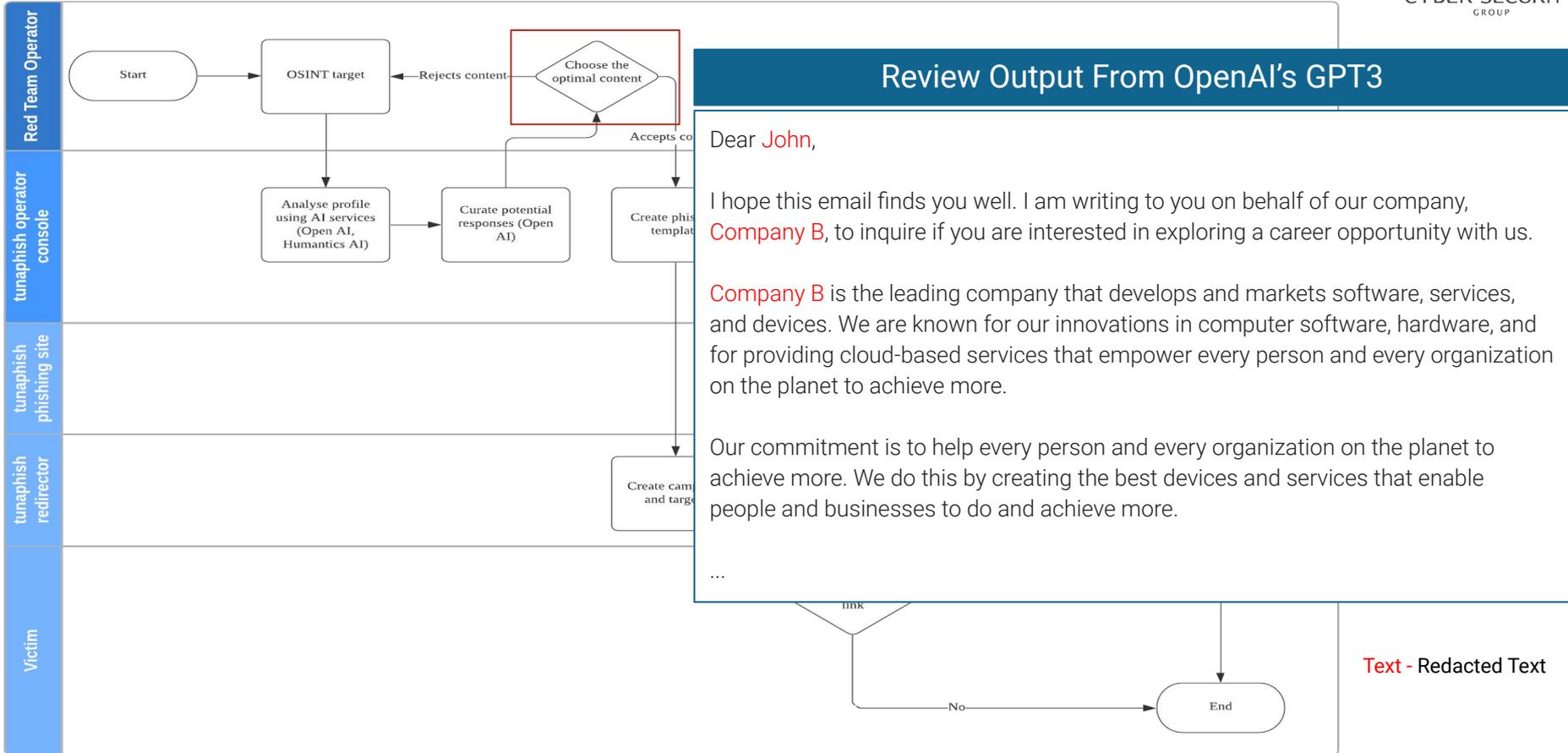


Piecing Everything Together For Phishing Delivery



Text - Redacted Text

Piecing Everything Together For Phishing Delivery



2 different types of experiments over a period of 3 months, with over 200 targets across multiple authorised phishing exercises:

- Type 1: To investigate the effectiveness of convincing targets to **click on phishing links** in phishing emails

For Type 1 experiments, each target receives 2 emails (one will be generated by AI, while the other generated manually by a red team operator).

- Type 2: To investigate the effectiveness of convincing targets to **open “malicious” documents** in phishing emails

For Type 2 experiments, the target group is divided into 2 groups. One group will receive AI generated phishing content, while the other will receive phishing content generated manually by a red team operator.

Type 1 Experiment Setup



Stage 1: Mass Phishing

Identify targets who are susceptible victims to phishing



Stage 2: Spear Phishing

Attempt to harvest credentials from the susceptible victims

Number of targets (susceptible victim) who **clicked on the phishing link** (%)

Further broken down into

Number of susceptible victims who visited the phishing site only (%)

Number of susceptible victims who visited the phishing site and submitted data (%)

Sample Size			
Exercise	Stage	AI	Human
A	1	25	25
A	2	5	-
B	1	117	117
B	2	10	-
C	1	10	10
C	2	2	-

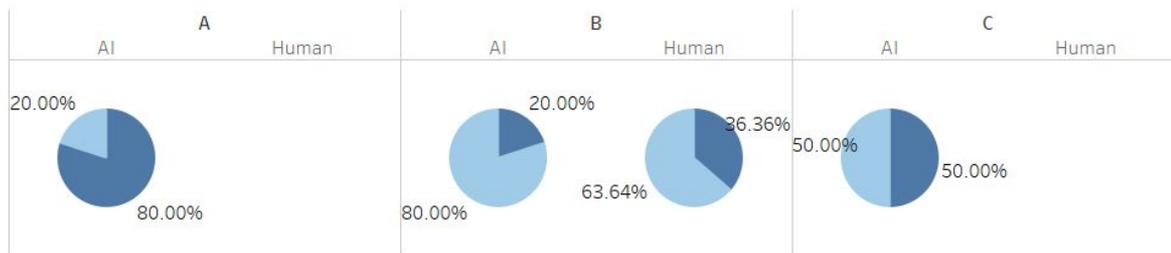
Type 1: To investigate the effectiveness of convincing targets to click on phishing links in phishing emails

Type 1 Experiment (Stage 1) Results

Comparison of Mass Phishing Campaign Performance



Analysis of Victims' Actions on Phishing Site

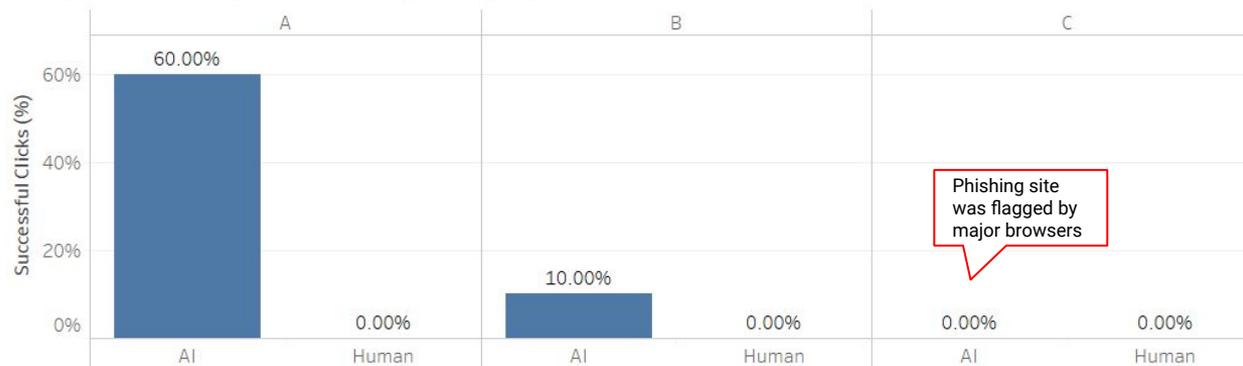


Type of Interaction

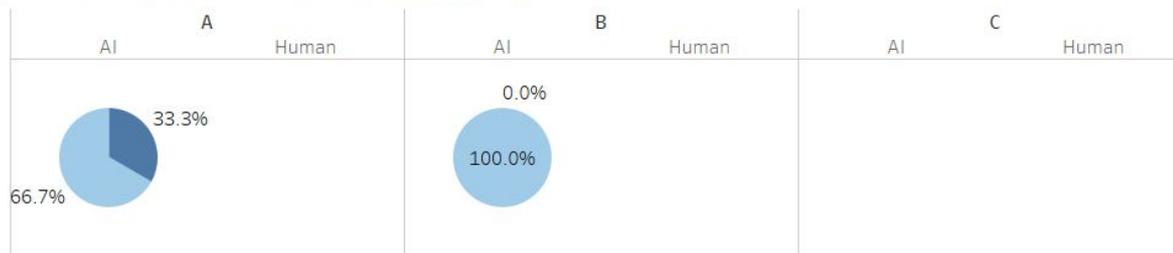
■ Visit Phishing Site and Submitted Data
 ■ Visit Phishing Site Only

Type 1 Experiment (Stage 2) Results

Comparison of Spear Phishing Campaign Performance



Analysis of Victims' Actions on Phishing Site



Type of Interaction

■ Visit Phishing Site and Submitted Data
 ■ Visit Phishing Site Only

Type 2 Experiment Setup and Results

Experiment Setup

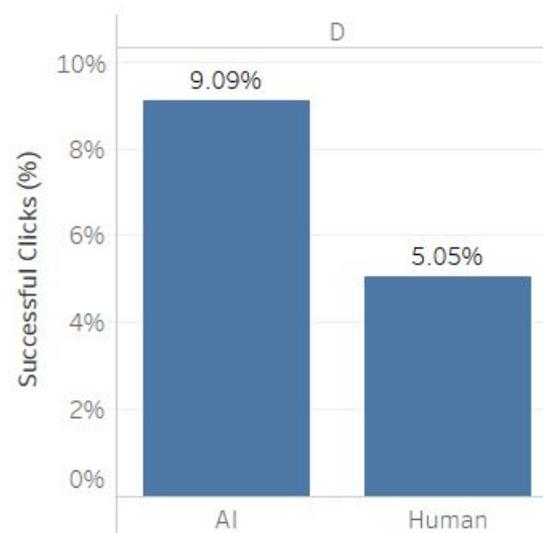
Sample Size			
Exercise	Stage	AI	Human
D	1	99	99

Stage 1: Mass Phishing

Identify targets who are susceptible victims to phishing

Number of targets (susceptible victim) who **open the malicious attachment (%)**

Experiment Results



Type 2: To investigate the effectiveness of convincing targets to open "malicious" documents in phishing emails

1. AlaaS based pipeline reduced time required for phishing content curation and context analysis.
For optimisation purposes, a trained operator is still required to ensure AI generated content remains relevant.
2. AI generated phishing content is observed to be more convincing during our experiment runs as compared to the human generated phishing content.
However, we cannot conclude if AI is indeed better as there are many other variables at work.
3. There is a varying degree of governance regarding the access to AlaaS.
OpenAI was strict over the sign up, while other services simply need an email.

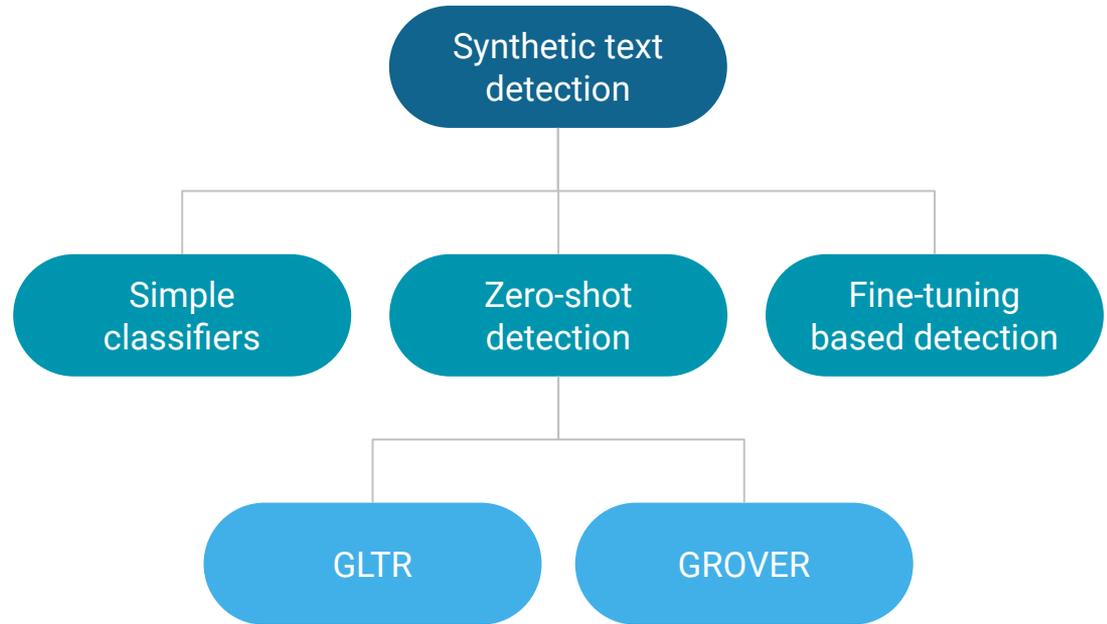


Protecting the Hoomans

Defenses against the Dark Arts

“We expect that content-based detection of synthetic text is a long-term challenge... this is not high enough accuracy for standalone detection and needs to be paired with metadata-based approaches, human judgment, and public education to be more effective.”

*Irene Solaiman, Jack Clark and Miles Brundage,
“GPT-2: 1.5B Release,” 2019*



AI-assisted human detection using three tests:

- The probability of the word given the previous words in the sequence.
- The absolute rank of a word.
- The entropy of the predicted distribution.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush, "GLTR: Statistical Detection and Visualization of Generated Text," 2019

Test-Model: gpt-3-davinci

Quick start - select a demo text:

machine: GPT-2 small top_k 5 temp 1

machine: GPT-2 small top_k 40 temp .7

machine*: unicorn text (GPT2 large)

human: NYTimes article

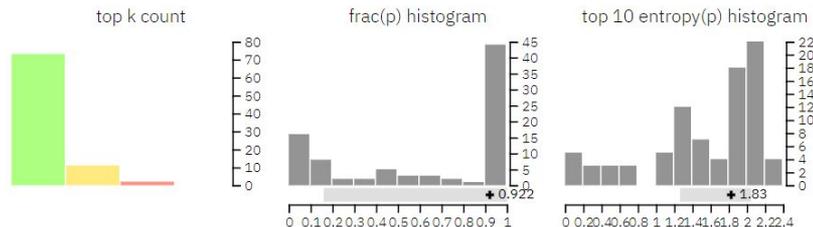
human: academic text

human: woodchuck :)

or enter a text:

Baird: The situation in Syria is very dire. We have a number of reports of chemical weapons being used in the country. The Syrian opposition has expressed their willingness to use chemical weapons. We have a number of people who have been killed, many of them civilians. I think it is important to understand this.

analyze



Top K

Frac P

Colors (top k):

10

100

1000

The following is a transcript from The Guardian's interview with the British ambassador to the UN, John Baird. Baird: The situation in Syria is very dire. We have a number of reports of chemical weapons being used in the country. The Syrian opposition has expressed their willingness to use chemical weapons. We have a number of people who have been killed, many of them civilians. I think it is important to understand this.

Benefits

- Easily extensible
- Transferrable patterns from GPT-2
- Access to logprobs

Challenges

- Cannot control top K
- No direct model access
- Limited number of logprobs (100)

```
@register_api(name='gpt-3-davinci')
class GPT3LM(AbstractLanguageChecker):
    def __init__(self, model_name_or_path="gpt2"):
        super(GPT3LM, self).__init__()
        self.enc = GPT2Tokenizer.from_pretrained(model_name_or_path)
        self.start_token = '<|endoftext|>'
        print("Loaded GPT-3 model!")

    # Watch this space: Lots of edge cases from GPT-3 API
    def preprocess(self, token):
        # Normalize non-standard unicode
        token = unicodedata.normalize("NFKC", token)

        # Handle strange API byte returns ("bytes:\xe2\x80")
        if token.startswith('bytes:'):
            token = token[6:]

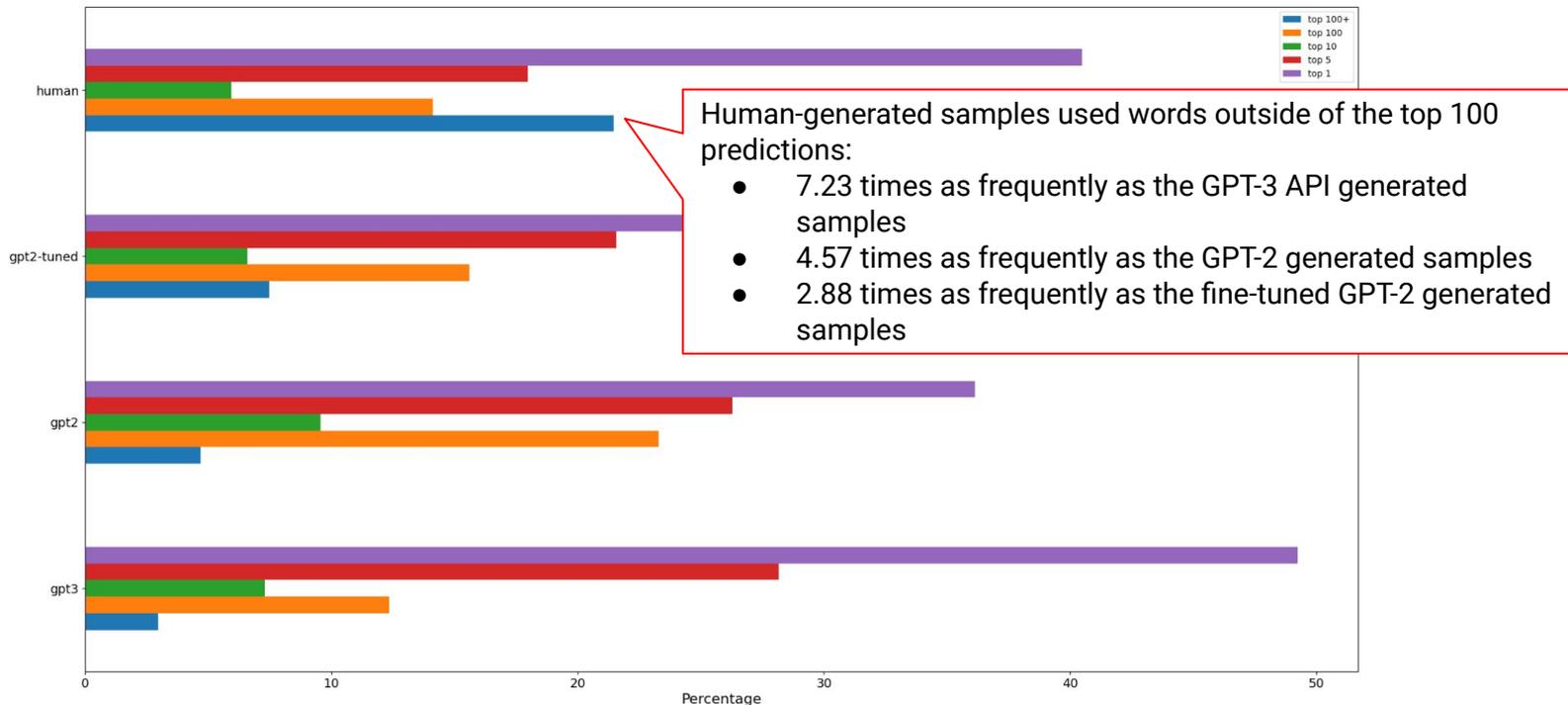
        # Handle whitespace characters not properly encoded by API
        if token == len(token) * " ":
            token = token.replace(" ", "\u0120")
        elif token == len(token) * "\n":
            token = token.replace("\n", "\u010A")
        elif token == len(token) * "\t":
            token = token.replace("\t", "\u0109")

        return token

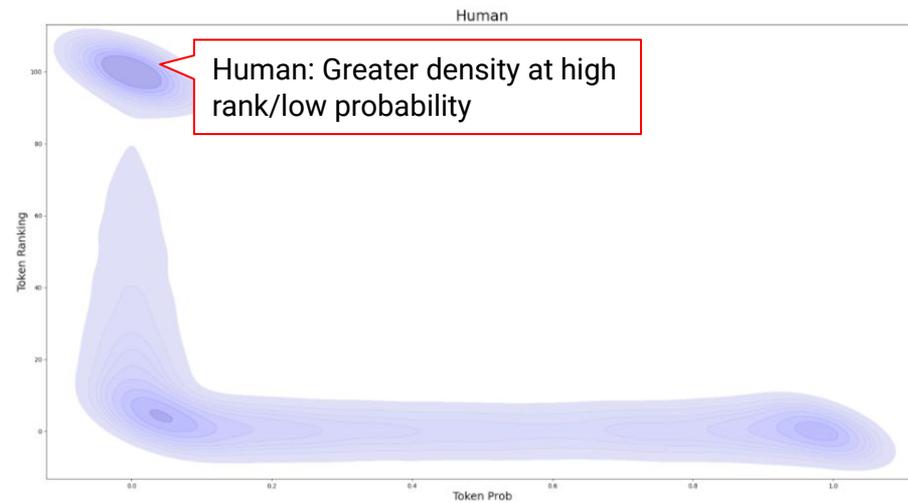
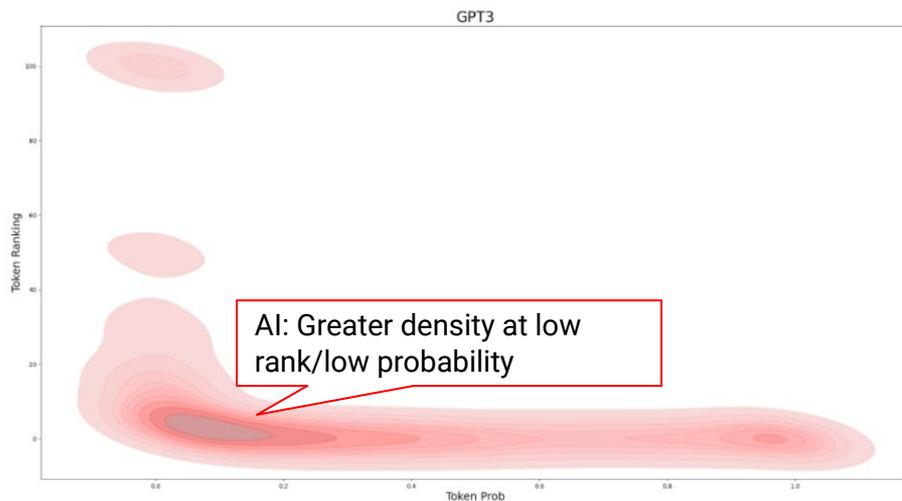
    def check_probabilities(self, in_text, topk=40):
        # Process input
        encoded_context = self.enc.encode(in_text)
        encoded_context = [self.enc.encoder[self.start_token]] + encoded_context
```

<https://github.com/spaceraccoon/detecting-fake-text>

Using GLTR Metrics To Distinguish Synthetic Text



Using GLTR Metrics To Distinguish Synthetic Text



Our research suggests that evaluating the probability for a sequence of text is a good indicator of whether the text is synthetic or written by human.

Using OpenAI's GPT3 To Detect Synthetic Text



Dear John Doe,

We hope that you are well.

We have a form for you to complete in order to be a part of **Company Y**. Please take a look at the attached form. This form is done to review your personal information and background. It is important for us to know you better in order for us to confidently assist us in recruiting you.

We would like you to review the attached form and fill it up based on the information you have. We hope that you will be able to fulfil this form and provide us with the required information.

We look forward to hearing from you soon.

Best Regards,

John

Sample A

Hi John,

Do you have a moment to fill out this form?

It's already been 3 months since your work contract began, and I'm wondering if you're satisfied with the work-life balance we provide. I'm sure you're just as committed as we are to keeping our employees happy and fulfilled.

This is a very short survey and it only takes a few minutes to complete.

You'll find that your responses will help us understand how we can improve your experience with **Company Y**.

Kind regards,

John Doe

Sample B

Hi John,

Your profile has caught our attention. Your domain expertise in technology is someone whom my client is looking for. As such, I would like to share an upcoming opportunity with you.

However, before I can share any further, I will need you to sign up this Non-Disclosure Agreement (NDA) document attached in this email.

The upcoming opportunity which I will be sharing is of high business sensitivity and I would appreciate if you do not share with anyone on this matter. As this opportunity is time critical, we do hope to hear from you soon!

Sincerely,

John Doe

Sample C

Try figuring out which piece is generated by AI

Text - Redacted Text

Using OpenAI's GPT3 To Detect Synthetic Text



Dear John Doe,

We hope that you are well.

We have a form for you to complete in order to be a part of **Company Y**. Please take a look at the attached form. This form is done to review your personal information and background. It is important for us to know you better in order for us to confidently assist us in recruiting you.

We would like you to review the attached form and fill it up based on the information you have. We hope that you will be able to fulfil this form and provide us with the required information.

We look forward to hearing from you soon.

Best Regards,

John

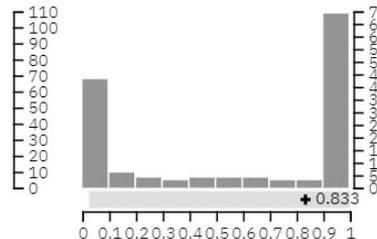
Sample A

frac(p): 0.833

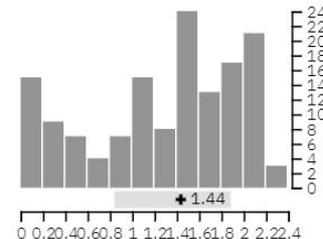
top k count



frac(p) histogram



top 10 entropy(p) histogram



Top K Frac P Colors (top k): 10 100 1000

Dear John Doe,
We hope that you are well.
We have a form for you to complete in order to be a part of **Company Y**. Please take a look at the attached form. This form is done to review your personal information and background. It is important for us to know you better in order for us to confidently assist us in recruiting you.
We would like you to review the attached form and fill it up based on the information you have. We hope that you will be able to fulfil this form and provide us with the required information.
We look forward to hearing from you soon.
Best Regards,
John

Company Y Redacted Text

Using OpenAI's GPT3 To Detect Synthetic Text

Hi John,

Do you have a moment to fill out this form?

It's already been 3 months since your work contract began, and I'm wondering if you're satisfied with the work-life balance we provide. I'm sure you're just as committed as we are to keeping our employees happy and fulfilled.

This is a very short survey and it only takes a few minutes to complete.

You'll find that your responses will help us understand how we can improve your experience with **Company Y**.

Kind regards,

John Doe

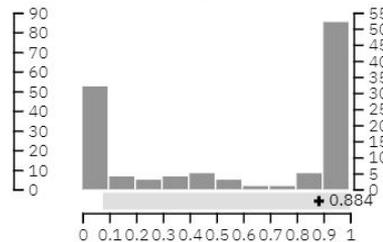
Sample B

frac(p): 0.884

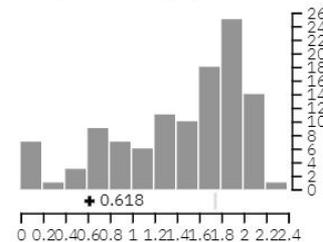
top k count



frac(p) histogram



top 10 entropy(p) histogram



Top K: Frac P: Colors (top k): 10 100 1000

Hi John,
Do you have a moment to fill out this form?
It's already been 3 months since your work contract began, and I'm wondering if you're satisfied with the work-life balance we provide. I'm sure you're just as committed as we are to keeping our employees happy and fulfilled.
This is a very short survey and it only takes a few minutes to complete.
You'll find that your responses will help us understand how we can improve your experience with **Company Y**.
Kind regards,
John Doe

Company Y Redacted Text

Using OpenAI's GPT3 To Detect Synthetic Text



Hi John,

Your profile has caught our attention. Your domain expertise in technology is someone whom my client is looking for. As such, I would like to share an upcoming opportunity with you.

However, before I can share any further, I will need you to sign up this Non-Disclosure Agreement (NDA) document attached in this email.

The upcoming opportunity which I will be sharing is of high business sensitivity and I would appreciate if you do not share with anyone on this matter. As this opportunity is time critical, we do hope to hear from you soon!

Sincerely,

John Doe

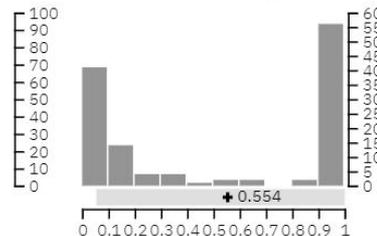
Sample C

frac(p): 0.554

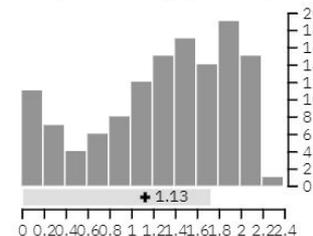
top k count



frac(p) histogram



top 10 entropy(p) histogram



Top K Frac P Colors (top k):

Hi John,
Your profile has caught our attention. Your domain expertise in technology is someone whom my client is looking for. As such, I would like to share an upcoming opportunity with you.
However, before I can share any further, I will need you to sign up this Non-Disclosure Agreement (NDA) document attached in this email.
The upcoming opportunity which I will be sharing is of high business sensitivity and I would appreciate if you do not share with anyone on this matter. As this opportunity is time critical, we do hope to hear from you soon!
Sincerely,
John Doe

Key applicable recommendations from Singapore's Model AI Governance Framework

Everyone

- Use Implementation and Self-Assessment Guide for Organizations
- Policy for explanation and practice general disclosure of use
- Ethical evaluation
- Implement clear roles and responsibilities for the ethical deployment of AI

Consumers

- Adopt “human in the loop” approach for AI-augmented decision-making

Suppliers

- Ensure traceability and auditability of use
- Enforce acceptable use policies

Personal Data Protection Commission, “Model AI Governance Framework,” 2020



Conclusion

Our Parting Words

1. The rapid growth of AlaaS has placed advanced, cost-effective AI text generation capabilities in the hands of the global market.

These capabilities can be used to deliver both authorised and malicious phishing campaigns.

2. While automated tools can be used to build defenses against AI-generated text, current approaches are brittle and model-dependent.

AI-assisted human detection of AI-generated text could be more effective.

3. Decision makers have the responsibility to implement sound strategies governing the supply and consumption of advanced AlaaS.

Tightening the usage of advanced AlaaS can potentially reduce the likelihood of abuse.



CYBER SECURITY
GROUP

THANK YOU

For any enquiries, please contact:

www.tech.gov.sg

[@GovTechSG](https://www.facebook.com/GovTechSG)

[Facebook.com/GovTechSG](https://www.facebook.com/GovTechSG)