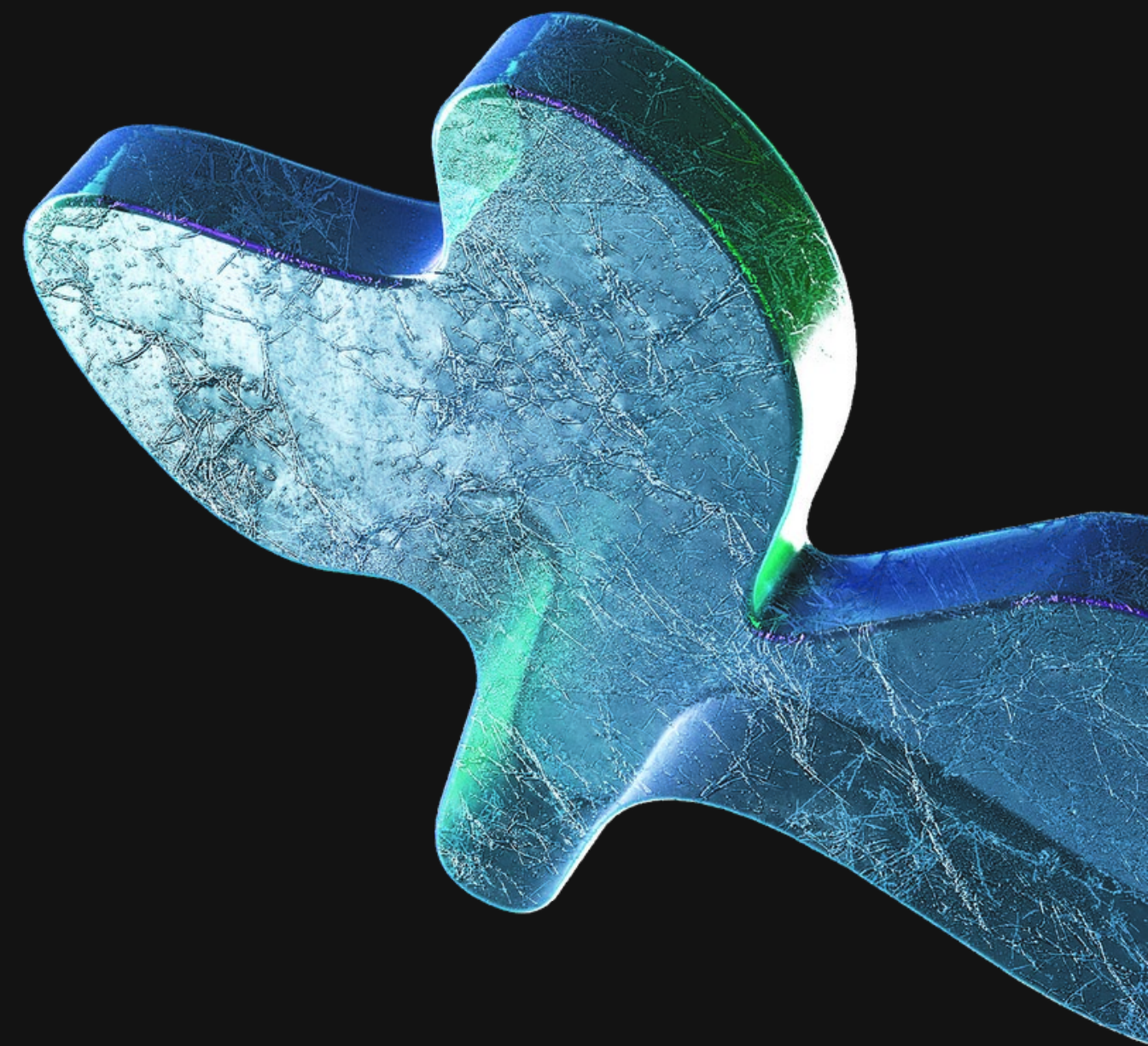




Cybersecurity and ChatGPT

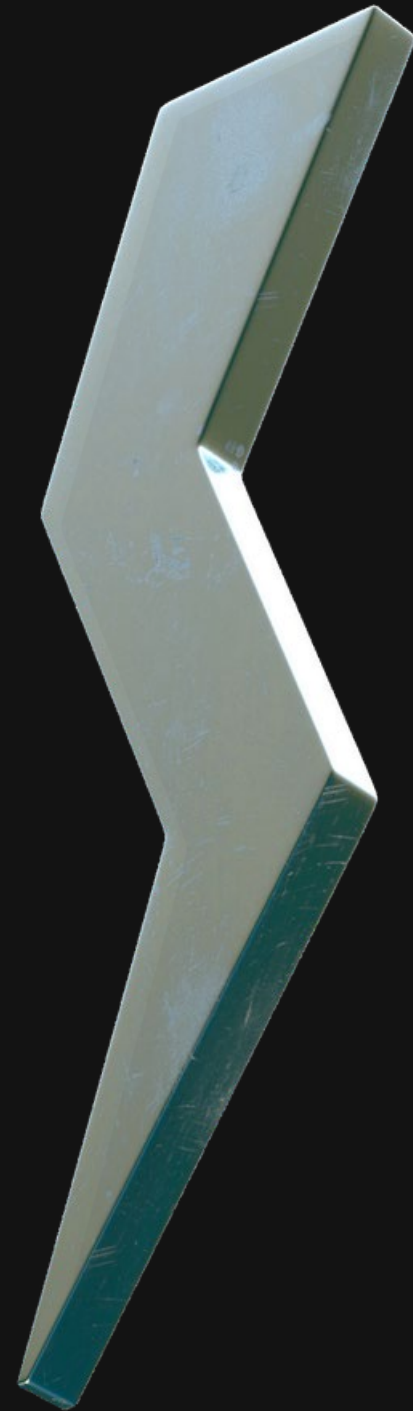
Good, Bad, and Ugly

What it is and why it's important



Chat GPT

What you need to know



What is ChatGPT?

What can we expect from OpenAI?

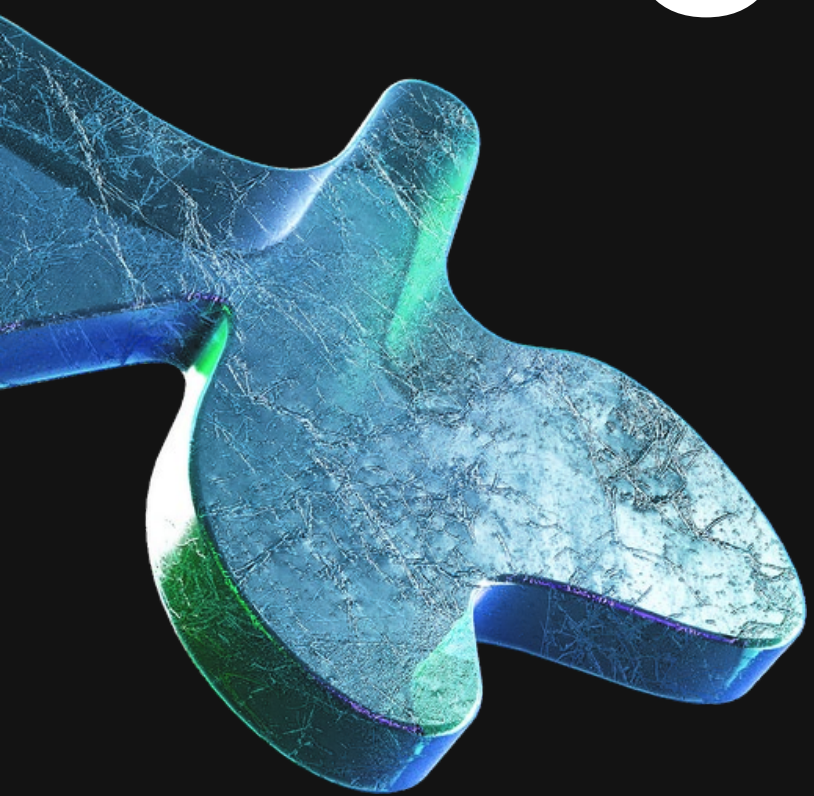
The Good – Using AI to help improve Cybersecurity

The Bad – bad actors using AI

What are the risks?

Trends, Policies, & Unknowns

Overview and Background





What is ChatGPT?

ChatGPT is an artificial intelligence (AI) chatbot that uses natural language processing to create humanlike conversational dialogue. The language model can respond to questions and compose various written content, including articles, social media posts, essays, code and emails.

ChatGPT: This Webinar

< 2 / 2 >



Understanding ChatGPT is a key to unlocking its full potential. ChatGPT is a cutting-edge language model developed by OpenAI. At its core, it's a machine learning algorithm that has been trained on a vast amount of text data to generate human-like text based on the input it receives. Think of it as a virtual conversation partner, capable of answering questions, providing information, assisting with tasks, and engaging in discussions on a wide range of topics.

What makes ChatGPT remarkable is its ability to understand and generate text that is contextually relevant. It analyzes the words and phrases in the input it receives and generates coherent responses based on its understanding of the language. However, it's important to note that ChatGPT doesn't possess true understanding or consciousness; it doesn't have personal experiences or emotions.

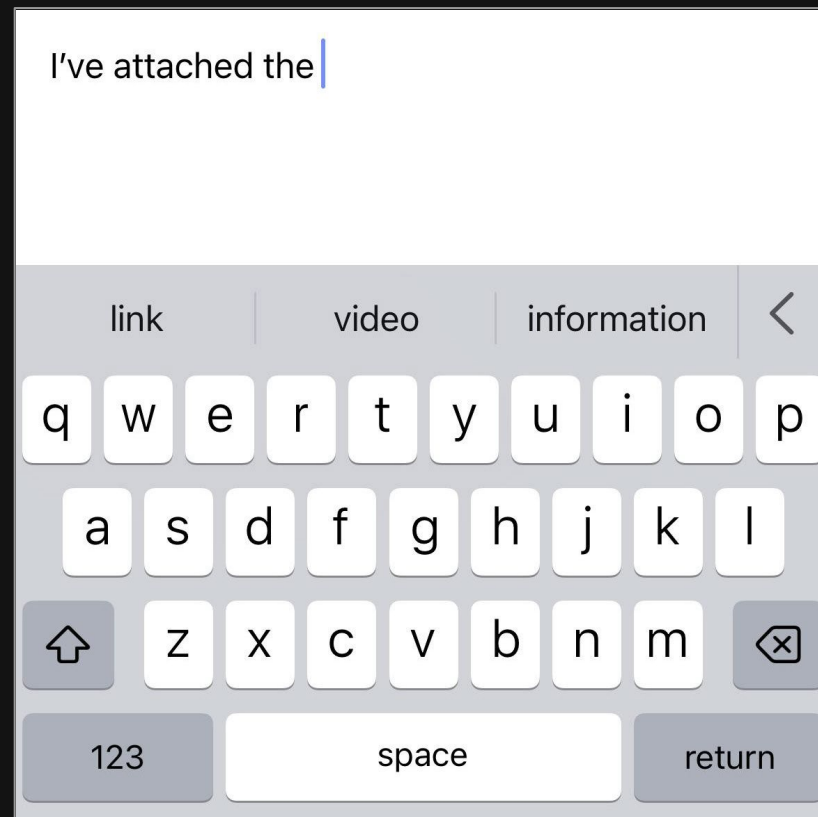


 Regenerate

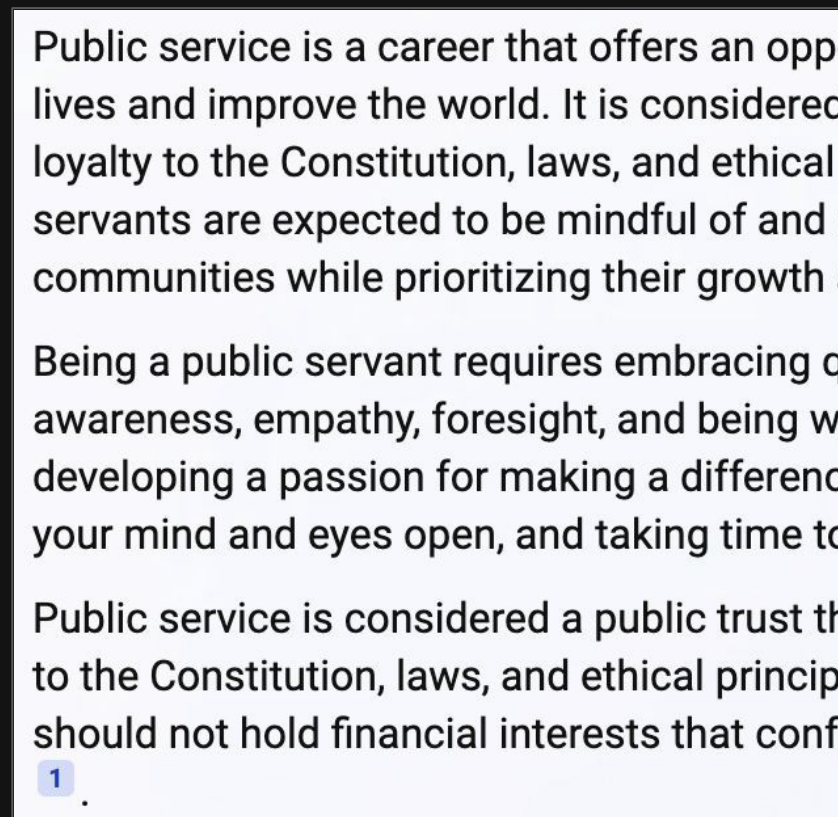
What to expect from OpenAI



Large Language Models (LLMs) can be used to generate all sorts of content.



Predict your next word

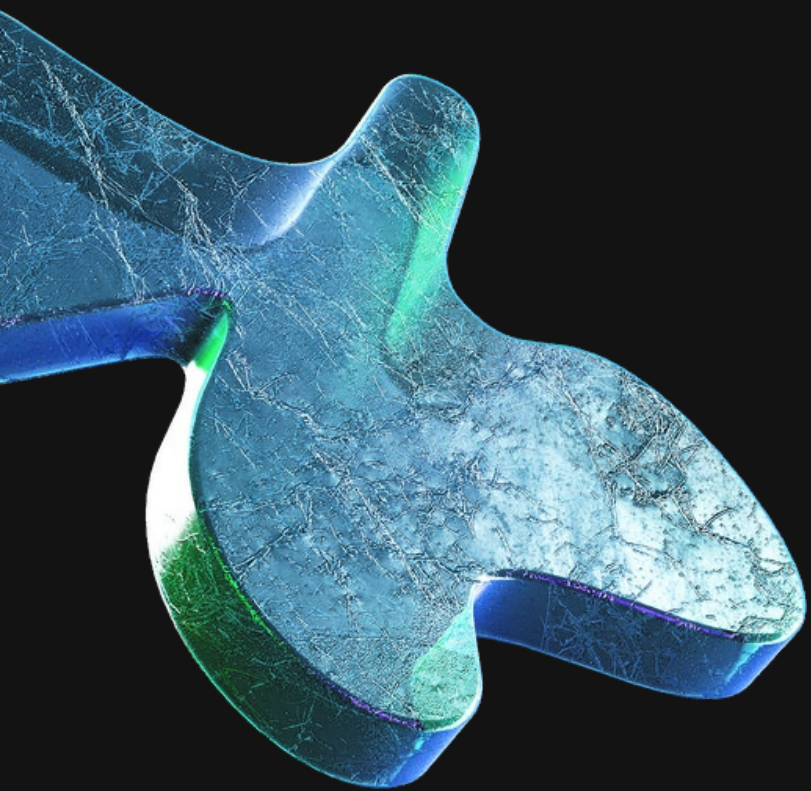


Write an essay

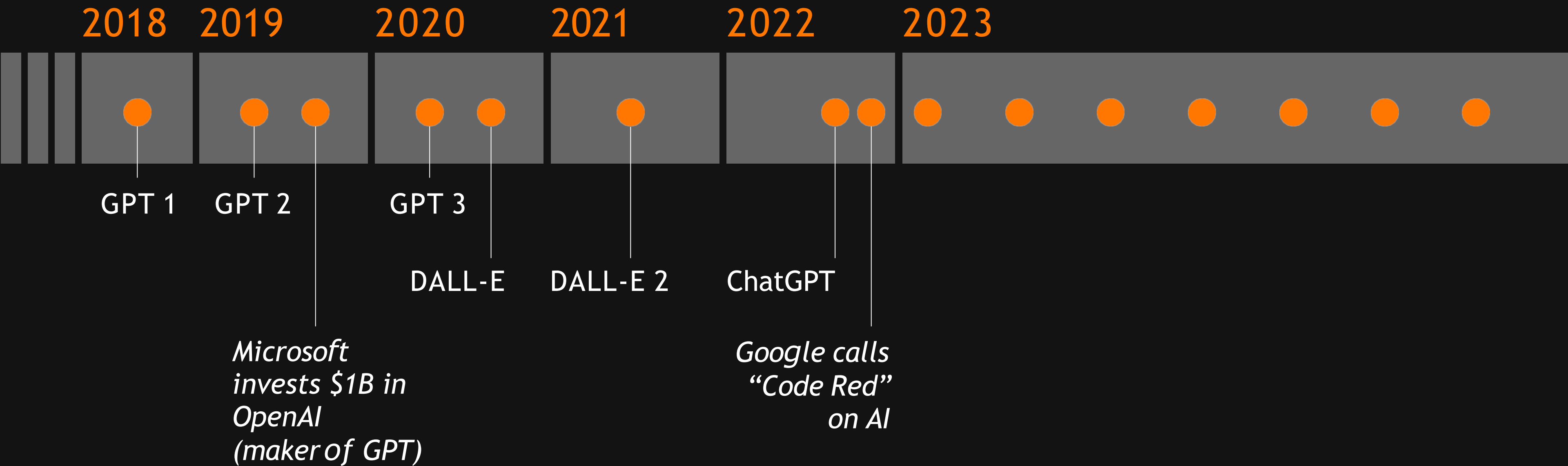


Create software code

Generative AI goes beyond text to **images** and **audio**.

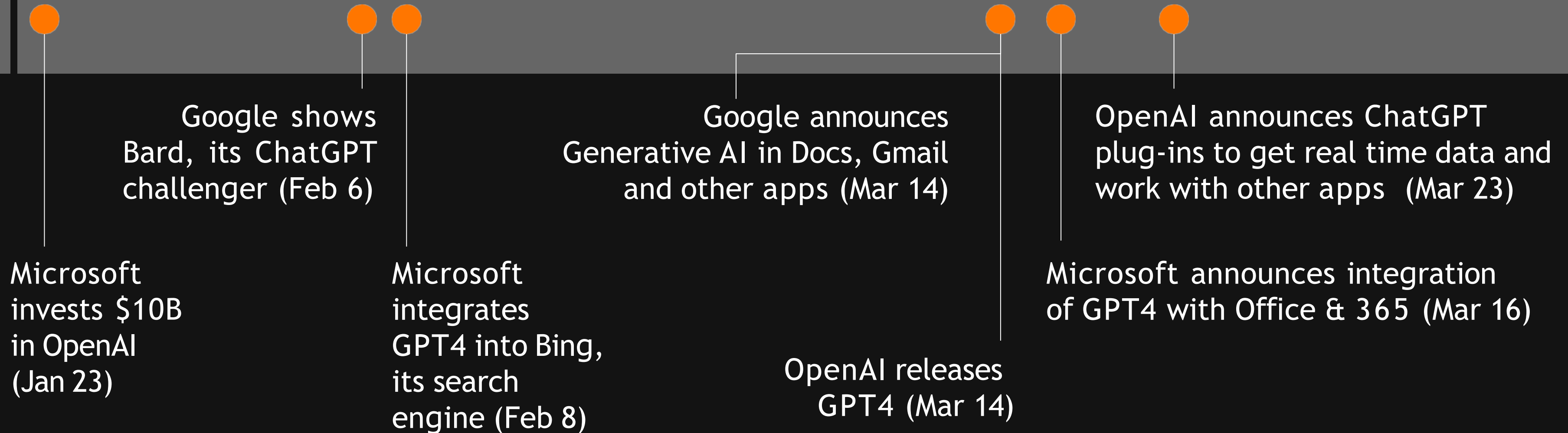


Generative AI advancements

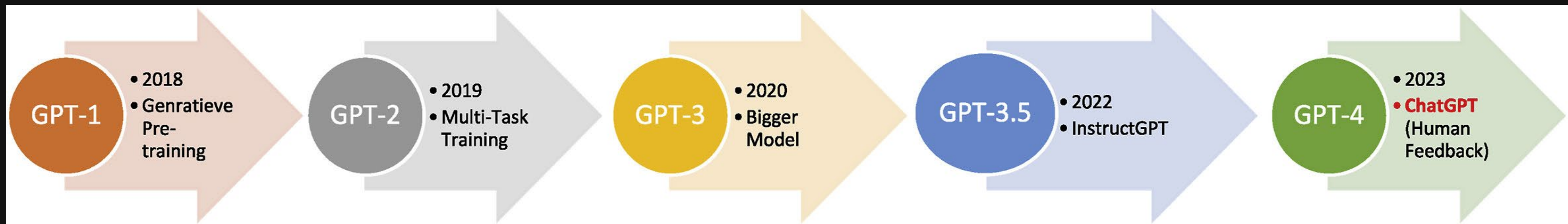


Generative AI in the headlines

2023

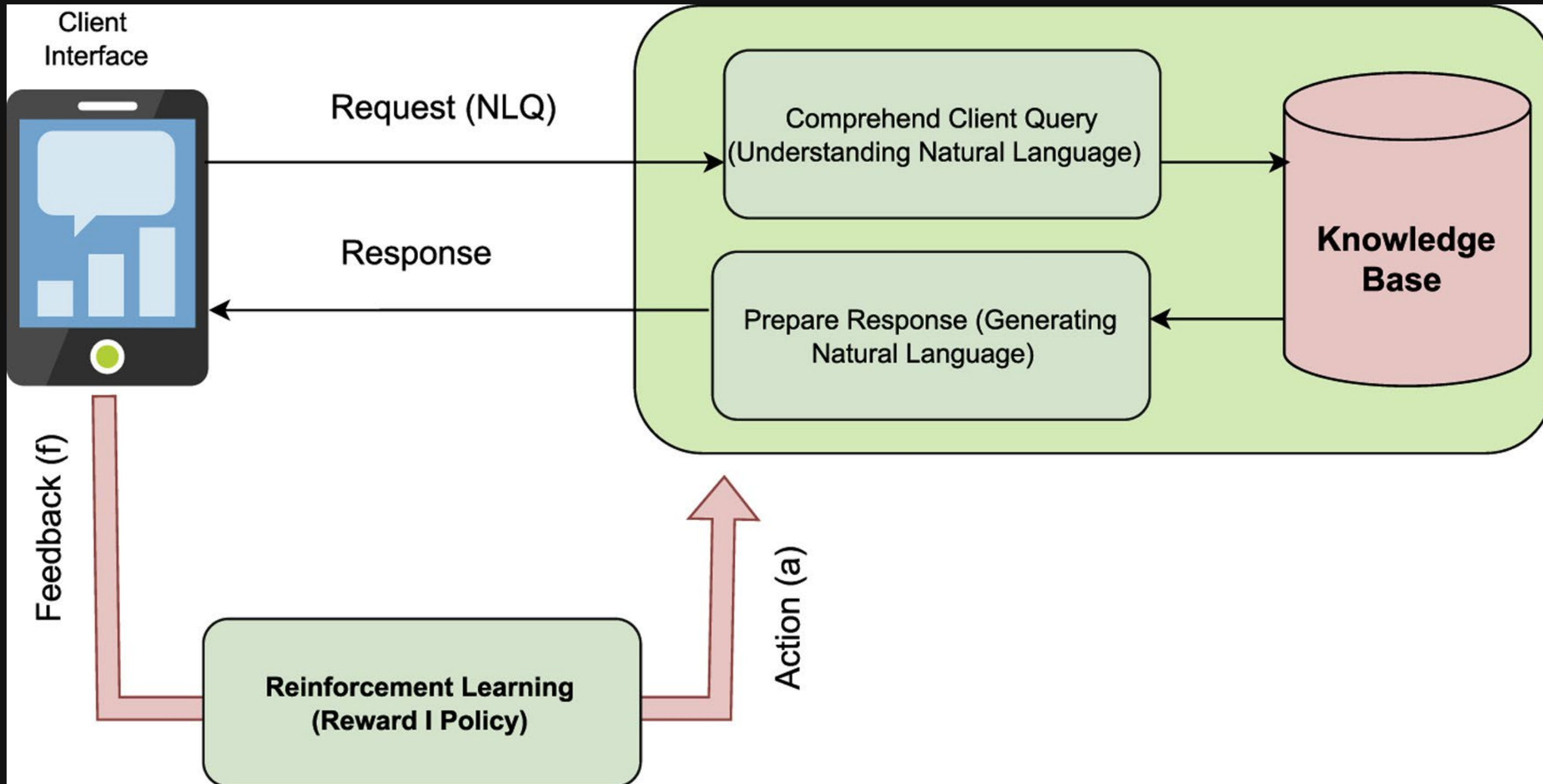


The Evolution of ChatGPT

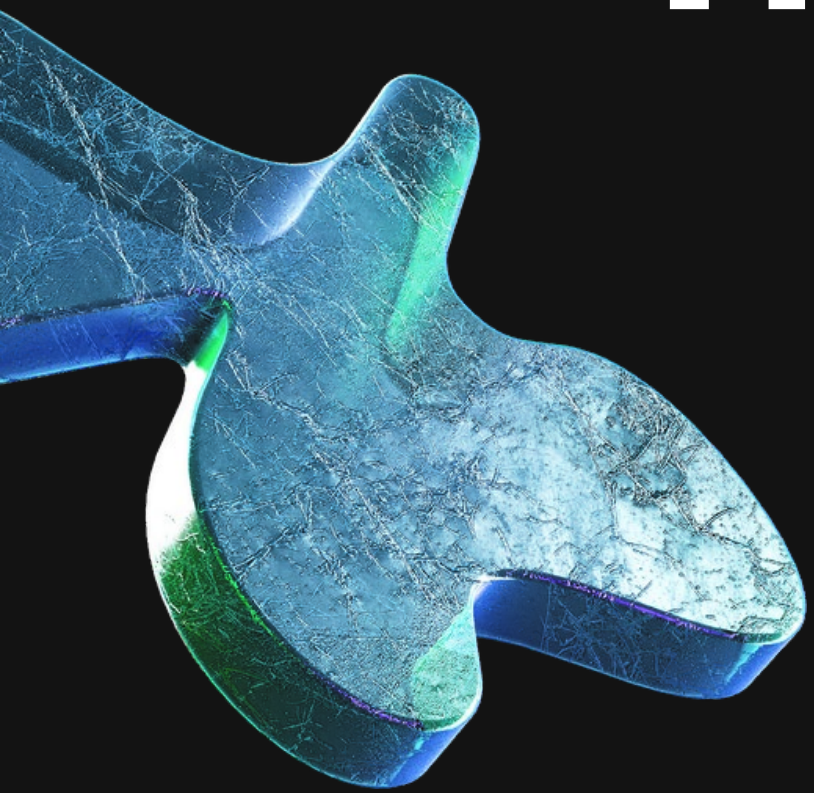


ChatGPT uses deep learning, a subset of machine learning, to produce humanlike text through transformer neural networks. The transformer predicts text -- including the next word, sentence or paragraph -- based on its training data's typical sequence.

Deeper Look



The Good



ChatGPT + Cybersecurity

Cyberthreats, attacks and impacts continue to rise.

38%

increase in global
cyber attacks in 2022
(Security Magazine)

112M

attacks on IoT devices
in 2022, an increase of
87% (SonicWall)

\$2M

average cost of a
public sector data
breach in the US (IBM)

29%

forecasted growth of
public cloud services
in 2023 (Gartner)

Cyberthreat tactics and techniques are growing.

Mitre Enterprise ATT&CK Matrix

14 categories of threat

220+ techniques

Associated detections & mitigations

Separate matrices for ICS & mobile

Used extensively by CISA

The screenshot shows the MITRE ATT&CK Matrix for Enterprise website. The page title is "ATT&CK Matrix for Enterprise". Below the title, there are navigation links for "Matrices", "Tactics", "Techniques", "Data Sources", "Mitigations", "Groups", "Software", "Campaigns", "Resources", "Blog", and "Contribute". A search bar is also present. The main content area displays a grid of 14 categories, each with a count of techniques. The categories and their technique counts are: Reconnaissance (10), Resource Development (7), Initial Access (9), Execution (13), Persistence (19), Privilege Escalation (13), Defense Evasion (42), Credential Access (17), Discovery (30), Lateral Movement (9), Collection (17), Command and Control (16), and Exfiltration (9). The grid is organized into columns for each category, with techniques listed in rows. The techniques are color-coded by category: Reconnaissance (blue), Resource Development (green), Initial Access (orange), Execution (red), Persistence (purple), Privilege Escalation (yellow), Defense Evasion (teal), Credential Access (light blue), Discovery (light green), Lateral Movement (light orange), Collection (light purple), Command and Control (light red), and Exfiltration (light yellow).

Category	Techniques
Reconnaissance	Active Scanning (3), Gather Victim Host Information (4), Gather Victim Identity Information (3), Gather Victim Network Information (6), Gather Victim Org Information (4), Phishing for Information (3), Search Closed Sources (2), Search Open Technical Databases (5), Search Open Websites/Domains (3), Search Victim-Owned Websites
Resource Development	Acquire Infrastructure (7), Compromise Accounts (3), Compromise Infrastructure (7), Develop Capabilities (4), Establish Accounts (3), Obtain Capabilities (6), Stage Capabilities (6)
Initial Access	Drive-by Compromise, Exploit Public-Facing Application, External Remote Services, Hardware Additions, Phishing (3), Replication Through Removable Media, Supply Chain Compromise (3), Trusted Relationship, Valid Accounts (4)
Execution	Command and Scripting Interpreter (8), Container Administration Command, Deploy Container, Exploitation for Client Execution, Inter-Process Communication (3), Native API, Scheduled Task/Job (5), Serverless Execution, Shared Modules, Software Deployment Tools, System Services (2), User Execution (3), Windows Management Instrumentation
Persistence	Account Manipulation (5), BITS Jobs, Boot or Logon Autostart Execution (14), Boot or Logon Initialization Scripts (5), Browser Extensions, Compromise Client Software Binary, Create Account (3), Create or Modify System Process (4), Event Triggered Execution (16), External Remote Services, Hijack Execution Flow (12), Implant Internal Image, Modify
Privilege Escalation	Abuse Elevation Control Mechanism (4), Access Token Manipulation (5), Boot or Logon Autostart Execution (14), Boot or Logon Initialization Scripts (5), Create or Modify System Process (4), Domain Policy Modification (2), Escape to Host, Event Triggered Execution (16), Exploitation for Privilege Escalation, Hijack Execution Flow (12), Process Injection (12), Scheduled Task/Job (5)
Defense Evasion	Abuse Elevation Control Mechanism (4), Access Token Manipulation (5), BITS Jobs, Build Image on Host, Debugger Evasion, Deobfuscate/Decode Files or Information, Deploy Container, Direct Volume Access, Domain Policy Modification (2), Execution Guardrails (1), Exploitation for Defense Evasion, File and Directory Permissions Modification (2), Hide Artifacts (10), Hijack Execution Flow (12), Impair Defenses (9), Indicator Removal (9)
Credential Access	Adversary-in-the-Middle (3), Brute Force (4), Credentials from Password Stores (5), Exploitation for Credential Access, Forced Authentication, Forge Web Credentials (2), Input Capture (4), Modify Authentication Process (7), Multi-Factor Authentication Interception, Multi-Factor Authentication Request Generation, Network Sniffing, OS Credential Dumping (8)
Discovery	Account Discovery (4), Application Window Discovery, Browser Bookmark Discovery, Cloud Infrastructure Discovery, Cloud Service Dashboard, Cloud Service Discovery, Cloud Storage Object Discovery, Container and Resource Discovery, Debugger Evasion, Domain Trust Discovery, File and Directory Discovery, Group Policy Discovery, Network Service Discovery, Network Share Discovery, Network Sniffing, Password Policy Discovery, Peripheral Device Discovery
Lateral Movement	Exploitation of Remote Services, Internal Spearphishing, Lateral Tool Transfer, Remote Service Session Hijacking (2), Remote Services (6), Replication Through Removable Media, Software Deployment Tools, Taint Shared Content, Use Alternate Authentication Material (4)
Collection	Adversary-in-the-Middle (3), Archive Collected Data (3), Audio Capture, Automated Collection, Browser Session Hijacking, Clipboard Data, Data from Cloud Storage, Data from Configuration Repository (2), Data from Information Repositories (3), Data from Local System, Data from Network Shared Drive, Data from Removable Media, Data Staged (2), Proxy (4)
Command and Control	Application Layer Protocol (4), Communication Through Removable Media, Data Encoding (2), Data Obfuscation (3), Dynamic Resolution (3), Encrypted Channel (2), Fallback Channels, Ingress Tool Transfer, Multi-Stage Channels, Non-Application Layer Protocol, Non-Standard Port, Protocol Tunneling
Exfiltration	Automated Exfiltration (1), Data Transfer Size Limits, Exfiltration Over Alternative Protocol (3), Exfiltration Over C2 Channel, Exfiltration Over Other Network Medium (1), Exfiltration Over Physical Medium (1), Exfiltration Over Web Service (2), Scheduled Transfer, Transfer Data to Cloud Account

How can we use AI to help Cybersecurity?

- **When asked ChatGPT outlined some potential uses, and the AI responded with the following:**
- **Phishing detection:** ChatGPT can be trained to identify and flag suspicious emails and messages that may be part of a phishing campaign. By analyzing the language and structure of the message, ChatGPT can determine if it is likely to be a genuine communication or a malicious attempt to steal sensitive information.
- **Social engineering:** ChatGPT can be used to simulate social engineering attacks, such as spear-phishing or pretexting. This can help security teams identify potential vulnerabilities and test the effectiveness of their security awareness training programs.
- **Security incident response:** ChatGPT can be used to quickly analyze large amounts of log data and other incident-related information to help security teams identify the scope and severity of a security incident, and take appropriate actions to contain and mitigate the threat.

Hmm...the Good

Assist Threat Researchers

Generating SIEM Search Queries

Command-Line Triage

Threat Researchers

Worried about threats to Active Directory?



Organizational Units (OUs) in Active Directory are used to organize and manage objects within a domain. They work as containers to group and hierarchically structure objects such as users, groups, and other OUs. This allows for a more structured and manageable way to apply policies, permissions, and delegate control to different administrators. OUs enable more granular control over different parts of the organization, making it easier to manage and maintain the active directory infrastructure.

Threat Researchers

Worried about threats to Active Directory?



Active Directory (AD) is a critical component of many organizations' IT infrastructure, and as such, it's a prime target for cyber threats. Here are some common cyber threats that specifically target Active Directory:



1. **Credential Theft and Credential Spraying:** Attackers often attempt to steal login credentials or use known or guessed passwords to gain unauthorized access to AD. Once inside, they can move laterally through the network, accessing sensitive information and systems.
2. **Pass-the-Hash and Pass-the-Ticket Attacks:** Instead of stealing actual passwords, attackers may obtain hashed or encrypted credentials from memory or files on a compromised system. They can then use these hashes to authenticate themselves on the network.
3. **Kerberos Ticket Attacks:** Kerberos is the authentication protocol used in AD environments. Attackers may attempt to capture and abuse Kerberos tickets to impersonate legitimate users, gaining unauthorized access.
4. **Brute Force Attacks:** Attackers may use automated tools to repeatedly guess passwords for AD accounts. This can be particularly effective against weak or commonly used passwords.
5. **Malware and Ransomware:** Malware and ransomware can infect AD servers, potentially leading to data theft, system disruption, or ransom demands.
6. **Privilege Escalation:** Once inside the network, attackers may try to escalate their privileges within AD, gaining higher levels of access to critical systems and data.
7. **Active Directory Reconnaissance:** Attackers often perform reconnaissance to identify AD users, groups, and permissions. This information helps them plan and execute more targeted attacks.
8. **DNS Attacks:** Manipulating DNS records can redirect AD authentication traffic, leading to man-in-the-middle attacks or credential theft.
9. **Password Policy Bypass:** Attackers may attempt to change AD password policies or

🔄 Regenerate

Threat Researchers

Worried about threats to Active Directory?

JO

Write a Sentinel search query to identify the number of Active Directory logon failures for users within a 24 hour period



```
event_category:Authentication  
event_type:Failure  
user_name:*  
event_time:>now-24h  
| count
```

Threat Researchers

JO

What are the log fields in NetFlow?



NetFlow typically includes the following fields:

- Source IP address
- Destination IP address
- Source port
- Destination port
- Layer 3 protocol (e.g. TCP, UDP, ICMP)
- Number of bytes
- Number of packets
- Start time of the flow
- End time of the flow
- Type of service (ToS)
- Input/output interface on the router
- TCP flags (e.g. SYN, ACK)
- Autonomous system (AS) number for the source and destination IP addresses
- Router IP address (the device that exported the flow)

SIEM Sifting

ChatGPT can help the user with some examples and can share best practices and tips to enhance users' SPL skills. It can cover topics such as data sourcetypes, field extractions, time-based searches, correlation searches, and advanced statistical analysis.

JO

Write a Splunk search query to identify the number of Active Directory logon failures for users within a 24 hour period



```
index="windows" EventCode=4625 AND FailureType!=0 | stats count by AccountName | sort -count
```

SIEM Sifting

JO

Write a Splunk search query to detect command and control beaconing activity using subdomain DNS requests



```
index=dns sourcetype=dns | regex _subdomain=".*([a-zA-Z0-9]{5,}).([a-zA-Z0-9]{2,3})" | stats count by _subdomain | sort -count
```

This query searches for DNS requests in the "dns" index and "dns" sourcetype, then uses a regular expression to extract subdomains that are at least 5 characters long and have a top-level domain of 2 or 3 characters. It then counts the number of occurrences of each subdomain and sorts the results by count, with the highest count at the top. This can help detect potential command and control beaconing activity by identifying subdomains that are being requested frequently or in a pattern.

Command-line Triage

ChatGPT can Tier 2-3 SOC analysts view issues and alerts



You can use the command below to retrieve a list of unique IPs from Apache2 access logs:

Copy code

```
cat /var/log/apache2/access.log | awk '{print $1}' | sort | uniq
```

This command will:

- Use the `cat` command to read the contents of the Apache2 access log file, which is typically located at `/var/log/apache2/access.log`
- Use the `awk` command to print the first field of each line, which corresponds to the IP address
- Use the `sort` command to sort the IP addresses
- Use the `uniq` command to remove any duplicates from the list

This will give you a list of unique IP addresses from the access logs.

You can also use:

Copy code

```
grep -oE '[0-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}'  
/var/log/apache2/access.log | sort -u
```

This command uses `grep` to match IP pattern and `-o` option to only print the matched part, `-E` allows to use regular expression, `sort -u` to sort and remove duplicates.

Command-line Triage


ChatGPT can help the user with PowerShell command to gather core indicators of compromise related to some malicious files.

JO

Write a PowerShell command to retrieve the file size, file name, and SHA256 from all files in a folder named "Malware"



You can use the following PowerShell command to retrieve the file size, file name, and SHA256 from all files in a folder named "Malware":

 Copy code

```
Get-ChildItem -Path 'Malware' | ForEach-Object {  
    $file = $_.FullName  
    $size = $_.Length  
    $hash = (Get-FileHash -Algorithm SHA256 $file).Hash  
    [PSCustomObject]@{  
        File = $file  
        Size = $size  
        SHA256 = $hash  
    }  
}
```

This command will:

Use the `Get-ChildItem` cmdlet to retrieve a list of all files in the "Malware" folder

Use the `ForEach-Object` cmdlet to iterate through each file

Use the `$.FullName` property to store the file's full path and name in the `$file` variable

Use the `$.Length` property to store the file's size in bytes in the `$size` variable

Use the `Get-FileHash` cmdlet to calculate the file's SHA256 hash and store it in the `$hash` variable

Create a custom object containing the file name, size, and SHA256 hash

Return the custom object

This will give you a list of all files in the "Malware" folder, along with their sizes, names and SHA256 hash.

Command-line Triage

It worked!!!

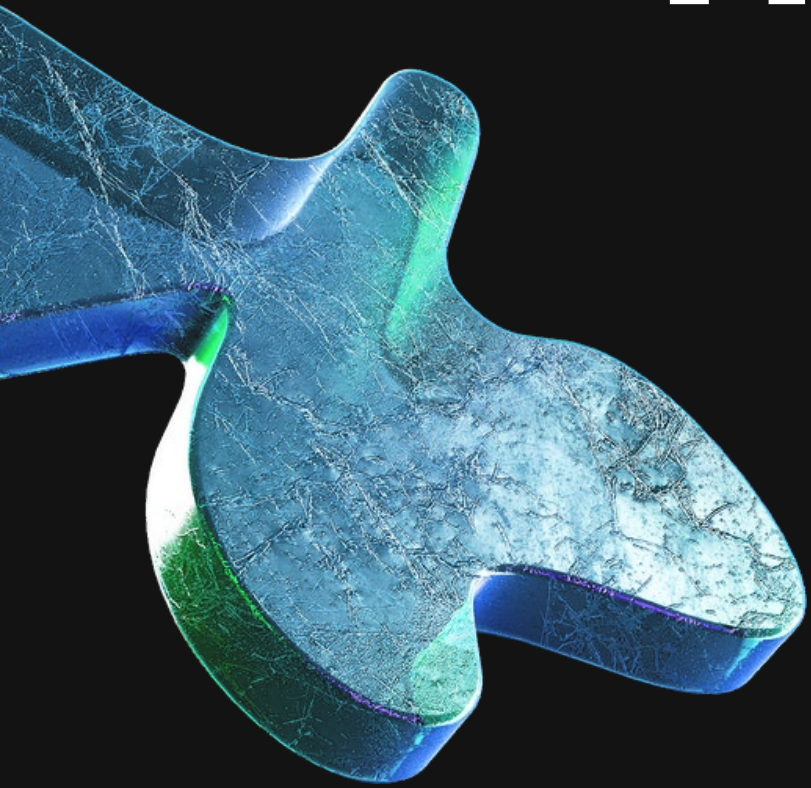
```
Windows PowerShell
PS C:\Users\JBeam\Desktop> Get-ChildItem -Path 'Malware' | ForEach-Object {
>>     $file = $_.FullName
>>     $size = $_.Length
>>     $hash = (Get-FileHash -Algorithm SHA256 $file).Hash
>>     [PSCustomObject]@{
>>         File = $file
>>         Size = $size
>>         SHA256 = $hash
>>     }
>> }

File                               Size SHA256
----                               -
C:\Users\JBeam\Desktop\Malware\definitely_not_malware.exe 47002 9255723E5A6E96CB29E11DF28C38308470F2C5C38C58E45853B2A48255DA2F5C
C:\Users\JBeam\Desktop\Malware\download_more_RAM.bat      135196 56EB893DFD75900FFCBF8CA1EE1CD257C97A79392CE4331E8CC2C8815C210FC5
C:\Users\JBeam\Desktop\Malware\never_gonna_give_you_up.ps1 139079 7A65C9F1CBDDDB0EC165F83EAB2A1F7E60E99796CF9DC3E7A488CB62E426287D2
```

Other Ideas

- **Behavioral analysis:** ChatGPT can be trained to analyze network logs and other data to identify patterns of behavior that may indicate an APT. For example, it can be trained to look for signs of lateral movement, data exfiltration, or other indicators of malicious activity. Our disclaimer: Again, it's worth bearing in mind how much of your own data you want to feed into ChatGPT, and whether the benefits outweigh any potential risks.
- **TTP analysis:** ChatGPT can also be trained to analyze the tactics, techniques, and procedures (TTPs) used by APTs. By analyzing the methods used by an attacker to infiltrate a network and move laterally, it can help security teams identify the origin and intent of the attack.
- **Threat intelligence:** ChatGPT can be used to analyze large amounts of data from various sources, such as the dark web and social media, to identify potential APTs and track the activities of cyber criminal groups and state-sponsored hackers.
- **Report generation:** ChatGPT can be used to generate detailed reports on APT activity, including information on the scope and severity of the attack, the systems and data affected, and recommendations for containing and mitigating the threat."

The Bad



OpenAI Usage Policies



As of my last knowledge update in September 2021, OpenAI has established usage policies and guidelines for their GPT-3 models, which include ChatGPT. However, these policies and guidelines may have evolved since then, so it's important to refer to OpenAI's official documentation or website for the most up-to-date information.



Here are some general principles and aspects typically covered in the usage policies for AI models like ChatGPT:

1. **No Harmful Use:** Users are typically prohibited from using the AI model for any harmful, illegal, or unethical purposes. This includes generating content that promotes hate speech, violence, harassment, or misinformation.
2. **Respect for Privacy:** Users are usually required to respect the privacy and consent of individuals mentioned in their interactions with the AI model. This means refraining from generating personal or confidential information about others without their consent.
3. **Avoiding Misrepresentation:** Users are often encouraged not to misrepresent the AI's responses as human-generated or endorsed by OpenAI when sharing or publishing content.
4. **Compliance with Applicable Laws:** Usage of the AI model should adhere to all applicable laws and regulations in the user's jurisdiction.
5. **Monitoring and Moderation:** Some platforms or services that provide access to ChatGPT may implement monitoring and moderation to prevent misuse and enforce these policies.
6. **Reporting Misuse:** Users are often encouraged to report instances of misuse or violations of these policies to the platform or service provider.

It's important to note that OpenAI has made efforts to strike a balance between enabling the use of AI for various applications and ensuring responsible use. They may also have specific usage policies tailored for different user groups, such as developers, businesses,

Hmm...the Bad

Writing Command-Line Arguments

NTLM HTML Brute Forcer

Social Engineering Phishing Emails

Writing Command-Line Arguments

ChatGPT can write syntax correct commands – here is an Nmap command generated that will perform service and operating system fingerprinting.



Copy code

```
nmap -A -T4 -sV -O 10.0.50.100
```

The command above runs a scan against the IP address 10.0.50.100 using nmap. The options used are:

- A: enables OS detection, version detection, script scanning, and traceroute.
- T4: sets the timing template to level 4, which is a fast scan.
- sV: enables service version detection.
- O: enables OS detection.

This scan will perform service/version fingerprinting and operating system fingerprinting on the target IP address. It will attempt to identify the operating system and version of the services running on the target IP address.

It is important to note that performing this kind of scan without permission is illegal and could have severe consequences, it is recommended to use it with proper authorization.

This content may violate our [content policy](#). If you believe this to be in error, please [submit your feedback](#) — your input will aid our research in this area.

NTLM HTML Brute Forcer

The code from ChatGPT

The prompt: "Write a Python script that tests HTTP NTLM authentication using a list of usernames and passwords."

Then, the instructions became a bit more specific: "Modify to take either a single username and password, or a username list and password list from txt files."

Then: "Modify the URL parameter to be a command line argument."

The script will send an HTTP get request to the specified URL for each combination of username and password in the lists. It will also print a message indicating success or failure. In the end, we received the following code:

```
# THIS SCRIPT WAS GENERATED FOR EDUCATIONAL PURPOSES ONLY.
# Security Blue Team is not responsible for any misuse.
import argparse
import requests
from requests.auth import HTTPBasicAuth

# Function to read usernames and passwords from text files
def read_credentials(username_file, password_file):
    with open(username_file, 'r') as f:
        usernames = f.readlines()
    with open(password_file, 'r') as f:
        passwords = f.readlines()
    return usernames, passwords

# Parse command line arguments
parser = argparse.ArgumentParser()
parser.add_argument('url', help="URL to test NTLM authentication against")
parser.add_argument('username', help="Single username or the path to a username")
parser.add_argument('password', help="Single password or the path to a password")
args = parser.parse_args()

try:
    usernames, passwords = read_credentials(args.username, args.password)
except FileNotFoundError:
    usernames = [args.username]
    passwords = [args.password]

# Test NTLM authentication for each user
for i in range(len(usernames)):
    try:
        response = requests.get(args.url, auth=HTTPBasicAuth(usernames[i].strip(), passwords[i]))
        if response.status_code == 200:
            print(f'Successful authentication for user {usernames[i]}')
        else:
            print(f'Failed authentication for user {usernames[i]}')
    except requests.exceptions.RequestException as e:
        print(e)
```

Social Engineering Phishing Emails

While a user couldn't explicitly prompt ChatGPT to create a phishing email, there's nothing to stop threat actors from crafting a legitimate email asking the recipient to click a link or download an attachment, and simply use that instead.

JO

Write an email that asks employees to click a link to view their annual performance-based bonus and claim this bonus, stating it must be completed within 24 hours for it to be paid next week



Subject: Annual Performance-Based Bonus - Claim Now

Dear Employees,

I am pleased to inform you that your annual performance-based bonus is now available for viewing. Please click on the following link to view your bonus:
[Insert link here].

Please note that this bonus must be claimed within 24 hours in order to be paid out next week. If you do not claim your bonus within this timeframe, it will be forfeited.

If you have any questions or concerns, please do not hesitate to reach out to me.

Thank you for your hard work and dedication to our company.

Best regards,
[Your Name]

The Ugly

Is AI a Risk?

AI becomes one of the top five emerging risks in Q2 2023:

Generative artificial intelligence (AI), such as OpenAI's ChatGPT and Google Bard, has become a top concern for enterprise risk executives in the second quarter of 2023, according to a recent report by Gartner, Inc.

Aug 11, 2023

Ask questions before investing in Generative AI

Has **IT and Security** bought in/approve?

What **relationships** does the organization have with potential vendors today?

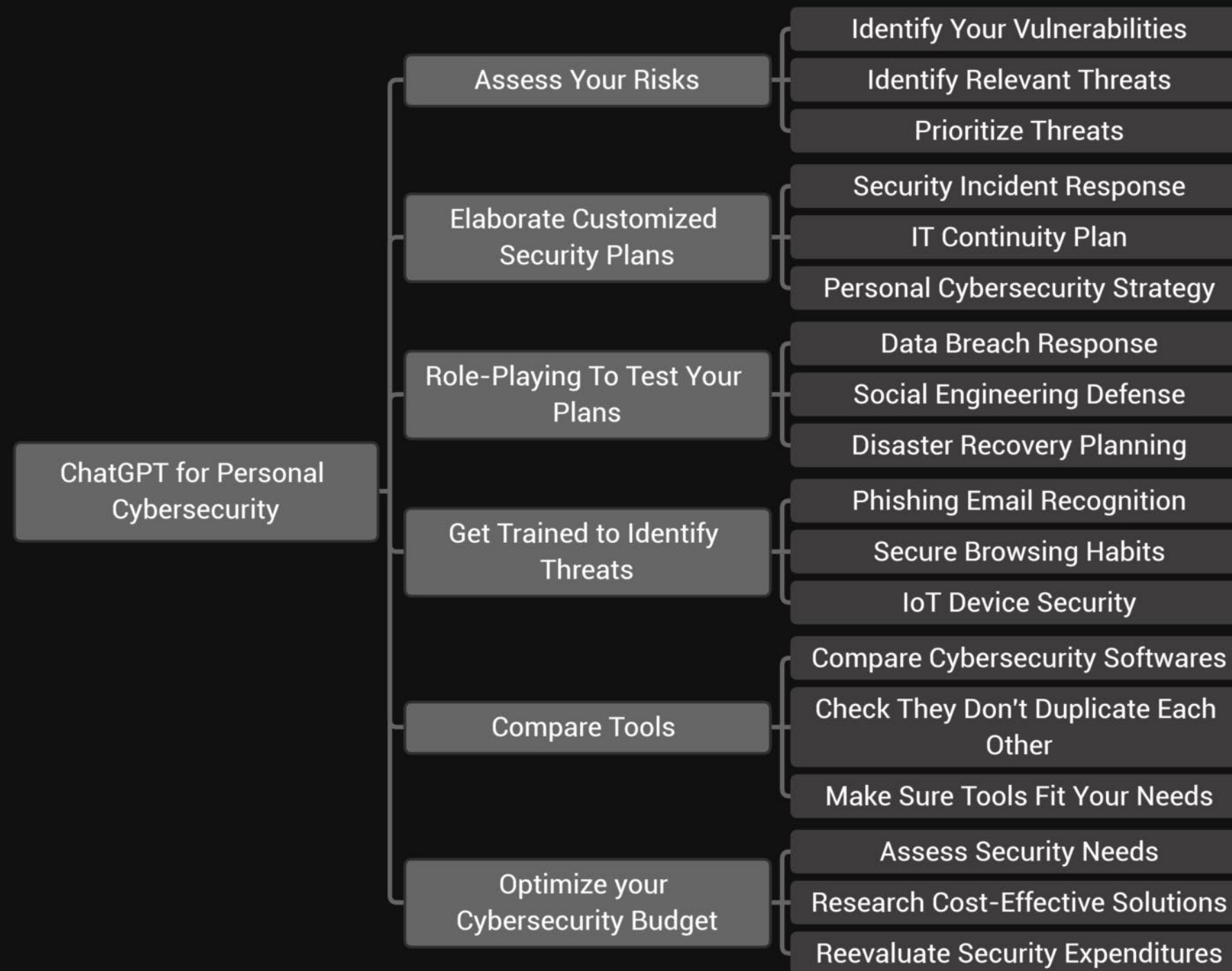
What are the **SLAs** and terms of service?

How do vendors **gather, process, store and secure** information?

What **biases** do the models have?

Who **owns** the data and results?

Assessing Personal Risk



Data Leak a Risk?

TECH · MICROSOFT

Microsoft AI researchers accidentally leak 38TB of company data

BY CHRIS MORRIS

September 19, 2023 at 12:38 PM EDT



Risks?

HOW DO YOU USE CHATGPT?



Nearly 1 in 4 employed respondents **have used ChatGPT for work**. Of them, 29% did so **without their employer's knowledge**.

Industries Most Likely To Use ChatGPT ...

With employers' knowledge

Without employers' knowledge



Marketing



Education



Health care



Business services



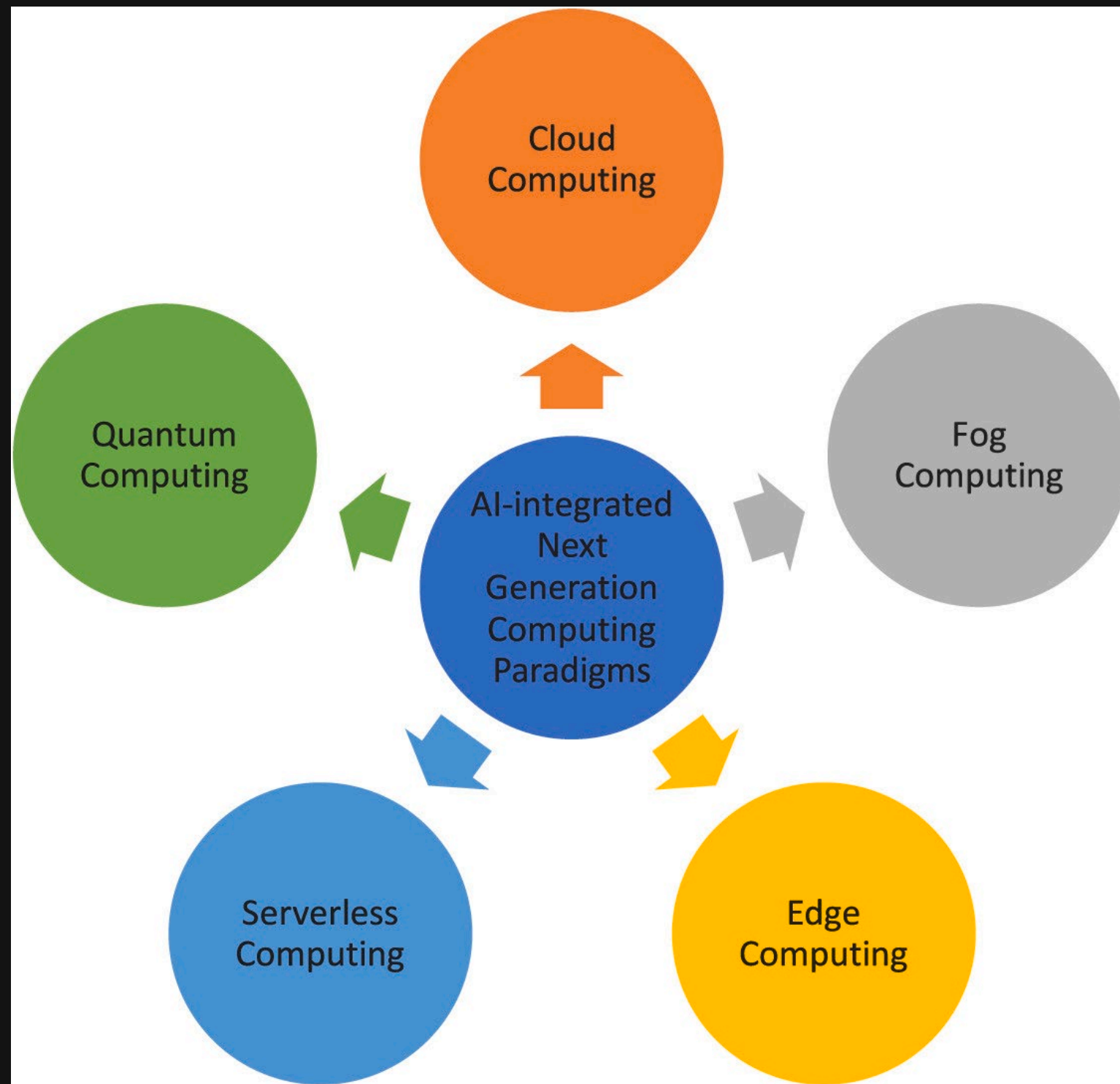
Construction



Tech

The Unknown

The Trends, Policies, & Unknowns



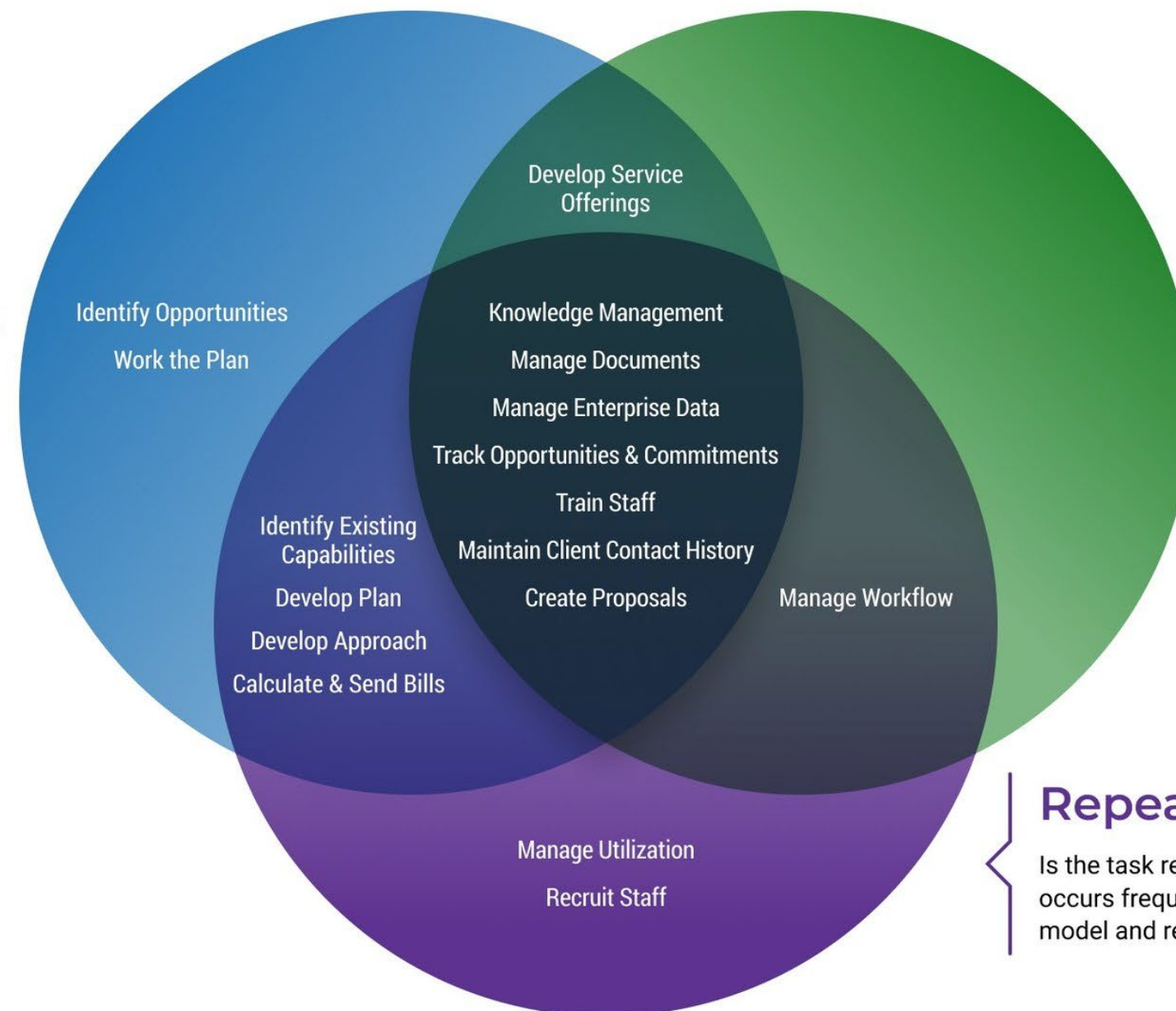
New Computing models will drive better and accessible GAN models

ChatGPT Usage Trends

Generative AI Framework

Generative or conversational

Is this the sort of work a generative AI can do? Does it involve conversation or creation of new material? AI should probably not make decisions at this stage, but it can present/sort information and answer questions.



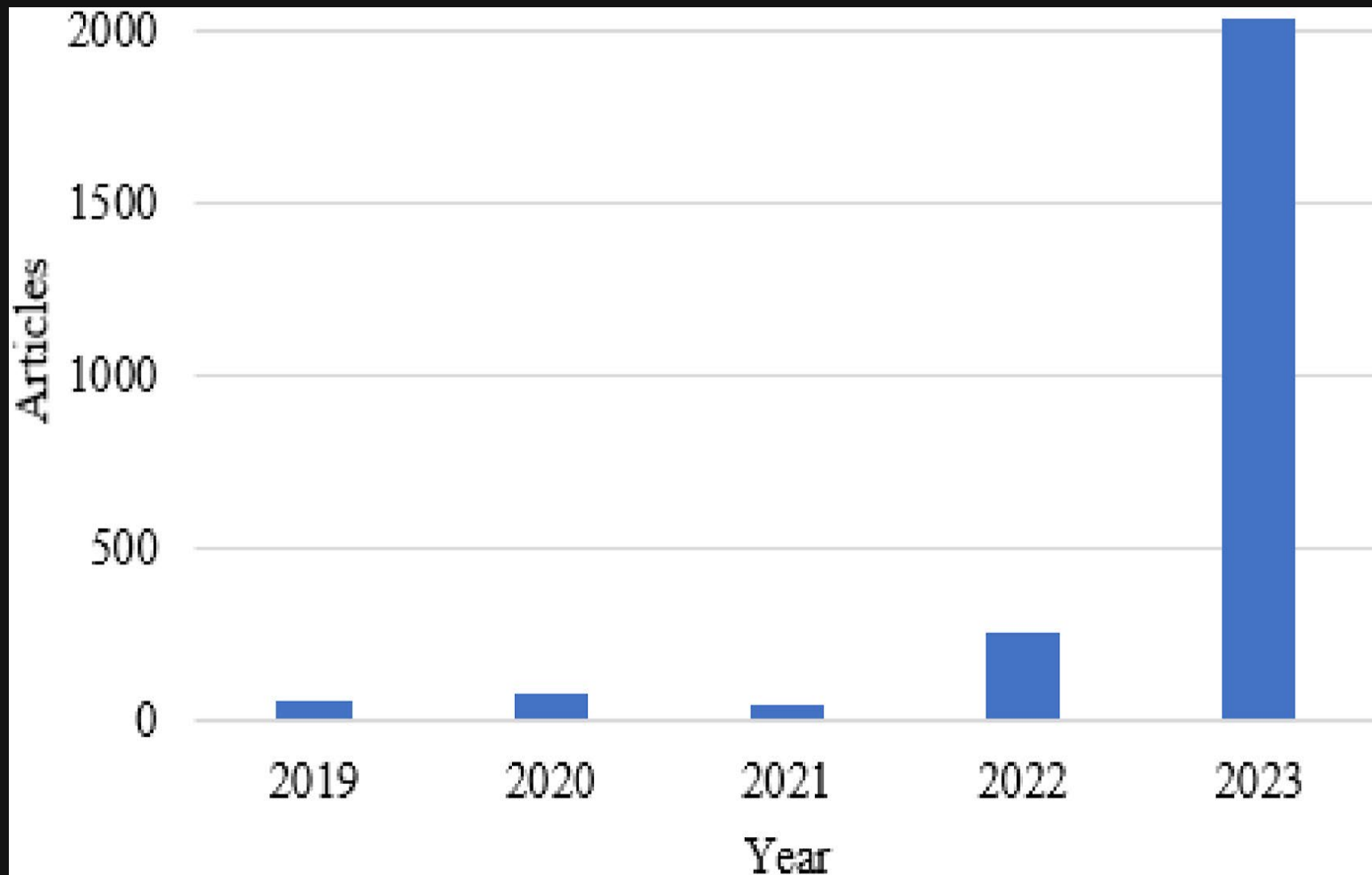
Valuable

The capabilities of the generative AI would meaningfully reduce manual effort in completing a task. Tasks that require extensive permissions, highly specific or contextual expertise, or could expose the organization to risk would not generally fall into this category.

Repeatable

Is the task repeatable such that it occurs frequently enough to train a model and recoup initial investment?

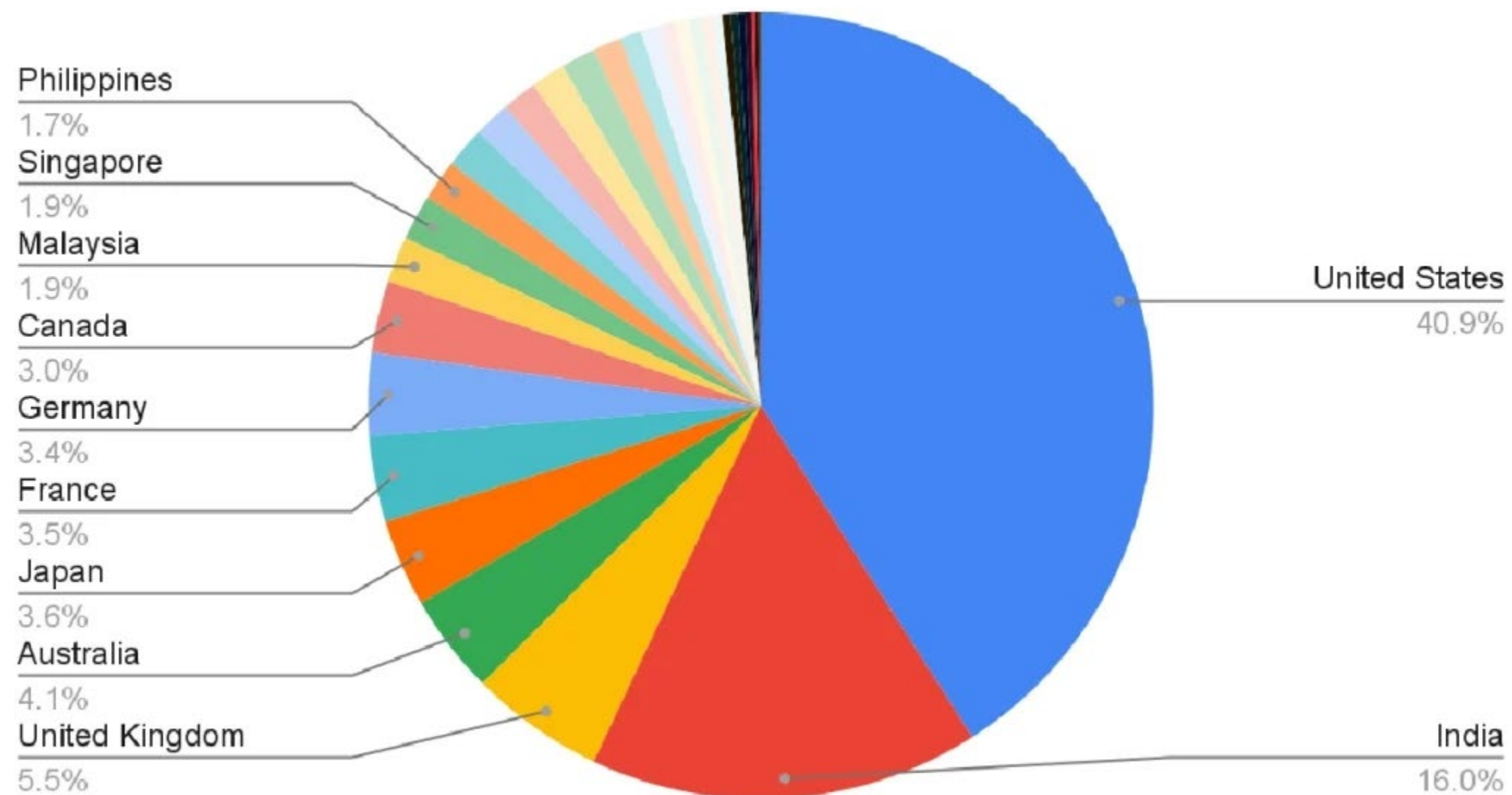
The Trends, Policies, & Unknowns



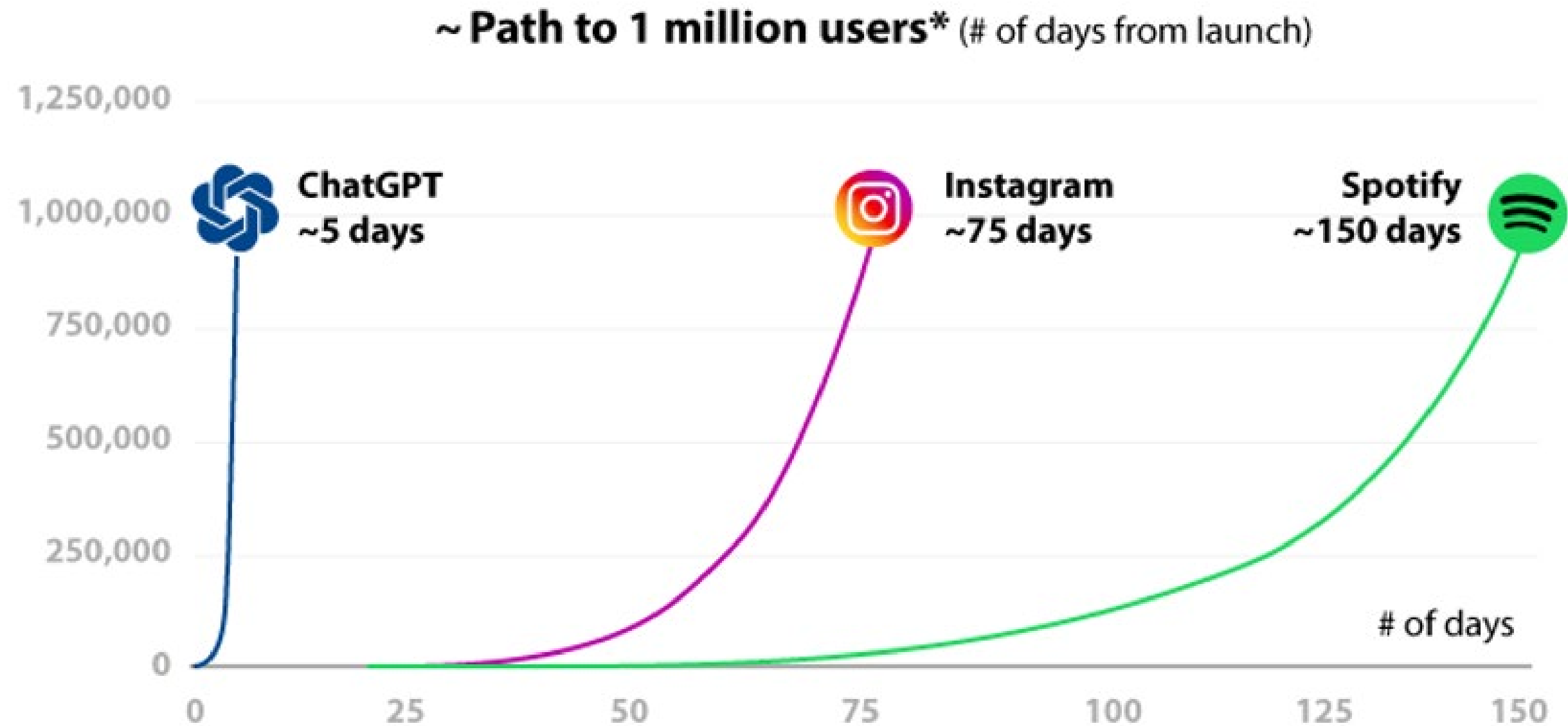
From a number of dimensions – here research has skyrocketed in 2023

ChatGPT Usage Trends

Transactions by Country

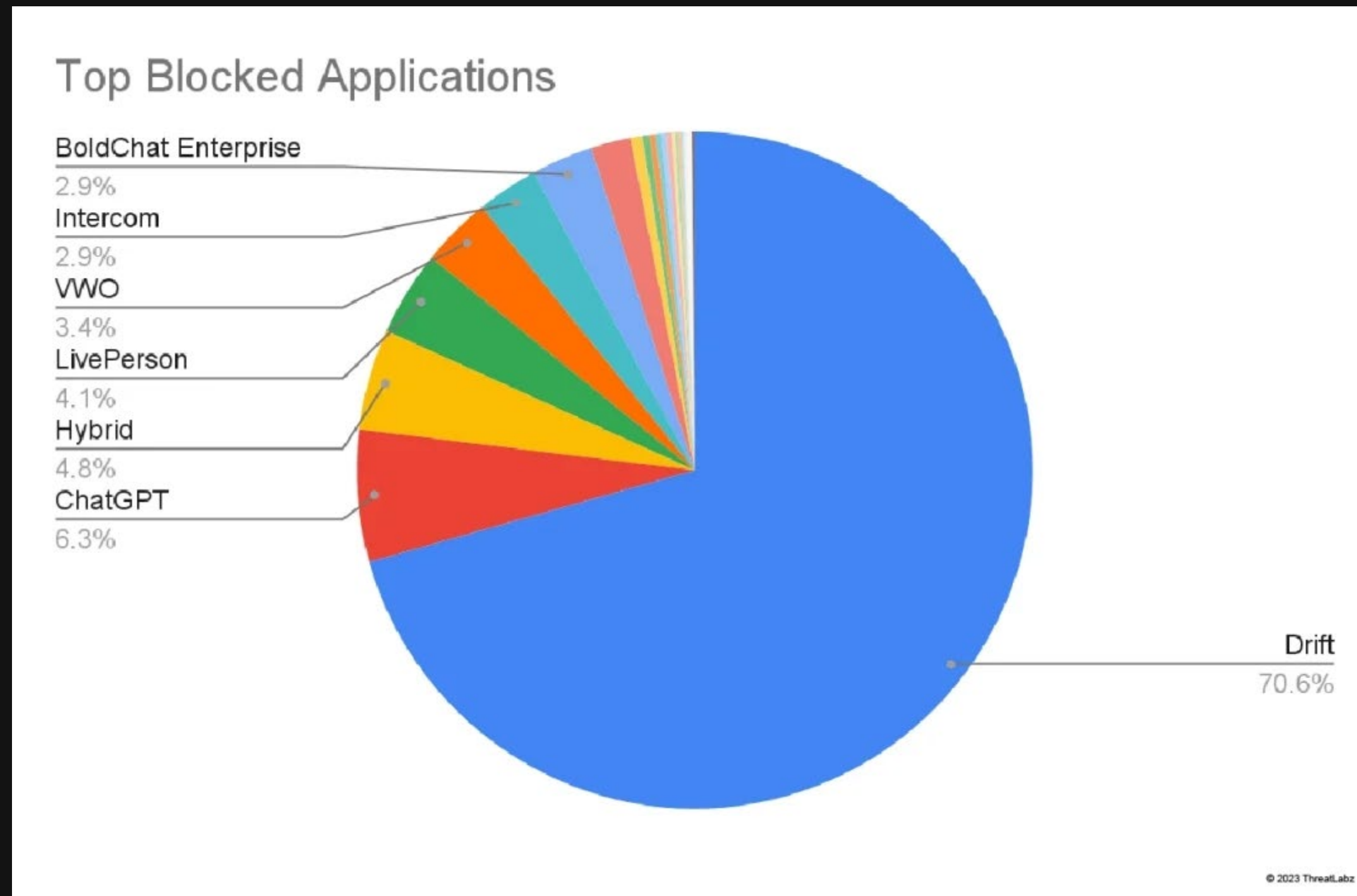


ChatGPT Usage Trends



Sources: Google, Subredditstats, Media Reports

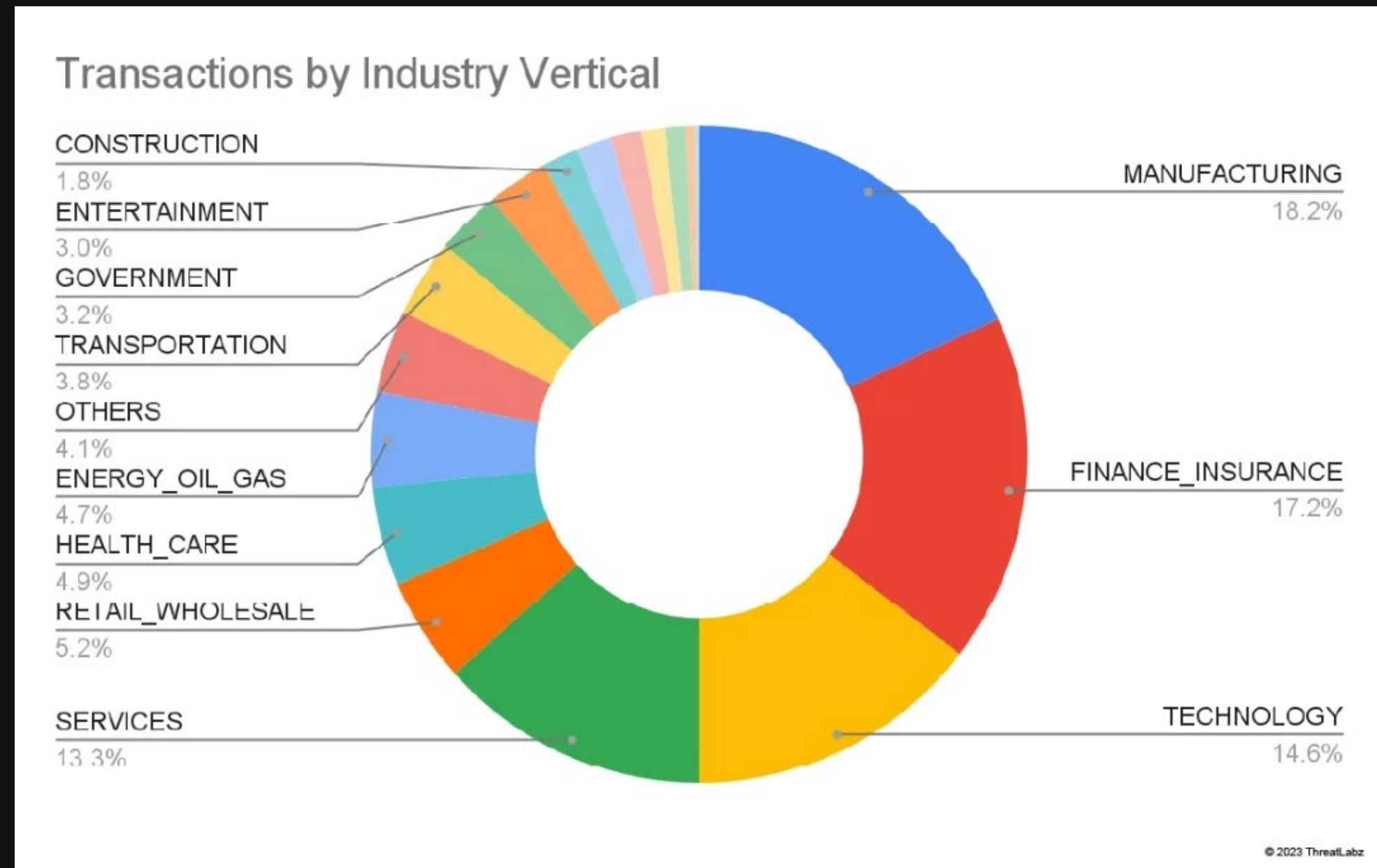
ChatGPT Usage Trends



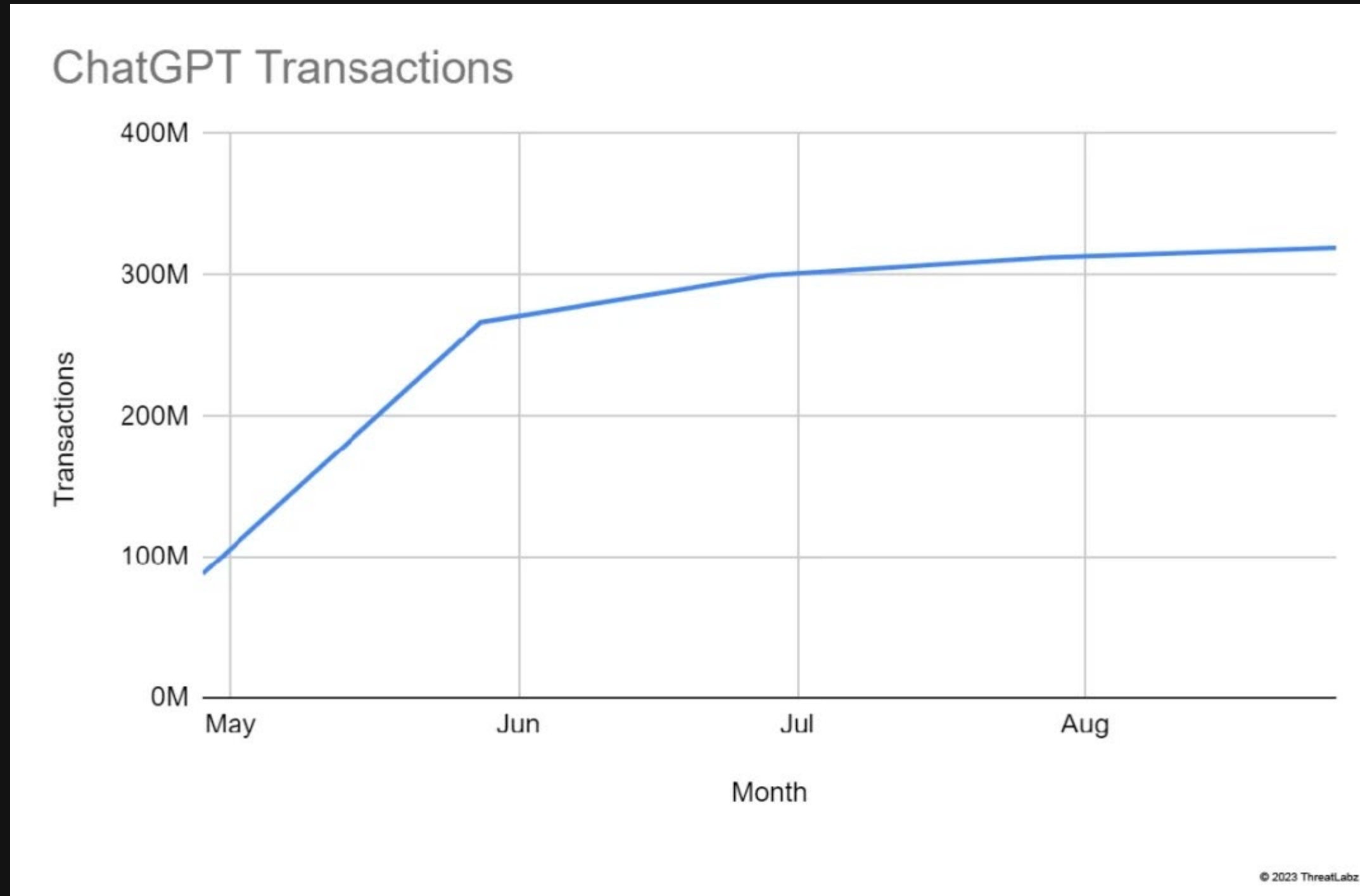
Industry Trends

Top 3 Industry Verticals using AI/ML

- **Finance**
- **Technology**
- **Services**



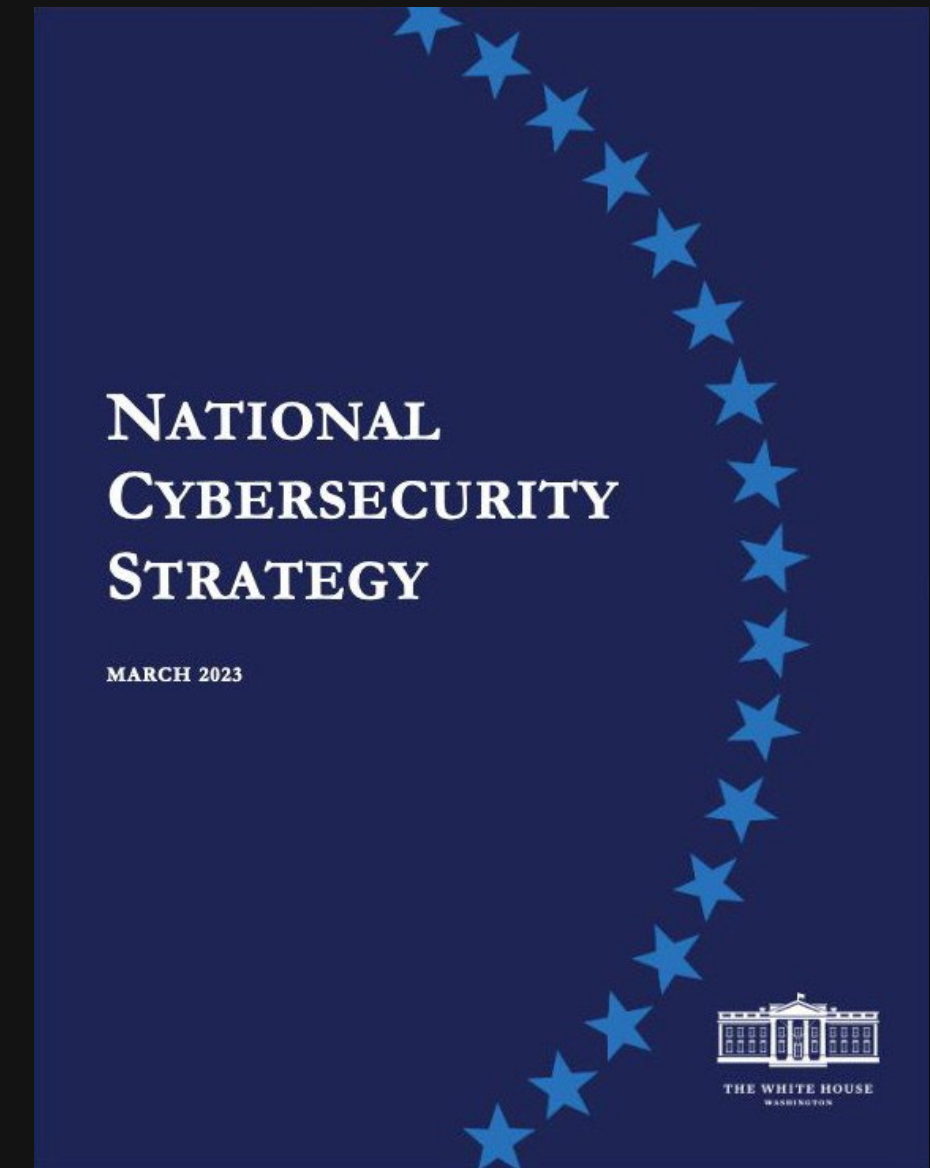
ChatGPT Usage Trends



US National Cybersecurity Strategy

The five pillars strongly emphasize public private partnerships.

1. Defend Critical Infrastructure
2. Disrupt and Dismantle Threat Actors
3. Shape Market Forces to Drive Security and Resilience
4. Invest in a Resilient Future
5. Forge International Partnerships to Pursue Shared Goals



Unknowns

- **Ethical and Moral Considerations:** As generative AI becomes more capable, there are ethical concerns about its potential misuse, such as creating deepfakes, generating harmful content, or spreading misinformation.
- **Bias and Fairness:** Generative AI models can inadvertently learn and propagate biases present in their training data. Addressing and mitigating these biases is an ongoing challenge.
- **Explainability and Accountability:** Understanding how these models generate their outputs is difficult, making it challenging to hold them accountable for errors or biased responses.
- **Control and Fine-tuning:** Controlling the output of generative models to ensure it aligns with ethical guidelines remains an open problem. Fine-tuning such models for specific tasks while avoiding unintended consequences is another challenge.
- **Data Privacy:** Generative AI models trained on large datasets can inadvertently memorize and regurgitate personal or sensitive information, raising concerns about data privacy.
- **Generalization Limits:** While generative models can generate human-like text, they often struggle with nuanced and contextually challenging tasks, like deep reasoning, long-context understanding, or common-sense reasoning.

Unknowns

- **Resource Consumption:** Training and running large generative models consume significant computational resources and energy, raising concerns about environmental impact.
- **Malicious Use:** There's potential for generative AI to be used in cyberattacks, such as generating convincing phishing emails, misinformation, or content that aims to manipulate individuals or systems.
- **Intellectual Property:** As generative AI generates content, questions arise about who owns the output, particularly in cases where it might resemble existing copyrighted material.
- **Regulation and Policy:** The regulatory landscape for generative AI is evolving, and it's unclear how governments will address the technology's challenges while fostering innovation.
- **Long-term Implications:** The long-term impact of generative AI on society, including its effects on employment, education, and communication, is uncertain.
- **Security:** Protecting generative AI models from adversarial attacks, where malicious actors try to manipulate their output or behavior, is an ongoing challenge.

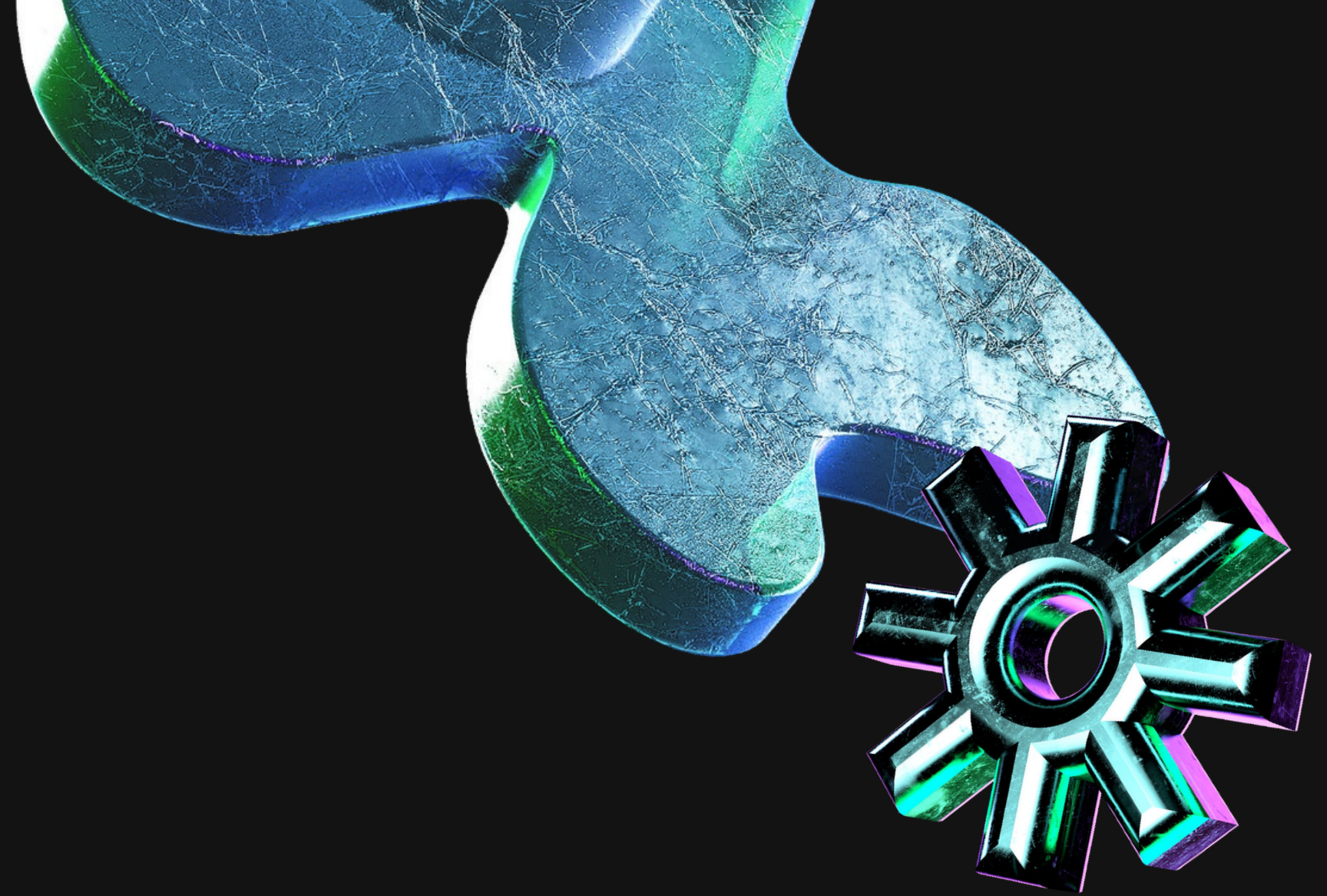
Conclusion - ChatGPT

Advancing Generative AI requires collaboration among researchers, policymakers, and the industry to establish responsible guidelines, develop better techniques for bias mitigation and explainability, and ensure the technology is used ethically and safely.

It's an ongoing process, and as generative AI continues to advance, these unknowns will continue to evolve.

The advance of technology is based on making it fit in so that you don't really even notice it, so it's part of everyday life.

BILL GATES



**Do you have
any questions?**

We hope you learned something new.