# Your AI Assistant has a Big Mouth

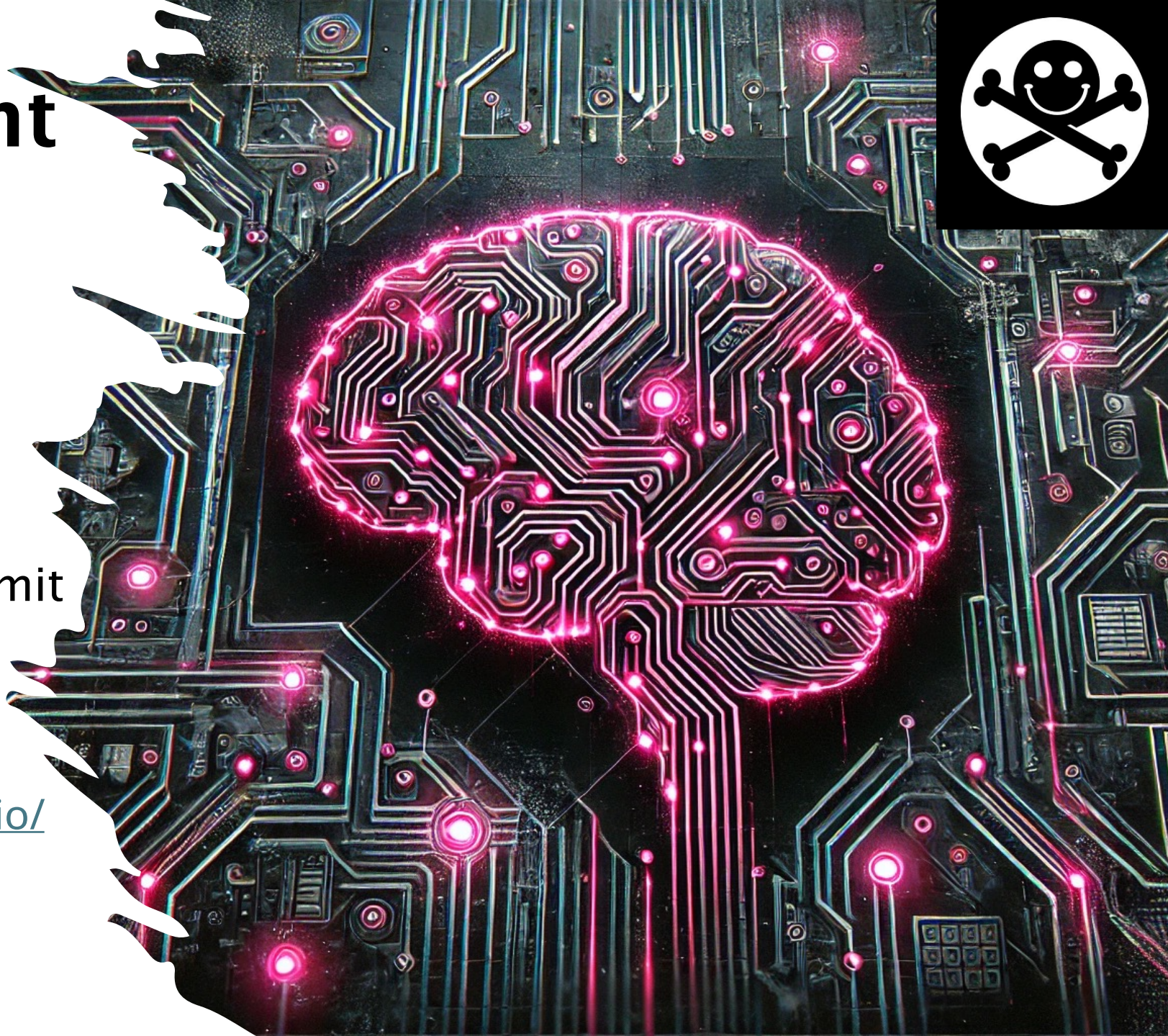
## A New Side-Channel Attack

Yisroel Mirsky, Roy Weiss,
Daniel Ayzenshteyn, Guy Amit

Offensive AI Research Lab
Ben-Gurion University
https://offensive-ai-lab.github.io/

CBG
Cyber@Ben-Gurion University of the Negev

INCD
Israel National Cyber Directorate











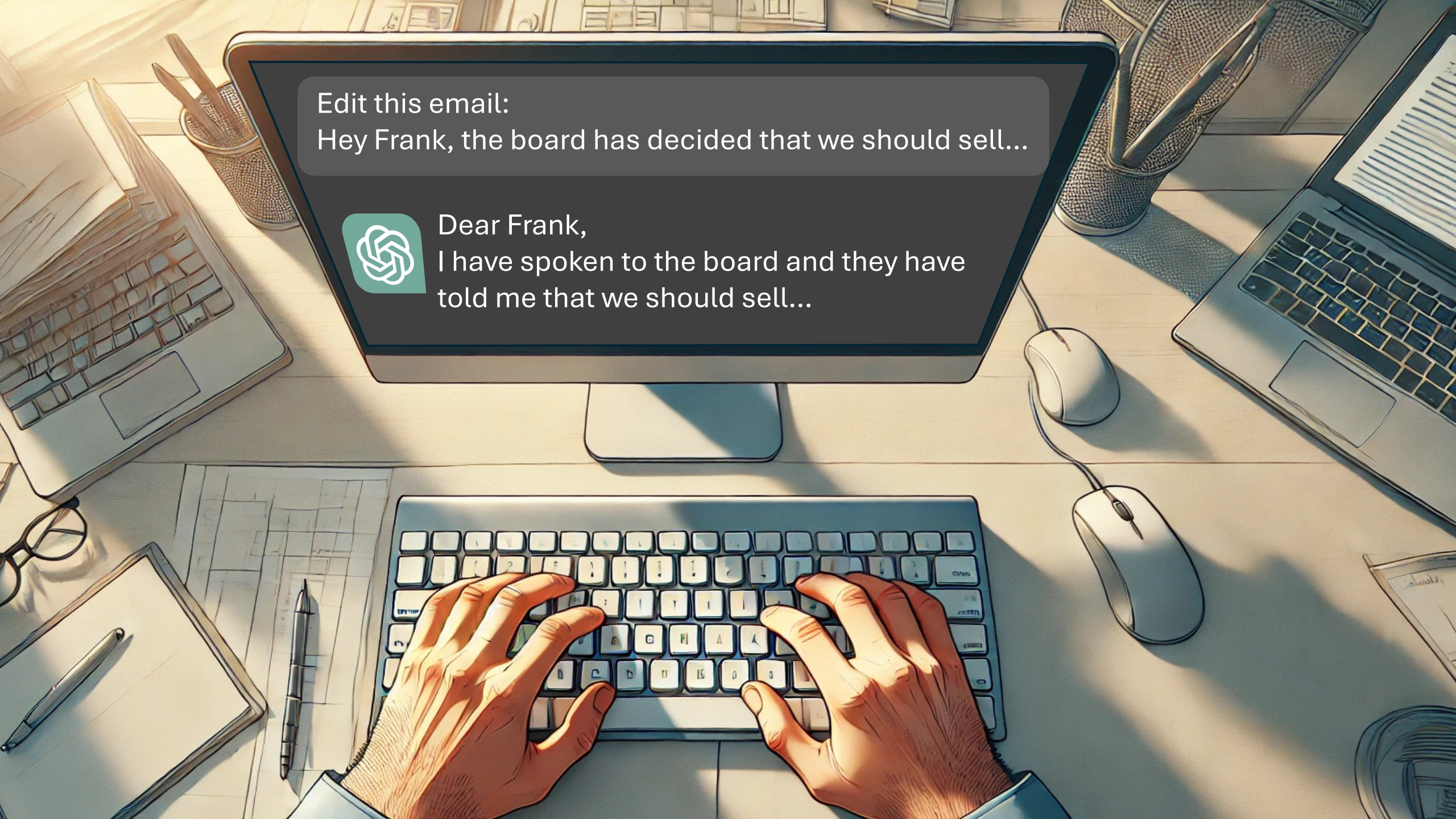
# The Age of AI Assistants




Why do I have a rash on my back?
I’m sorry to hear that you have a rash on your back...


What are signs my wife has an affair?
Your wife may be having an affair if you notice that she...


How can I hack my neighbors WIFI?
There are several ways you can hack...


Edit this email:
Hey Frank, the board has decided that we should sell...
Dear Frank,
I have spoken to the board and they have told me that we should sell...

# What if someone could read them?



Why do I have a rash on my back?

I'm sorry to hear you have a rash...





Eavesdropper

# But isn’t it Encrypted?



5D7C93BA97606ABBC1AF0AE37

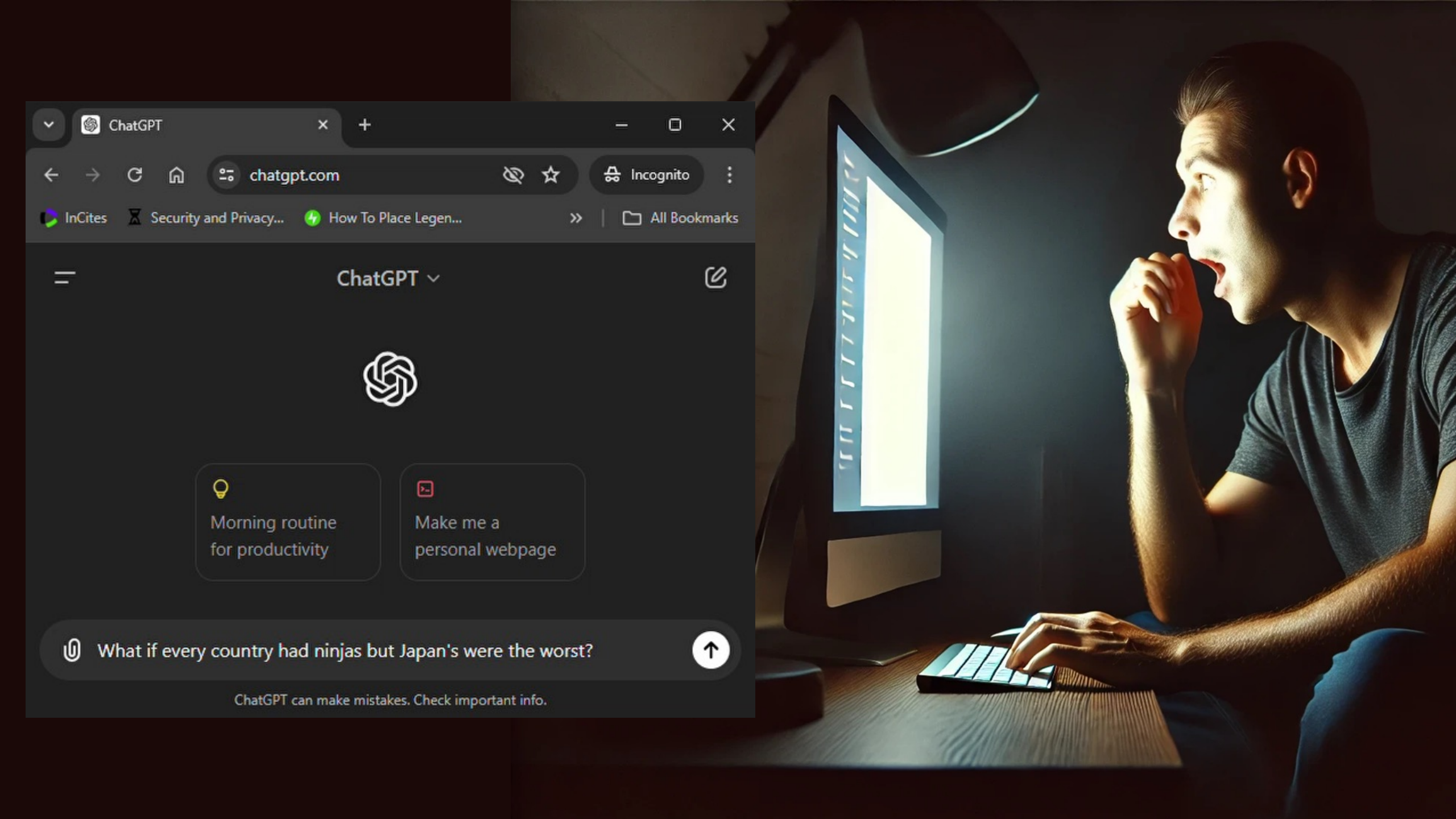
F1428BB4E81D1A3C57979143D1





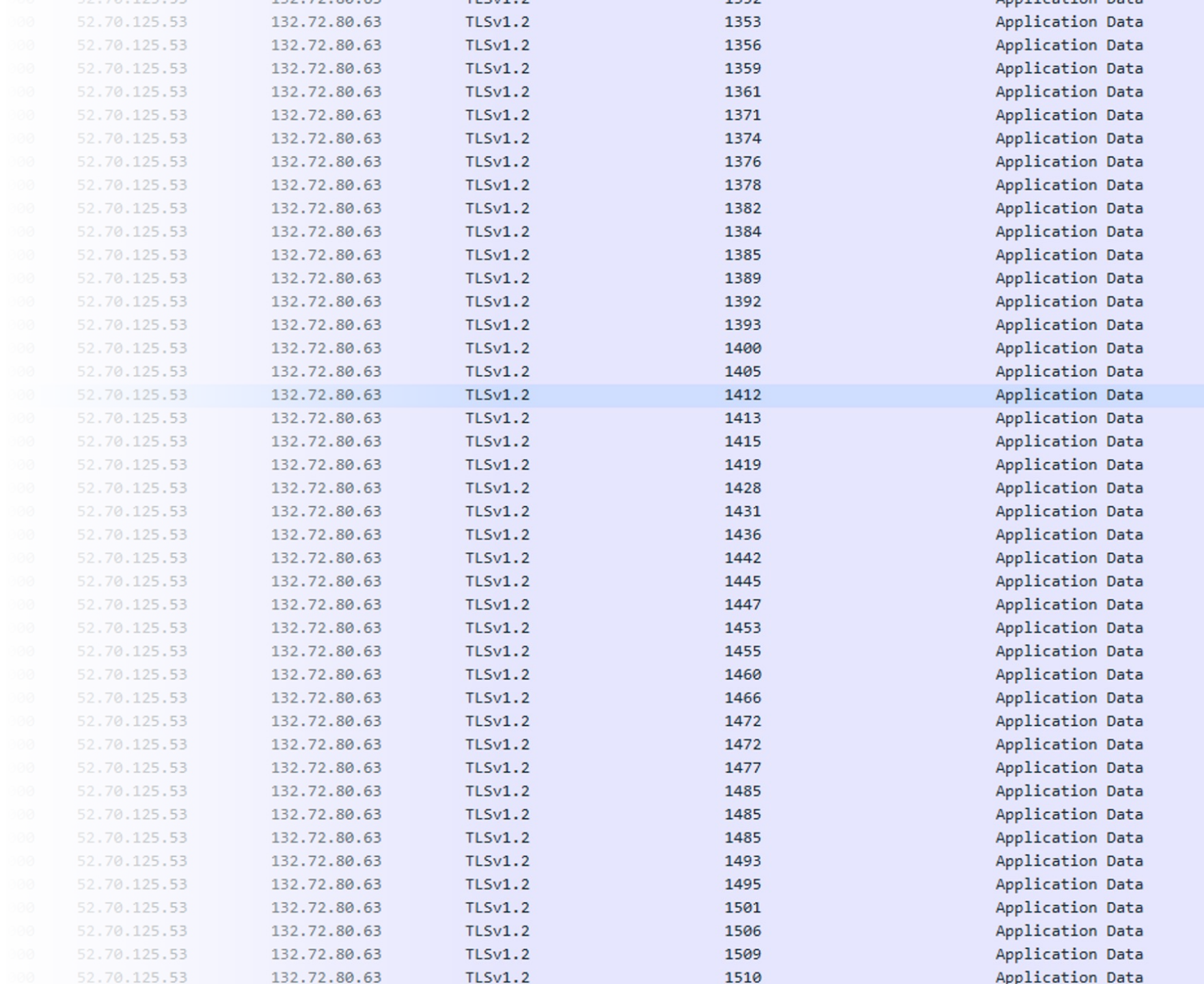
Eavesdropper



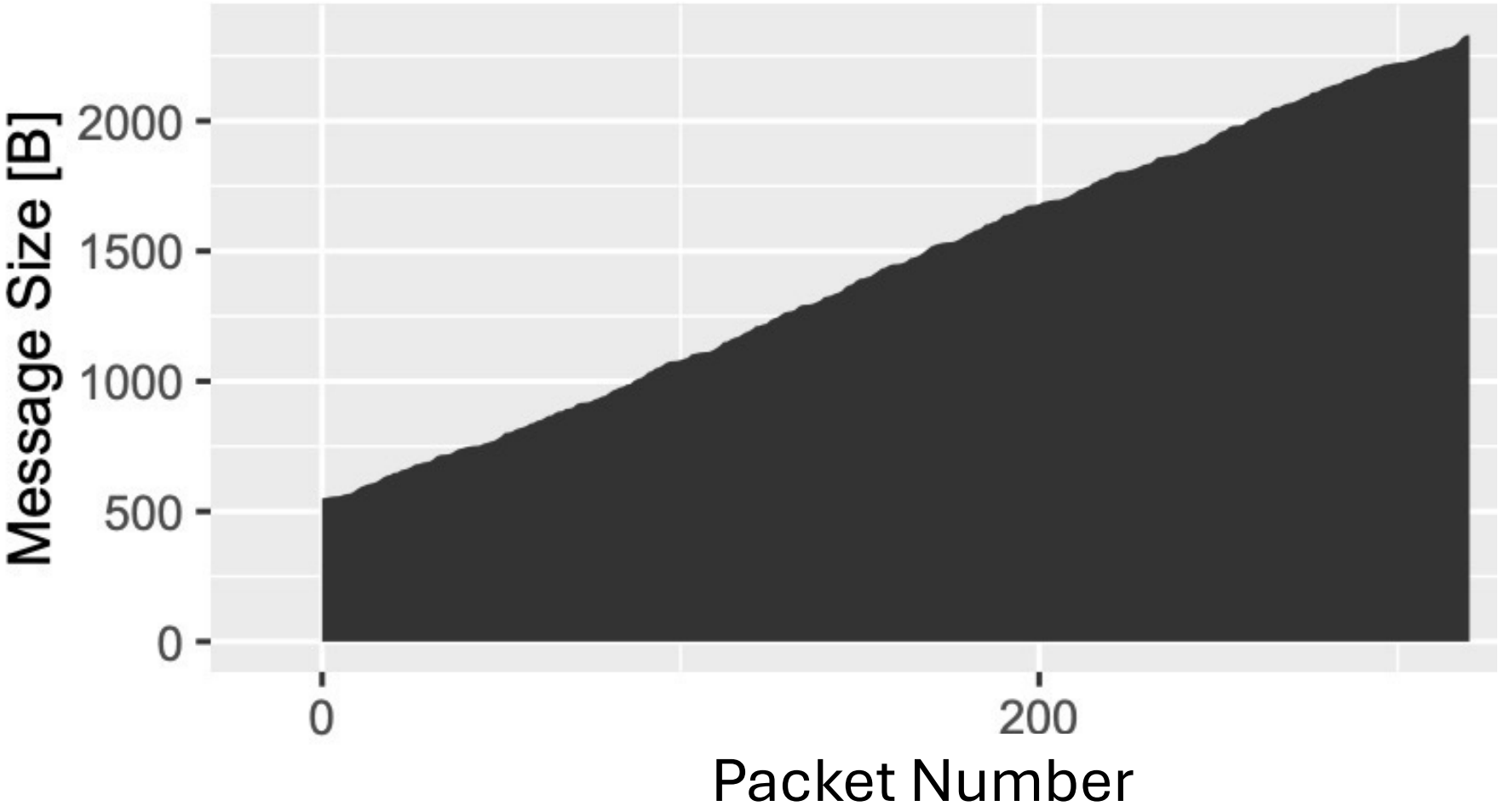

ChatGPT
chatgpt.com
Incognito
InCites
Security and Privacy...
How To Place Legen...
All Bookmarks
ChatGPT
Morning routine for productivity
Make me a personal webpage
What if every country had ninjas but Japan's were the worst?
ChatGPT can make mistakes. Check important info.

# Streaming Responses

- Words are sent one at a time
- Packets contain all previous words too
- No padding



| | | | | |
|---|---|---|---|---|
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1353 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1356 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1359 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1361 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1371 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1374 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1376 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1378 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1382 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1384 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1385 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1389 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1392 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1393 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1400 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1405 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1412 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1413 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1415 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1419 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1428 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1431 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1436 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1442 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1445 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1447 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1453 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1455 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1460 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1466 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1472 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1472 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1477 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1485 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1485 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1485 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1493 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1495 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1501 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1506 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1509 | Application Data |
| 52.70.125.53 | 132.72.80.63 | TLSv1.2 | 1510 | Application Data |




Difference in message sizes ≈ word length

# Tokens

- AI assistants use Large language models (LLM)
- LLMs read and generate text as tokens
- Tokenizers are standard and open source

**GPT-4** 


**LLAMA-2** 

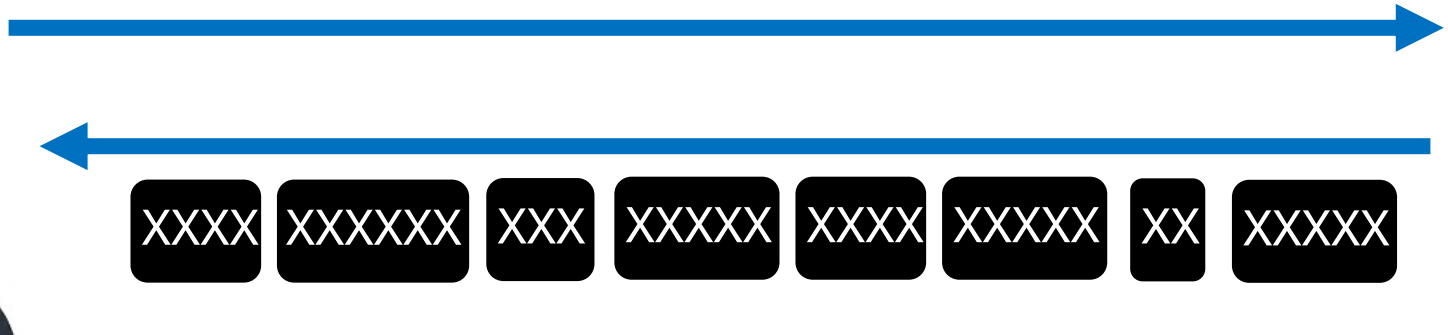

**Bottom line**: Tokens = Words

*and some punctuation*

# The **Token-Length** Side Channel



XXXXXXXXXXXXXXXXXXXXXXXXX





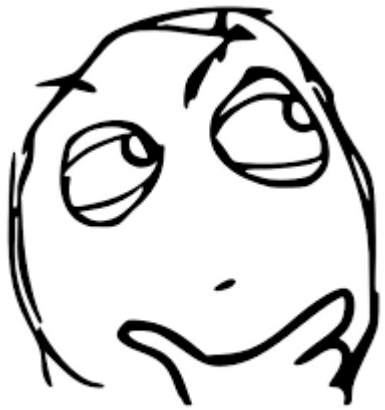
XXXX XXXXXX XXX XXXXX XXXX XXXXX XX XXXXX

I'm_ sorry_ to_ hear_ you_ have_ a_ rash_



Eavesdropper

# **The Threat** - Responses Reveal A lot



I’m sorry to hear you have a rash…

Your wife may be having an affair if you notice that she…

You may be experiencing drug addiction if…

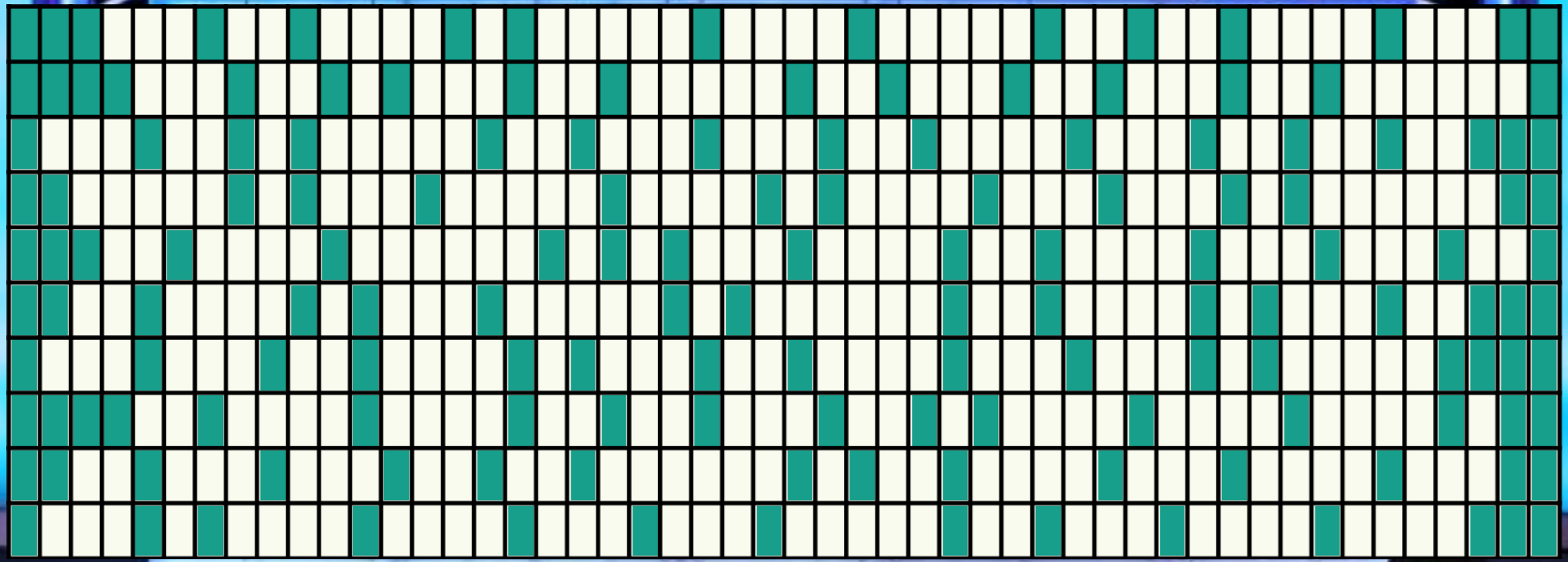
There are several ways you can hack…

Dear Frank,
I have spoken to the board and they have told me that we should sell…


WHEEL OF FORTUNE
GUESS THE GPT RESPONSE!
THE WORD FINDER

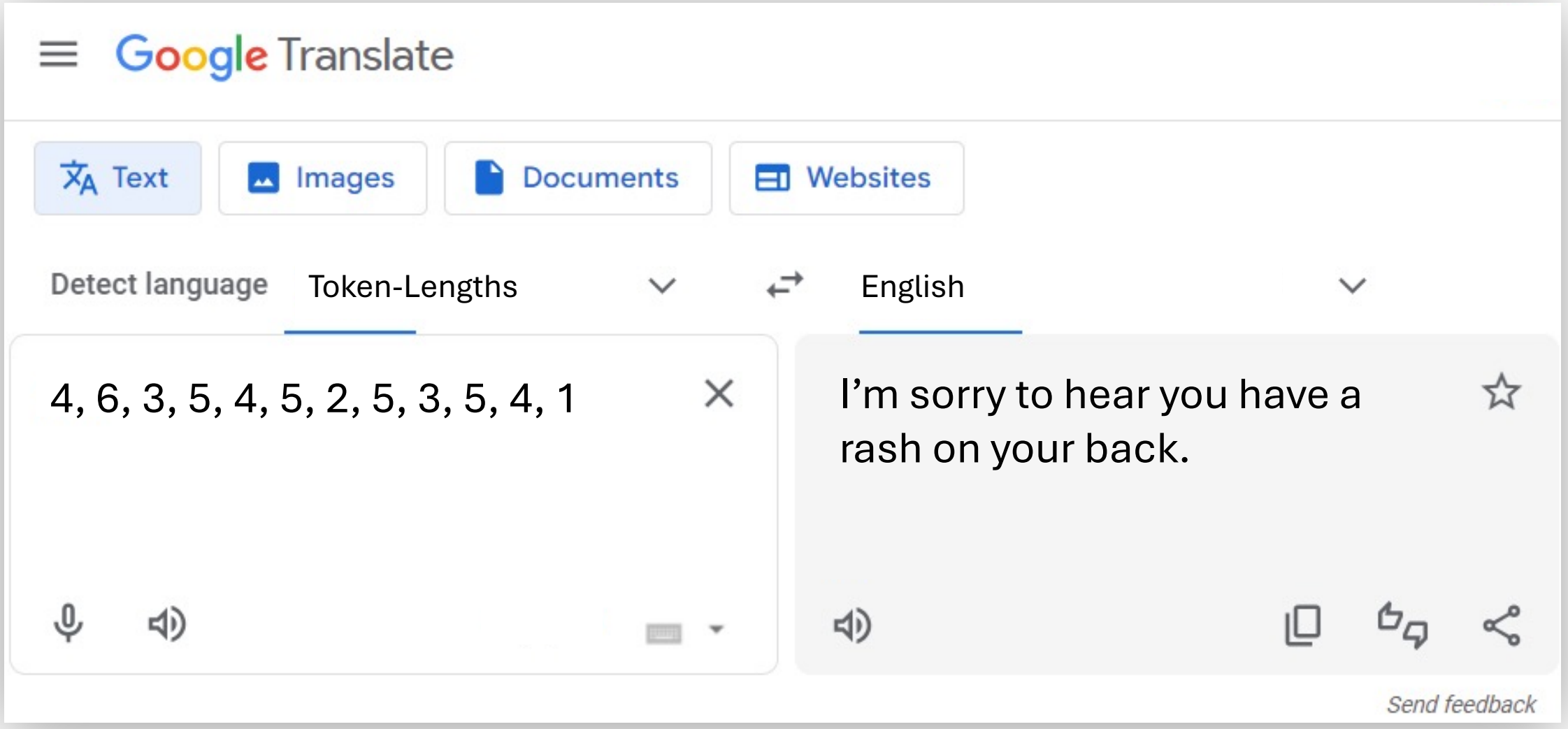

I ' M S O R R Y T O
H E A R T H A T Y O U
H A V E A R A S H O N
Y O U R B A C K
WHEEL OF FORTUNE
GUESS THE GPT RESPONSE!
THE WORD FINDER

# Guessing Responses is HARD!



WHEEL OF FORTUNE GUESS THE GPT RESPONSE! THE WORD FINDER

# Our Task - Translation



# Machine Learning to the Rescue!

**Why ML**

- Translation tools use ML
- ML learns from examples
- Excellent at pattern recognition

**Large Language Models (LLM)**

- They power ChatGPT and other services
- Huge context windows

# GPT-Keylogger

I'm sorry to hear you have a rash on your back.

**Token Lengths**

4, 6, 3, 5, 4, 5,
2, 5, 3, 5, 4, 1

**Attack Model**

*LLM trained on many examples*

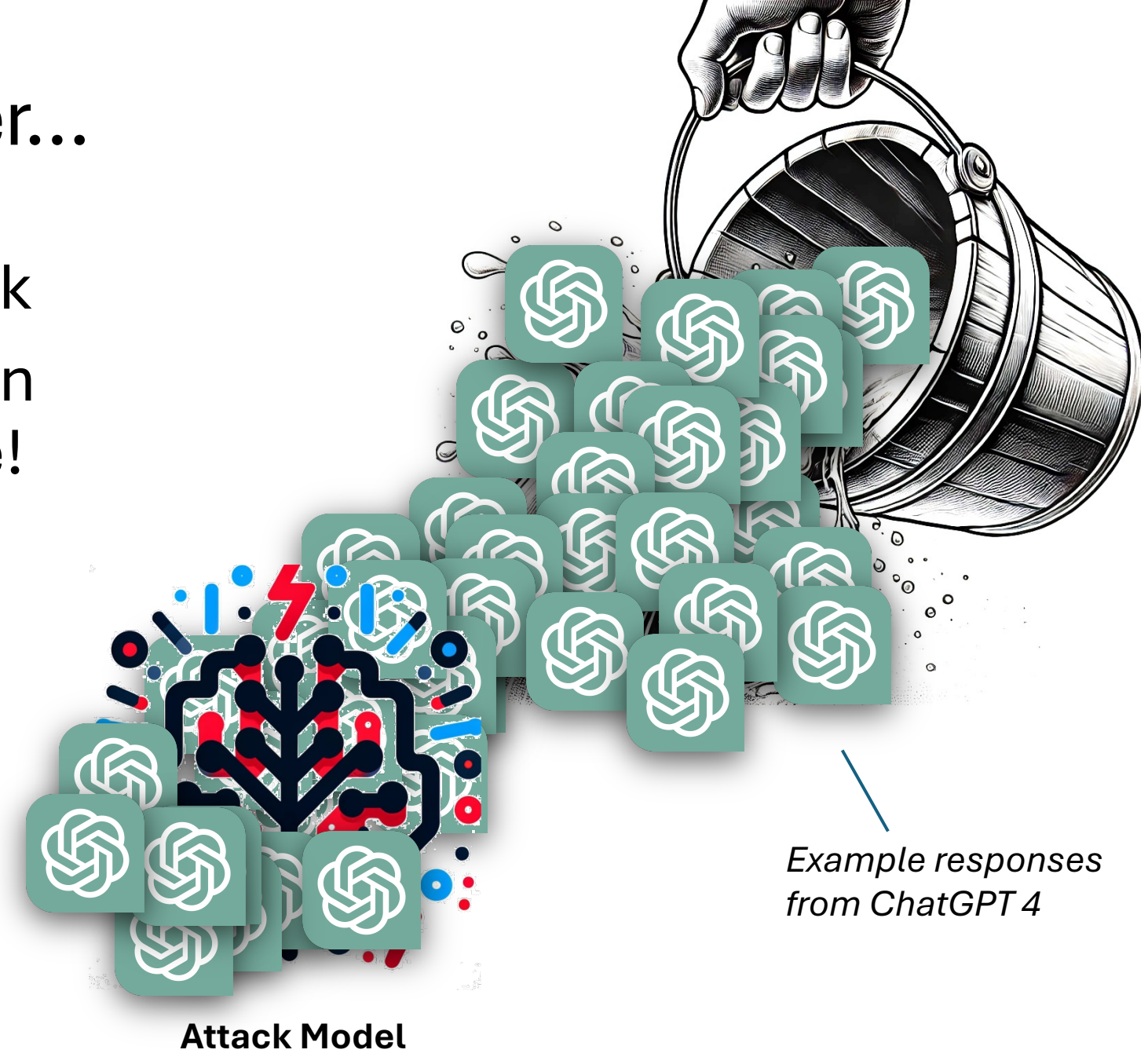


**Predicted Text**

I'm sorry to hear you have a rash on your **face**.

# Making it Even Better…

- Known-plaintext attack
- Fine-tune the model on the target chat service!



# Attack Success Rates on **OpenAI GPT-4**



Predicting First Sentence: 55%
Predicting Entire Text: 38%

More stats to follow...

# Successes – First Sentence

$\phi$ : 1.00 ROUGE-1: 1.00 Edit Distance: 0.00

The most common signs and symptoms of depression in young adults include:

The most common signs and symptoms of depression in young adults include:

$\phi$ : 0.80 ROUGE-1: 0.83 Edit Distance: 0.21

I don't have personal experience in **qualifying** for a career **option.**

I don't have personal experience in **preparing** for a career **change,**

$\phi$ : 0.92 ROUGE-1: 0.88 Edit Distance: 0.12

Yes, there are several **important** legal considerations that **couples** should be aware of when considering a divorce,

Yes, there are several **potential** legal considerations that **someone** should be aware of when considering a divorce.

$\phi$ : 0.66 ROUGE-1: 0.80 Edit Distance: 0.17

Alcohol consumption can have a harmful effect on **liver function**.

Alcohol consumption can have a harmful effect on **sleep patterns**.

# Successes – First Paragraph

$\phi$ : 0.93 ROUGE-1: 0.75 Edit Distance: 0.26

**Original**: As an AI language model, I don't have personal experience or emotions, but below are some strategies that can help in developing a self-compassionate attitude:
1. Practice mindfulness - Mindfulness can help you identify and observe your thoughts and emotions without judging or reacting to them. This can help you develop a more compassionate attitude towards yourself.
2. Treat yourself with kindness - In stressful situations, treat yourself the way you would treat a friend or loved one. Do something kind for yourself, such as taking a break, self-care activities, or giving yourself a compliment.
3. Change your self-talk - Instead of being self-critical or negative, try to reframe your self-talk in a positive and supportive manner.
4. Accept your imperfections - Acknowledge that everyone makes mistakes and has flaws, including yourself. Learn to accept yourself for who you are, imperfections and all.
5. Be patient with yourself - Give yourself time to process and bounce back from stressful situations. Remember that healing and growth take time.
6. Seek support - Reach out to friends, family, or a mental health professional for support, guidance,
**Prediction**: As an AI language model, I don't have personal experience or emotions, but below are some strategies that can help in developing a self-compassion attitude:
1. Practice mindfulness - Mindfulness can help you identify and process your emotions and thoughts without judging or reacting to them. This can help you develop a more compassionate attitude towards yourself.
4. Treat yourself with kindness - By embracing compassion,treat yourself the way you would treat a friend or loved one. By accepting your own mistakes, such as taking a break,resuming activities, or simply avoiding a temptation.
4. Change your self-talk - Instead of going into a battle or argument,try to reframe your self-talk in a positive and supportive manner.
3. Accept your imperfections - Acknowledge that everyone makes mistakes and has flaws. Challenge yourself, learn to accept yourself for who you are, treat oneself kindly.
5. Be patient with yourself - Your personal life is complex and unique with many different challenges. Practice self-control and manage your time.
6. Seek support - Reach out to friends,family, or a mental health professional for support, guidance,

# Fails – First Sentence

$\phi$ : 0.44 ROUGE-1: 0.64 Edit Distance: 0.36

I would suggest the following strategies for **team leaders** to balance the needs of **individual team members with** the **needs** of **the team as a whole**:

I would suggest the following strategies for **film studios** to balance the needs of **production with staying true to** the **story** of **a movie**:

$\phi$ : 0.12 ROUGE-1: 0.00 Edit Distance: 0.78

**Cider and other apple-based beverages hold great significance in Normandy's culture and cuisine.**

**Coral bleaching occurs after organisms lose their reproduction or grow out a suitable new habitat.**

# How common is this vulnerability?

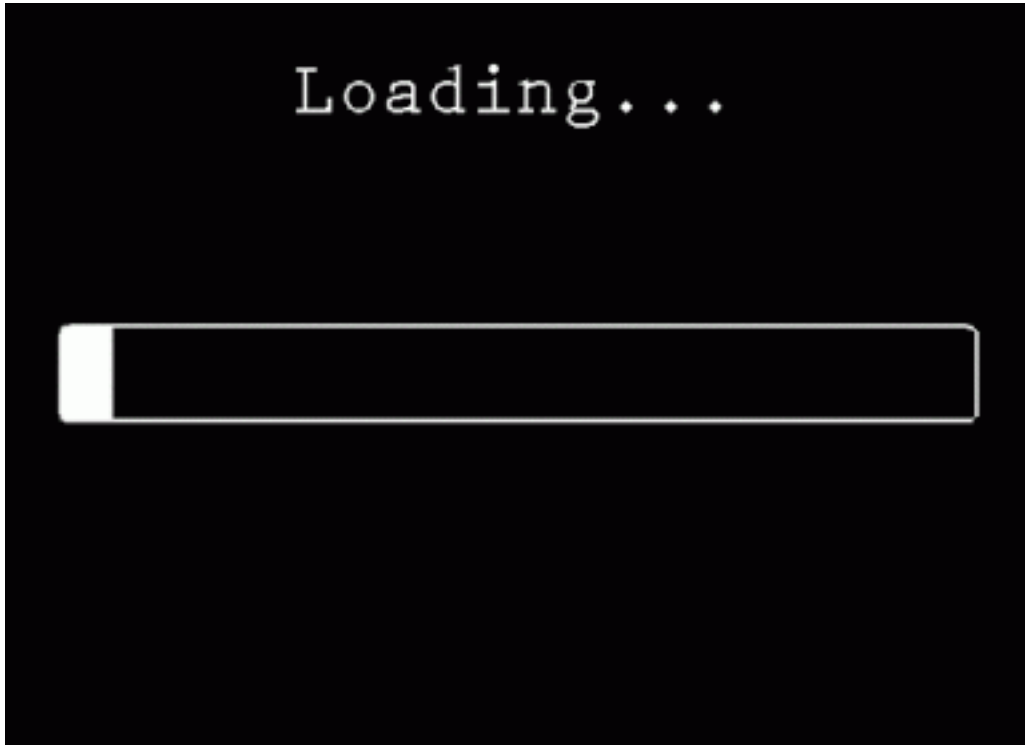
Before disclosure Feb ‘24



# Why is this vulnerability overlooked?

1. Token lengths ≠ word lengths
2. Problem considered too hard for adversary
   - Guessing with no context
   - Long word length sequences
3. AI Gold Rush

# Okay, but why stream the tokens?



# Recap so far…

**Vulnerability**

- Discovery of the token-length side channel
- Common in LLM services

**Exploitation**

- GPT-Keylogger
  - A custom AI model that can **use** the side channel
  - Known-plaintext attack

# The Rest of this Talk

*Deep Dive*

Finding the Side Channel → Token-Length Extraction → Exploitation with AI → Demos & Defences

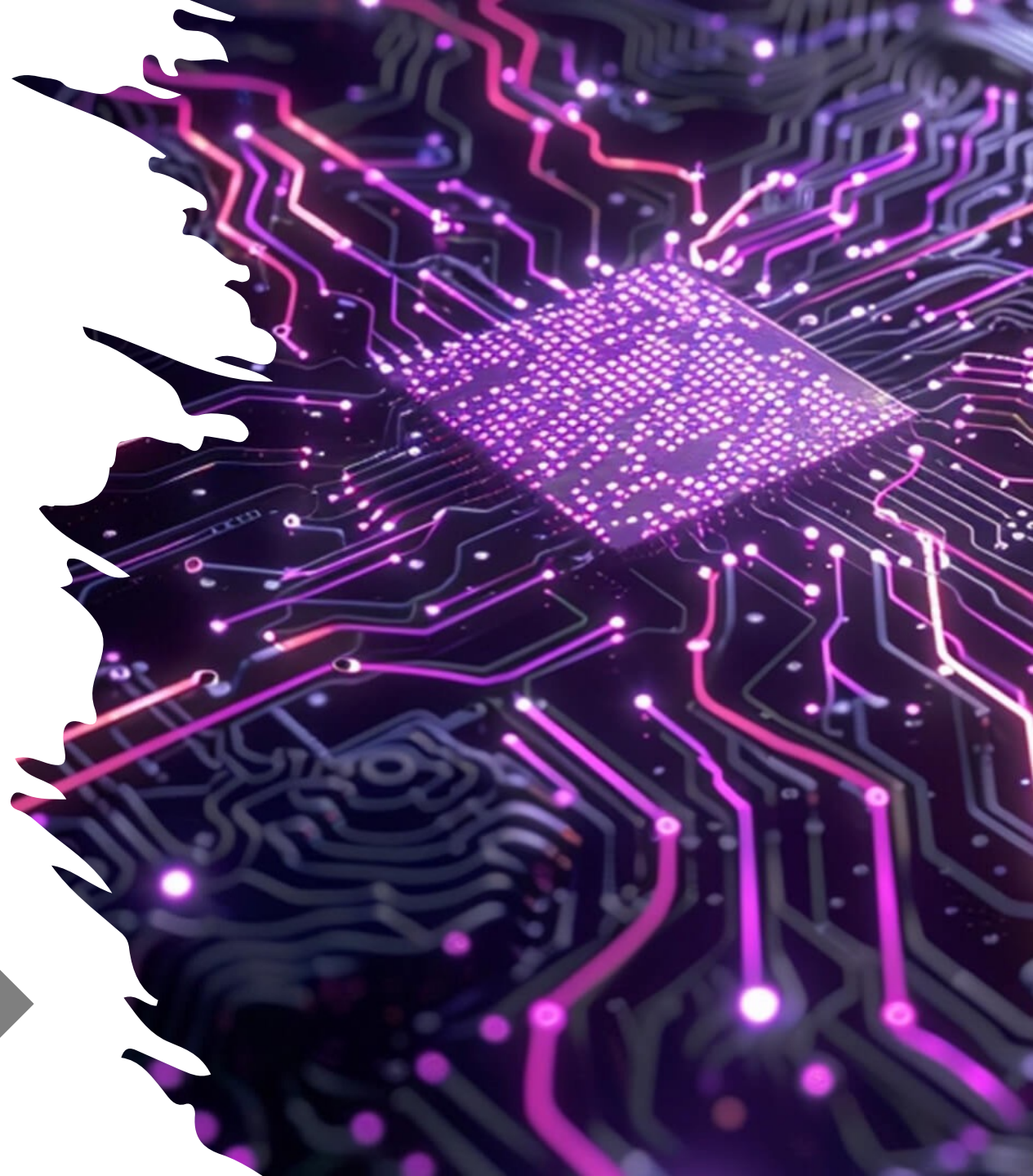








# Finding the Side Channel



Finding the Side Channel | Token-Length Extraction | Exploitation with AI | Demos & Defences

# Traffic Capture

The attacker needs to be able to sniff the victims incoming traffic.

- Same local network (e.g. Café, Library, etc…)
- ISP
- …



# Filtering the Traffic

**Objective**: Find Response Traffic

- Filter by IP address
  - OSINT
  - Connect as client
- Filter by protocol



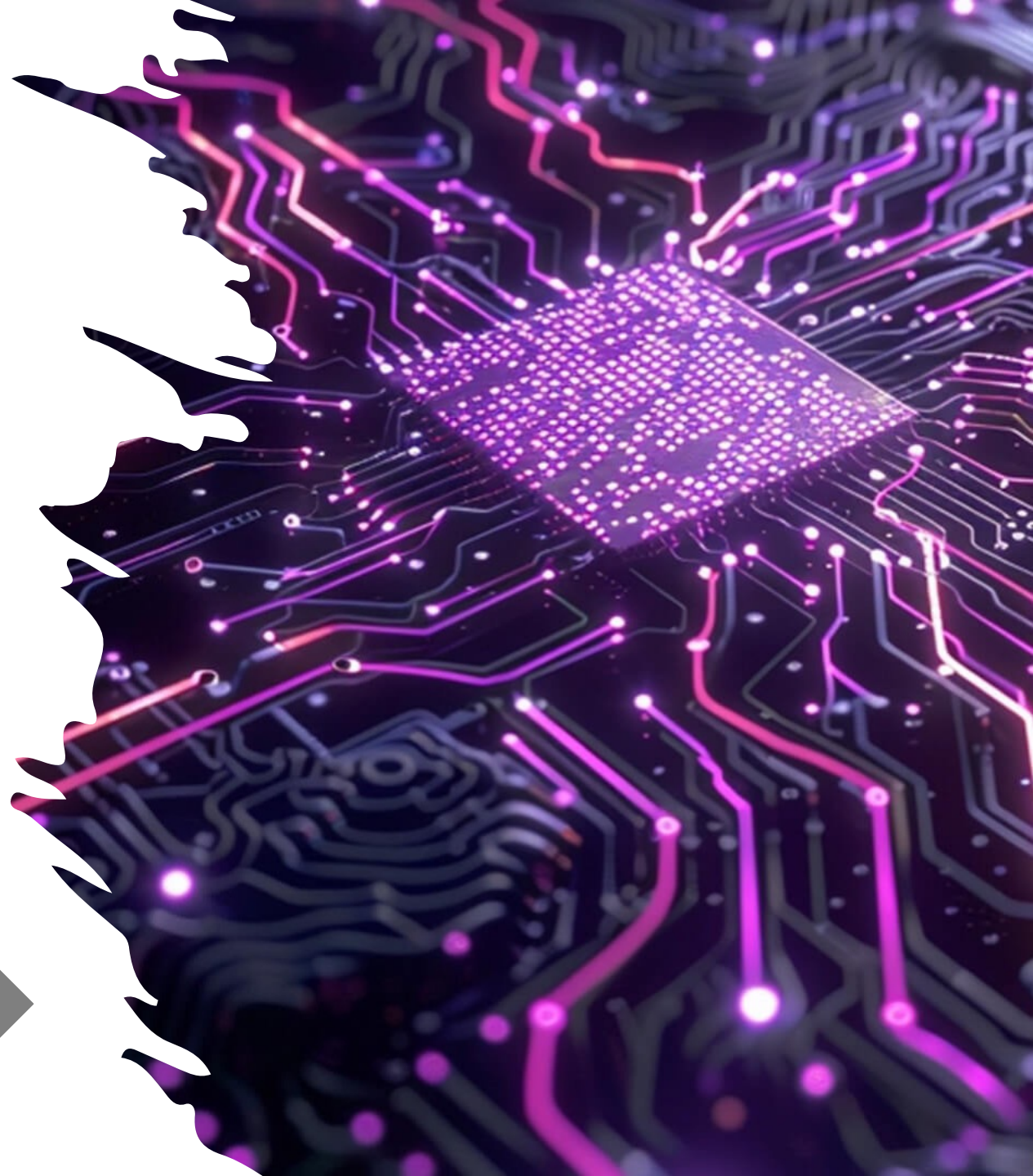
# Protocols

- Different Protocols are used to stream tokens
- Depends on the vendor

| Protocol | Vendor(s) |
|---|---|
| QUIC | ChatGPT |
| WebSocket | Microsoft CoPilot, Preplexity |
| HTTP2 | Claude, Writesonic, Notion AI, ClickUp AI, Text Cortex AI |
| TCP | Quora Poe, OpenAI API, CloudFlare AI API |

# Token-Length Extraction

Finding the Side Channel | Token-Length Extraction | Exploitation with AI | Demos & Defences



# From Traffic to Token-length Sequence

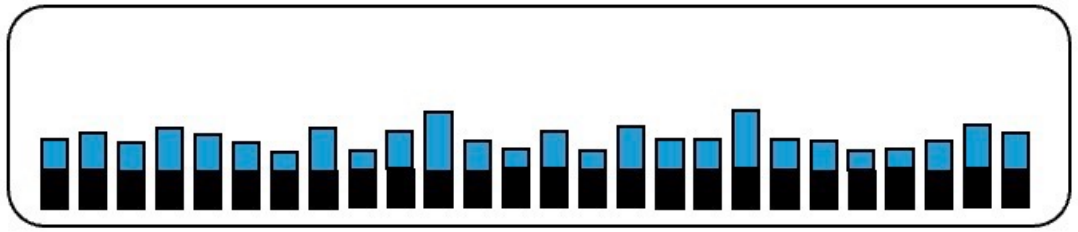
**Objective**: find the token-lengths

- Traffic is encrypted
- But, we can see bandwidth patterns

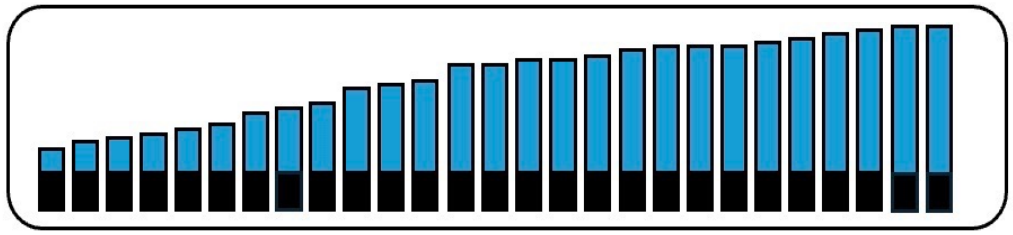
# How Tokens are Sent

We have observed 2 ways vendors send the next token.

## Base + Next Token



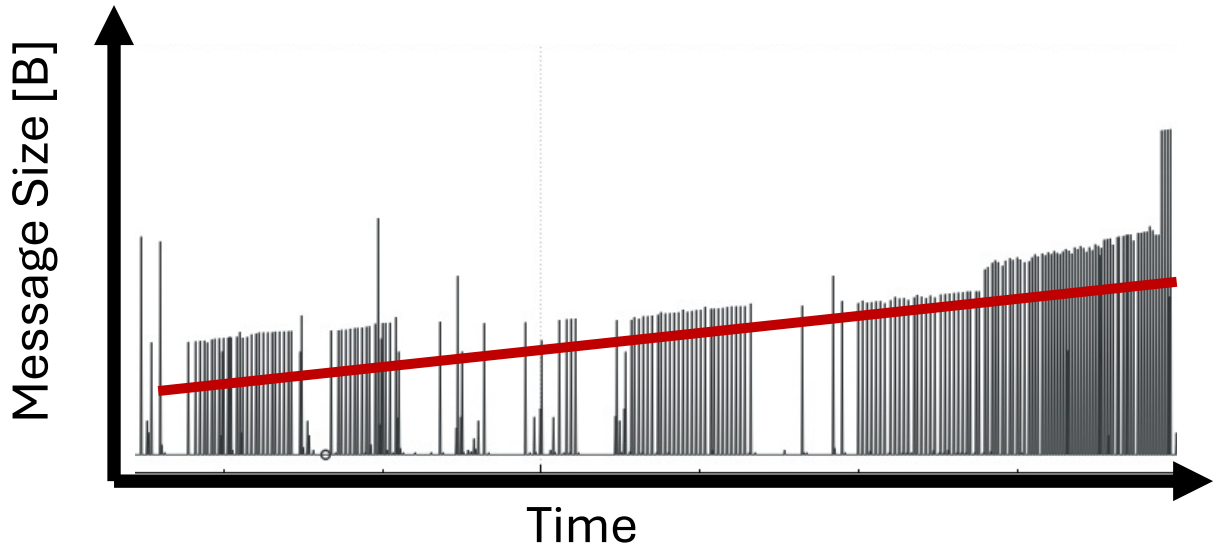
## Token Accumulation



# Identifying the Messages from a Response

Finding the start and end of a response

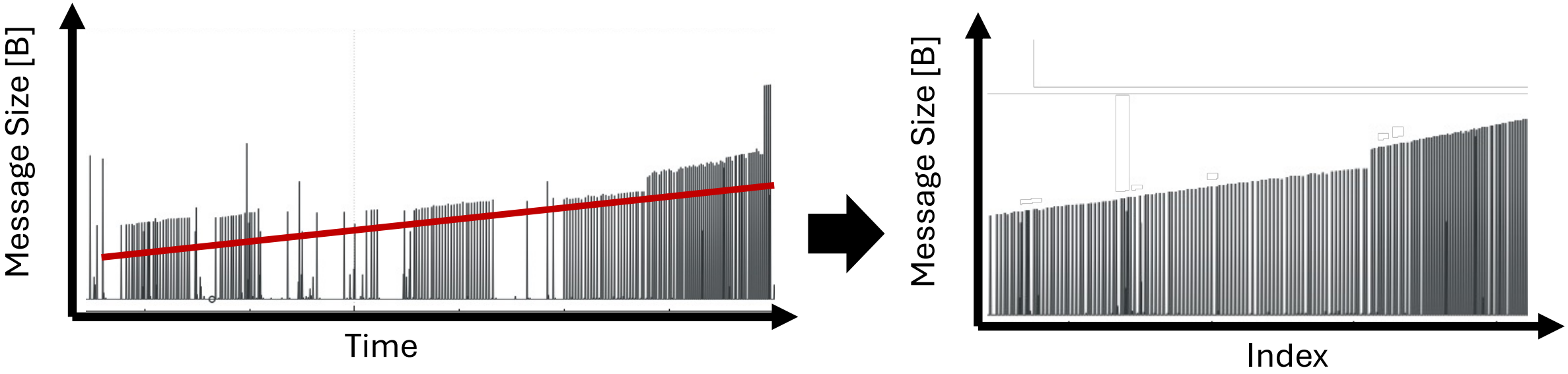
💡 Remember: every vendor is different.




* Traffic is **encrypted**, so we must find message tokens based on bandwidth patterns alone

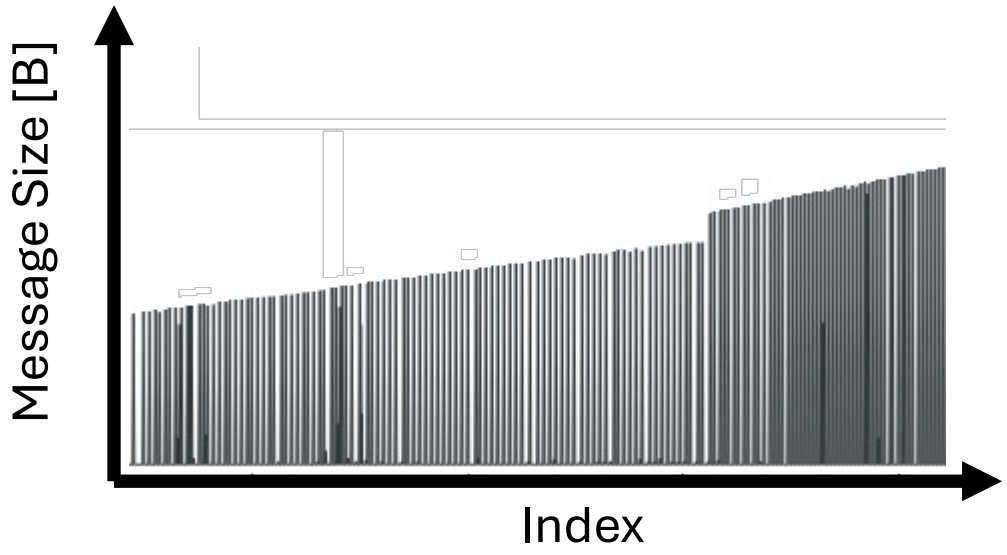
# Cleaning it Up

- Some messages **do not** contain tokens
  - Handshakes, control, …
- We remove these by following the monotonic trend
- …there is probably a better way



# Token Length Extraction

To get the token lengths, we compute the delta between messages



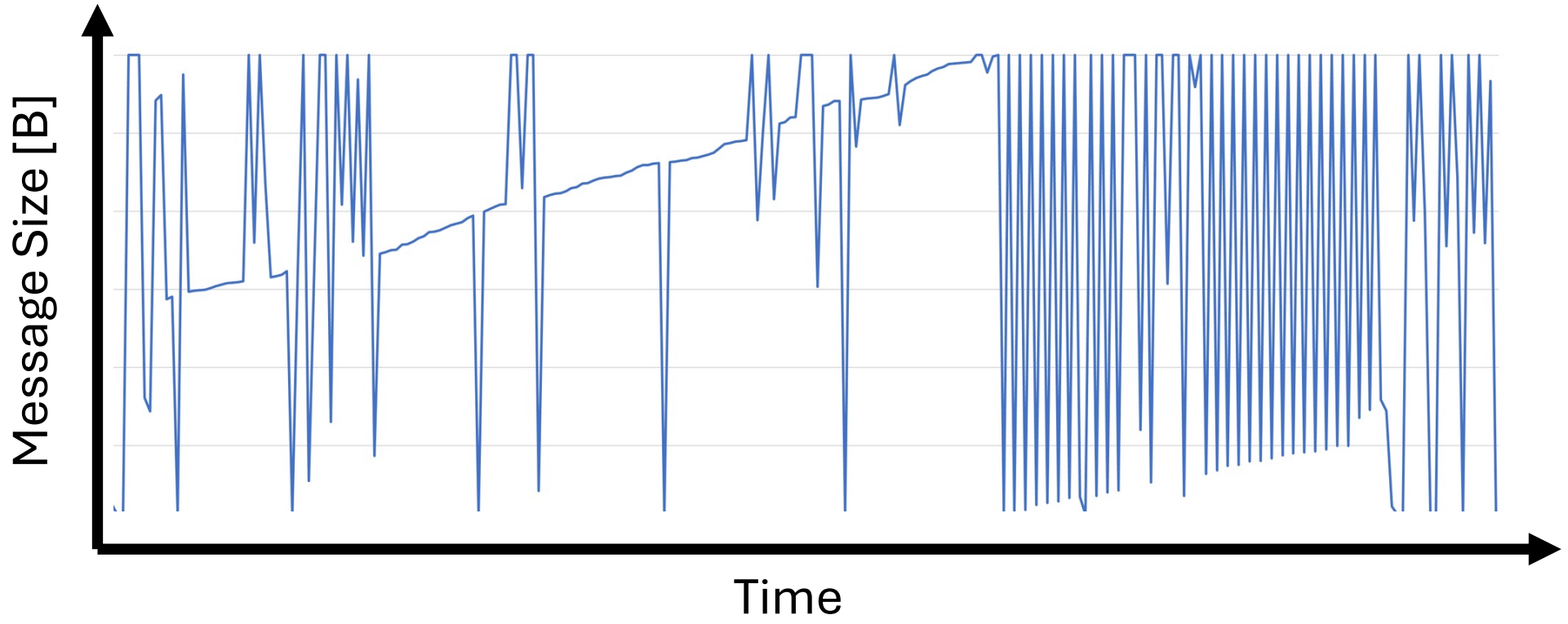

4, 6, 3, 5, 4, 5, 2, 5, 3, 5, 4, 1, …

# Real World Challenges

- Message Size Limits
- Message Buffering -> Message Grouping

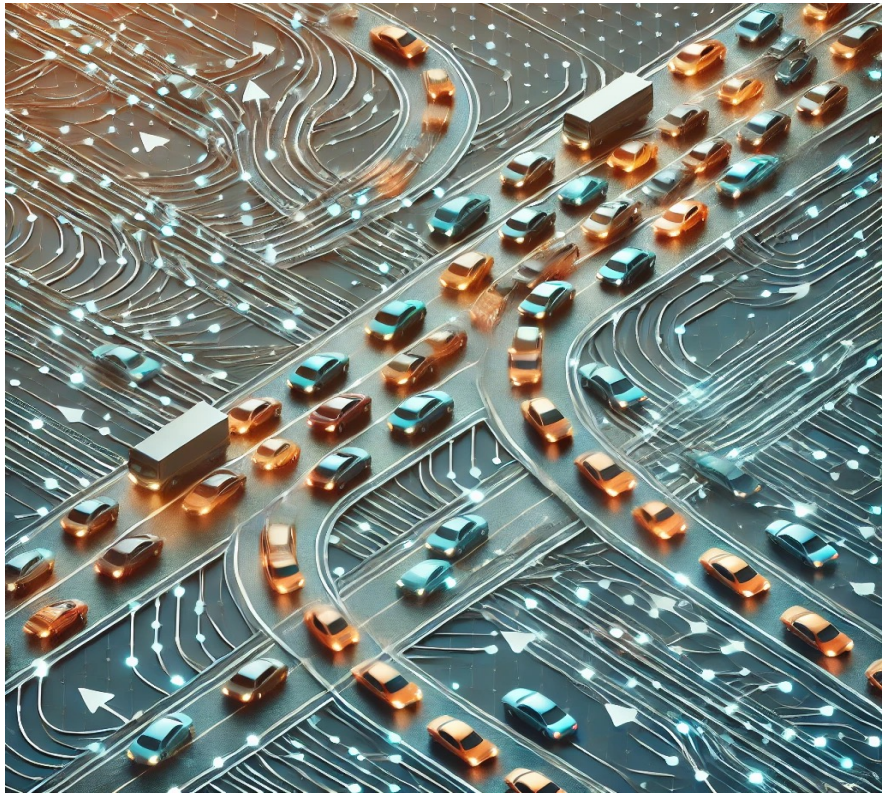
# Message Size Limits

- Maximum transmission unit, QUIC has even lower max message size
- Results in a saw-tooth bandwidth pattern

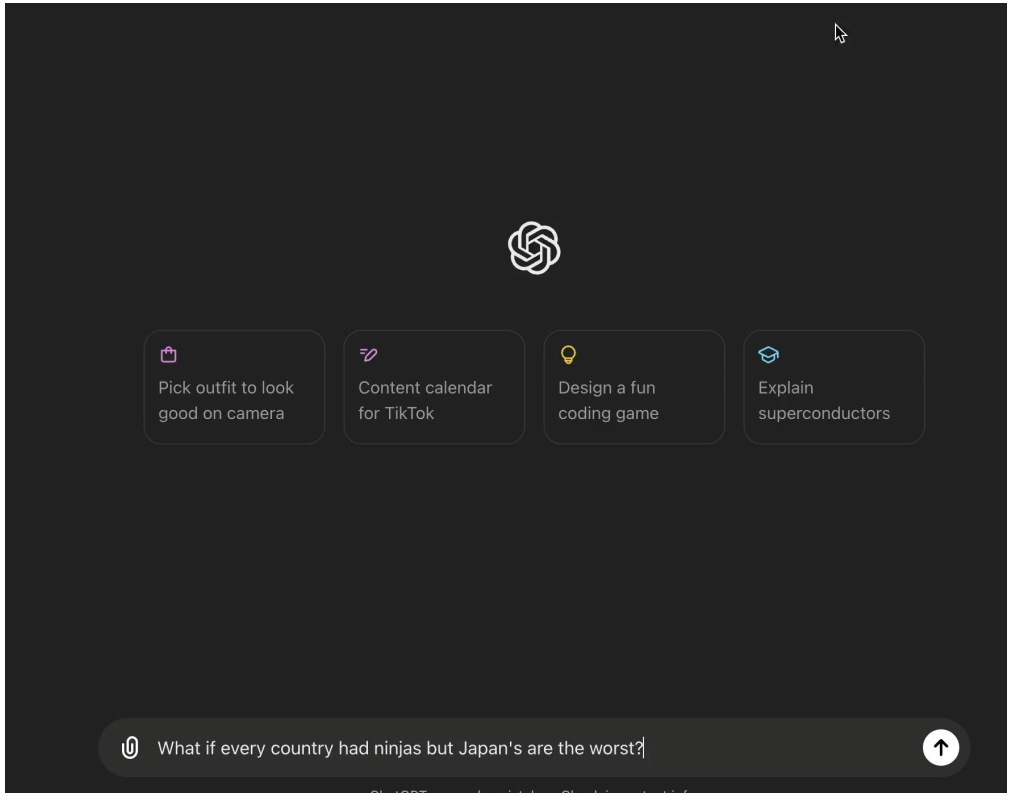


# Message Buffering

## Busy Network

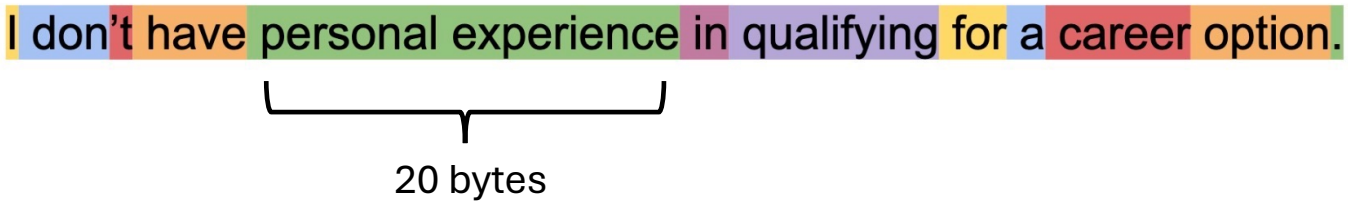


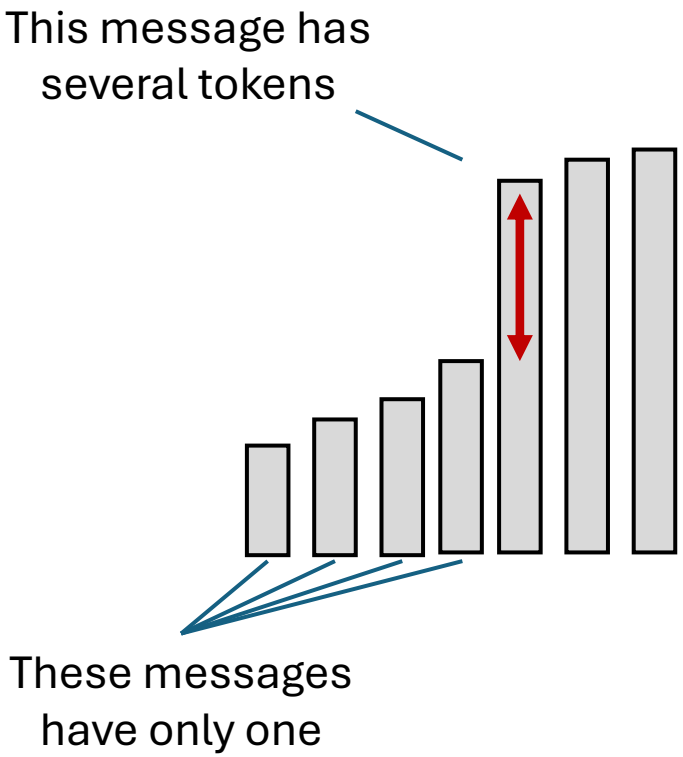
## More Buffering



# Message Buffering

- Our main target was **OpenAI's ChatGPT**
- A challenge with QUIC
  - Messages can be assembled on client side
- Sometimes we get grouped messages



# How did we figure this all out?

## Reverse Engineering

**Step 1:** Connect as legitimate client

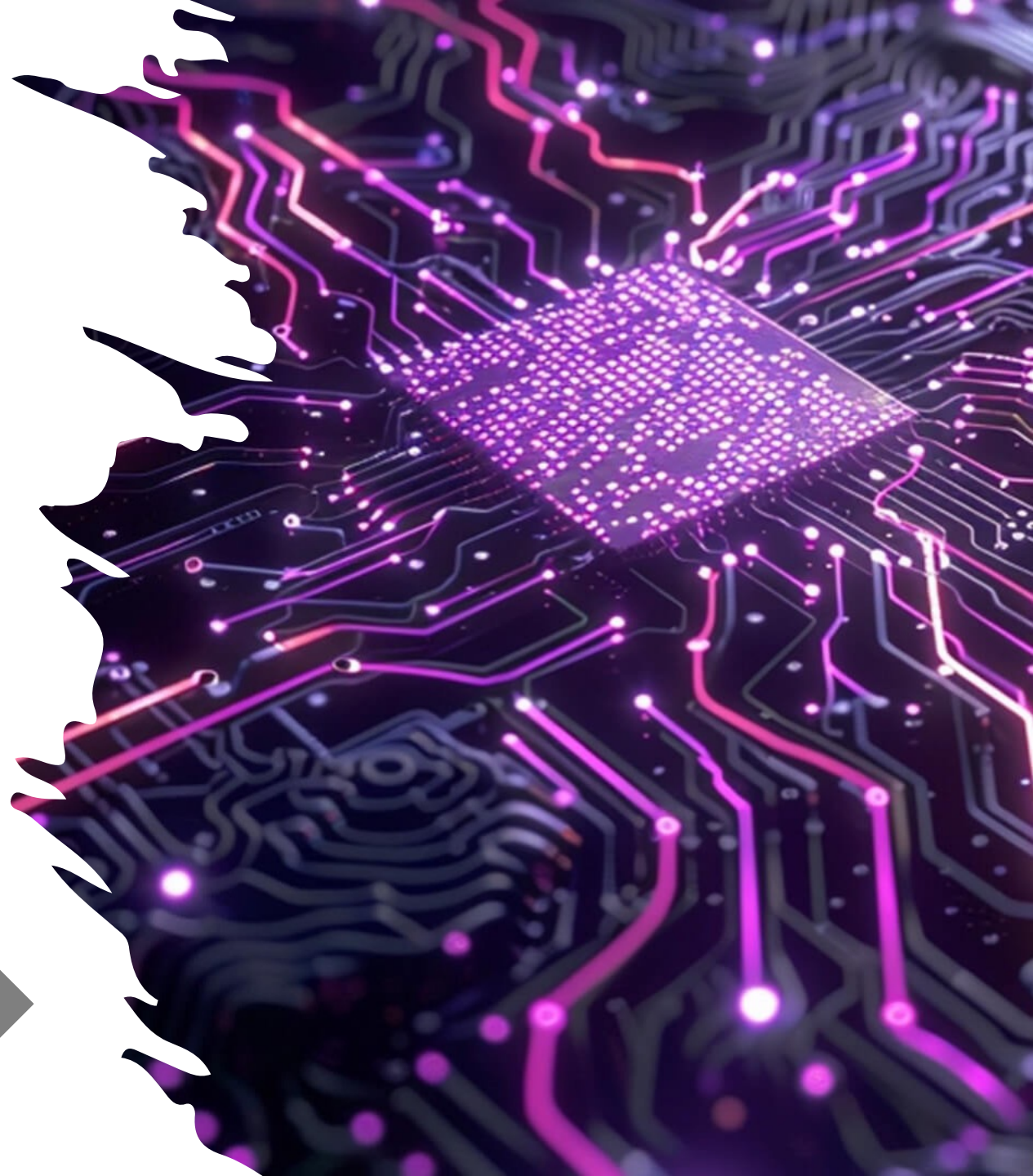
**Step 2:** Grab SSL key

**Step 3:** Plug it into Wireshark



# Exploitation with AI

Finding the Side Channel | Token-Length Extraction | Exploitation with AI | Demos & Defences



# How to Train your ML Model



# The Procedure

## Create Dataset

| | |
|---|---|
| The best way to save money... | 3, 5, 4, 3, 5, 6 ... |
| Planning a trip to Europe... | 8, 2, 5, 3, 7 ... |
| Finding a job you love requires... | 7, 2, 4, 4, 5, 9 ... |
| ⋮ | ⋮ |

## Train Model

3, 5, 4, 3, 5, 6 ...
8, 2, 5, 3, 7 ...
7, 2, 4, 4, 5, 9 ...
⋮



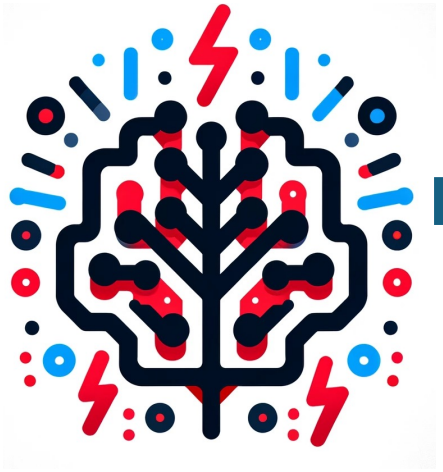
The best way to save money...
Organizing a trip to Europe...
Securing a job you love requires...
⋮

# The Procedure

**Use The Model**



3, 4, 5, 2, 5, 5...



**Generate Options & Rank them**

If you have a rash then...

If you have a **pain** then...

⋮

If **one cuts** a **pear five**...

⋮

# And...

Dealing with a rash on your back involves several steps:

1. Identify the cause: Determine if the rash is due to an allergy, irritation, infection, or other condition.
2. Keep the area clean: Gently wash the rash with mild soap and water.
3. Apply a soothing lotion: Use calamine lotion or an anti-itch cream to relieve discomfort.
4. Avoid scratching: Scratching can worsen the rash and lead to infection.
5. Wear loose clothing: Prevent further irritation by wearing loose, breathable fabrics.
6. …

6, 5, 2, 5, 3, 5, 5, 7, 8,
6, 1, 1, 8, 4 ,6, 8, 3, 7,
3, 2, 1, 1, 5, 8, 4, 3 …

**Token-length Sequence**



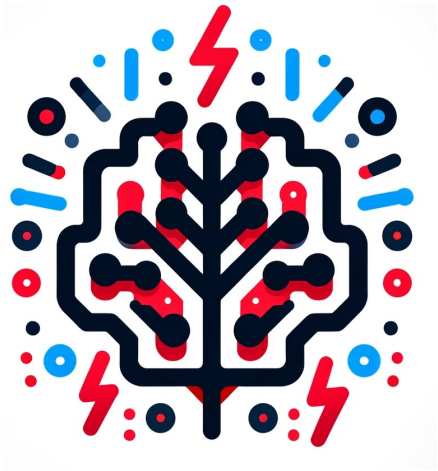
# And… didn't work : (

Dealing with a rash on your back involves several steps:

1. Identify the cause: Determine if the rash is due to an allergy, irritation, infection, or other condition.
2. Keep the area clean: Gently wash the rash with mild soap and water.
3. Apply a soothing lotion: Use calamine lotion or an anti-itch cream to relieve discomfort.
4. Avoid scratching: Scratching can worsen the rash and lead to infection.
5. Wear loose clothing: Prevent further irritation by wearing loose, breathable fabrics.
6. …

6, 5, 2, 5, 3, 5, 5, 7, 8,
6, 1, 1, 8, 4 ,6, 8, 3, 7,
3, 2, 1, 1, 5, 8, 4, 3 …

**Token-length Sequence**



Preparing for a job interview involves several steps:

1. Research the company: Learn about their mission, values, and recent news.
2. Review the job description: Understand the skills and qualifications required.
3. Practice common questions: Rehearse answers to typical interview questions.
4. Plan your outfit: Choose professional and appropriate attire.
5. Bring necessary documents: Have copies of your resume and references ready.
6. …

# Eureka!

**Supporting a friend with depression involves being present and understanding.** Listen without judgment and offer a safe space for them to express their feelings. Encourage them to seek professional help from a therapist or counselor. Avoid giving unsolicited advice or trying to "fix" their problems. Instead, reassure them of your support and remind them that they are not alone. Engage in activities they enjoy, and be patient as they navigate their journey to...

**If you suspect your child is being bullied, the first step is to talk to them openly and calmly.** Create a safe environment for them to share their experiences without fear of judgment. Document any incidents they describe, noting dates, times, and details. Contact the school to discuss your concerns with teachers, counselors, or administrators. Advocate for your child by requesting a plan of action to address the bullying. Encourage your child to develop coping strategies...

**Discussing substance abuse with a loved one requires sensitivity and care.** Choose a private and comfortable setting to have the conversation. Approach the topic with empathy and concern rather than judgment or accusation. Use "I" statements to express your feelings and observations, such as "I've noticed you've been drinking more lately, and I'm really worried about you." Encourage them to share their perspective and listen actively without interrupting. Offer...

# Divide and Conqure

**Discussing substance abuse with a loved one requires sensitivity and care.** Choose a private and comfortable setting to have the conversation. Approach the topic with empathy and concern rather than judgment or accusation. Use "I" statements to express your feelings and observations, such as "I've noticed you've been drinking more lately, and I'm really worried about you." Encourage them to share their perspective and listen actively without interrupting.



# Divide and Conqure

- **Discussing substance abuse with a loved one requires sensitivity and care.**
- Choose a private and comfortable setting to have the conversation.
- Approach the topic with empathy and concern rather than judgment or accusation.
- Use "I" statements to express your feelings and observations, such as "I've noticed you've been drinking more lately, and I'm really worried about you."
- Encourage them to share their perspective and listen actively without interrupting.



# How do we find sentences?

Token-length sequence (3 sentences):

1 5 5 8 6 5 5 1 6 3 3 1 4 5 5 3 5 5 1 5 1 4 7 4 2 7 1

Most Tokenizers make punctuation marks (. , :) separate tokens

⇒meaning they are sent as a separate message

⇒meaning their length is 1

⇒Too short, probably a comma…

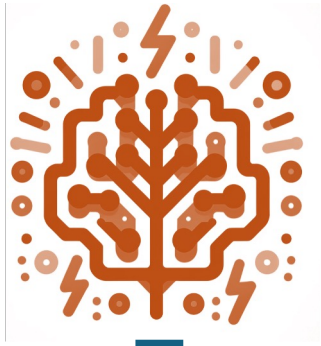
1 5 5 8 6 5 5 **1** 6 3 3 **1** 4 5 5 3 5 5 **1** 5 **1** 4 7 4 2 7 **1**

# The Full Model

## Parse Token-length Sequence

1 5 5 8 6 5 5 **1** 6 3 3 **1** 4 5 5 3 5 5 **1** 5 **1** 4 7 4 2 7 **1**

**Model A**



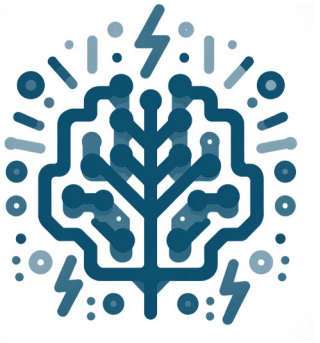
I need more details about your rash.

**Model B**



Where is it, and what does it look like?

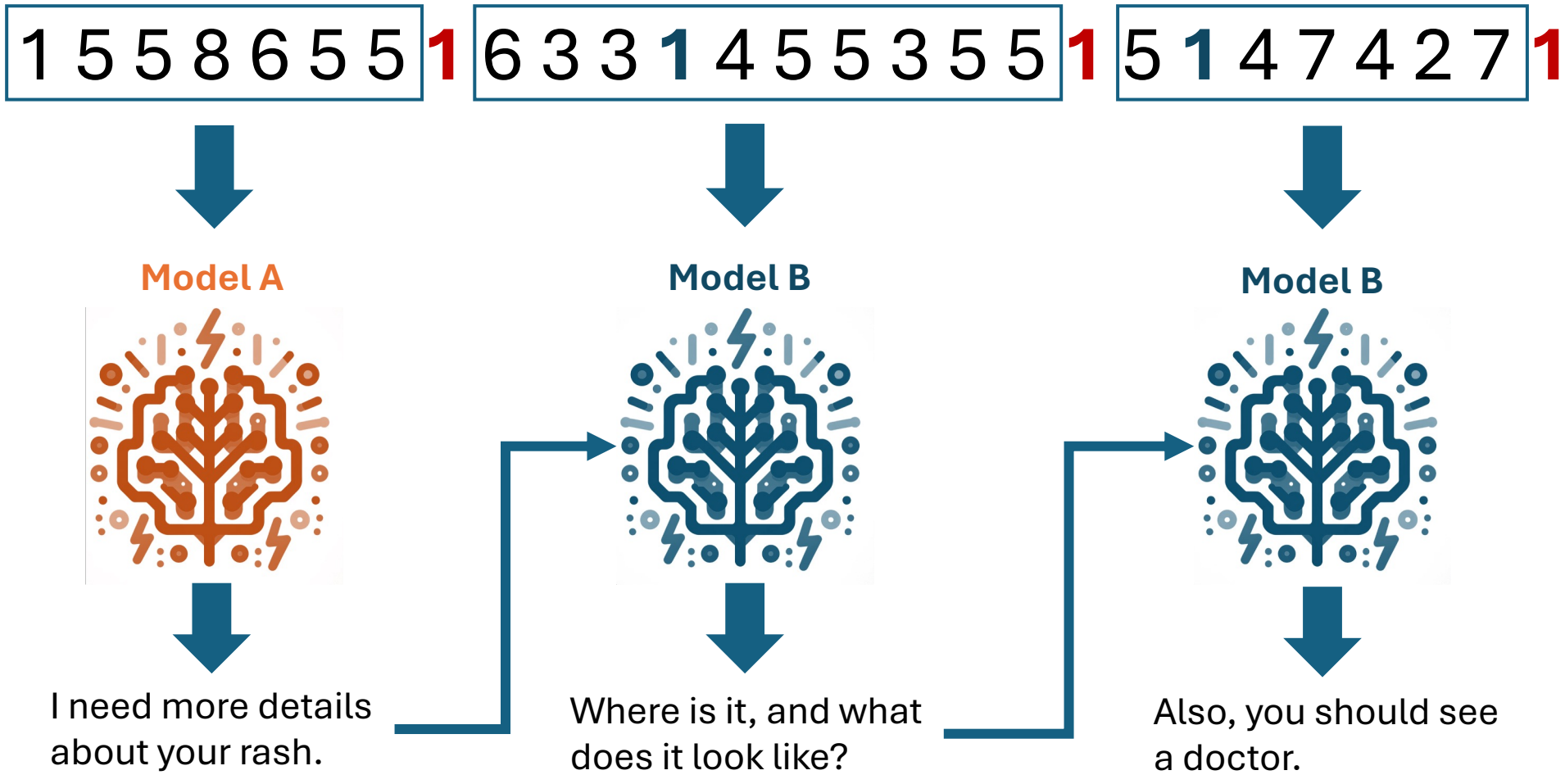
**Model B**



Also, you should see a doctor.

# The Full Model – more context

**Parse Token-length Sequence**



# Known-plaintext Attack

AI Assistants have unique speaking style

1. **Predictable Style**
   Mannerism, structure, …
2. **Low Response Diversity**
   Prompt variants lead to same response

…we can use this to our advantage

# Known-plaintext Attack

## Predictable Style

### Warnings

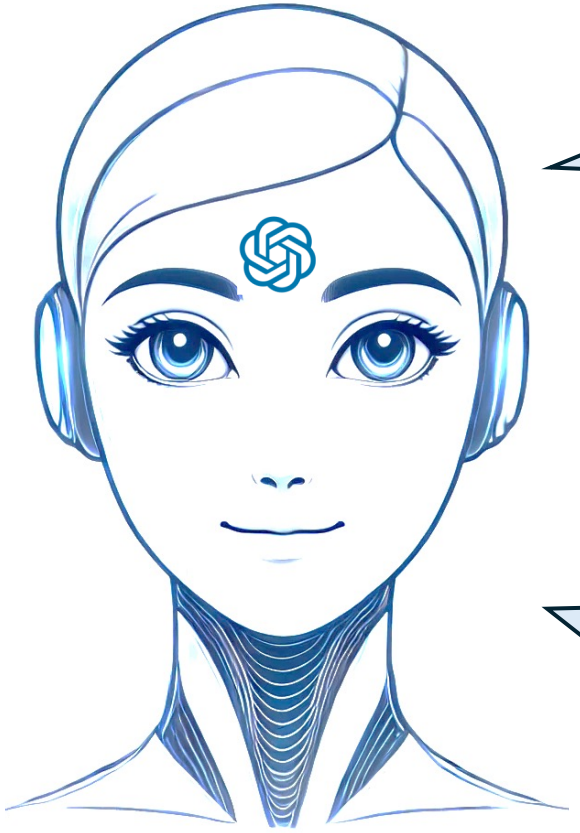
I don't have personal beliefs or interests, but..

### Unique Phrases

"Yellowstone National Park"

### Structure

- Apples\n
- Milk\n
- ...



### Templates

Certainly! Here's a simple and classic recipe for...

### AI Safeguards

I'm sorry, but I can't assist with creating content that could be used for phishing

# Known-plaintext Attack

## Low Response Diversity



# The Procedure – *known-plaintext*

## Create Dataset

How can I...
Help me plan...
How to...



The best way to save money...
Planning a trip to Europe...
Finding a job you love requires...

3, 5, 4, 3, 5, 6 ...
8, 2, 5, 3, 7 ...
7, 2, 4, 4, 5, 9 ...

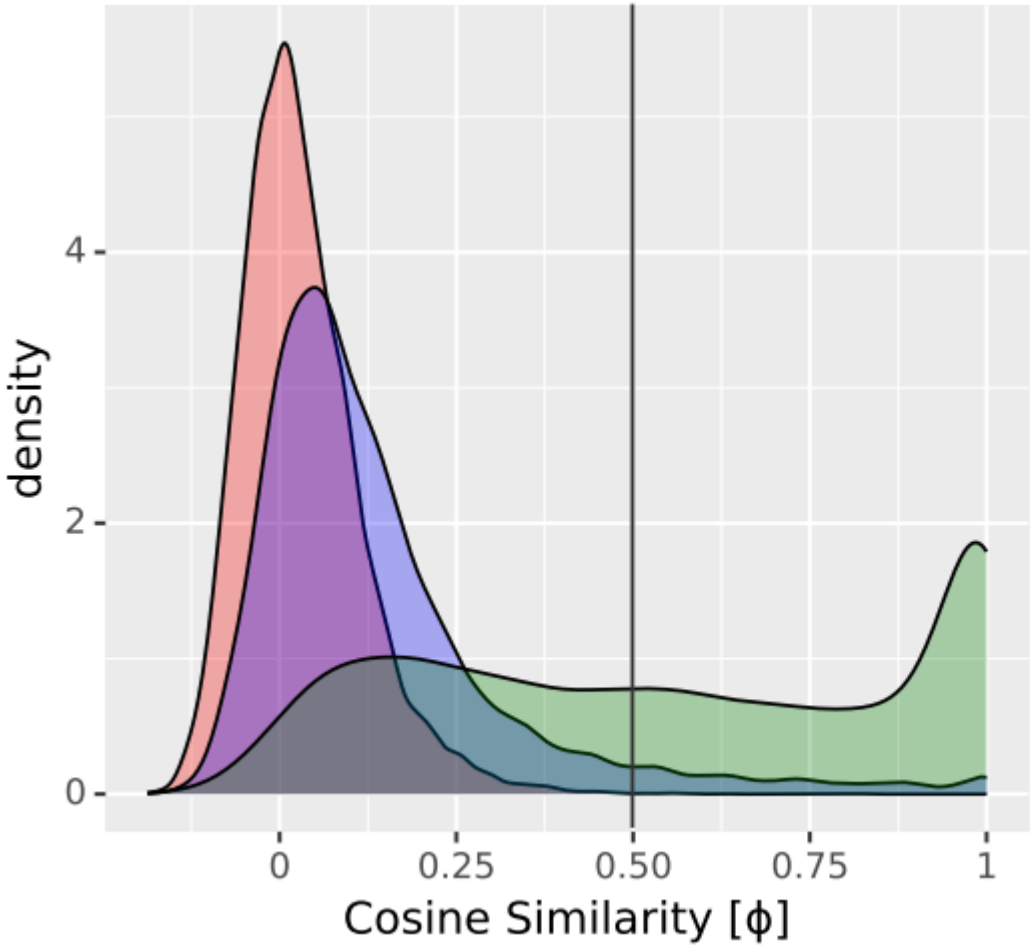
## Train Model

3, 5, 4, 3, 5, 6 ...
8, 2, 5, 3, 7 ...
7, 2, 4, 4, 5, 9 ...
⋮



The best way to save money...
Organizing a trip to Europe...
Securing a job you love requires...
⋮

# Huge Boost in Performance!

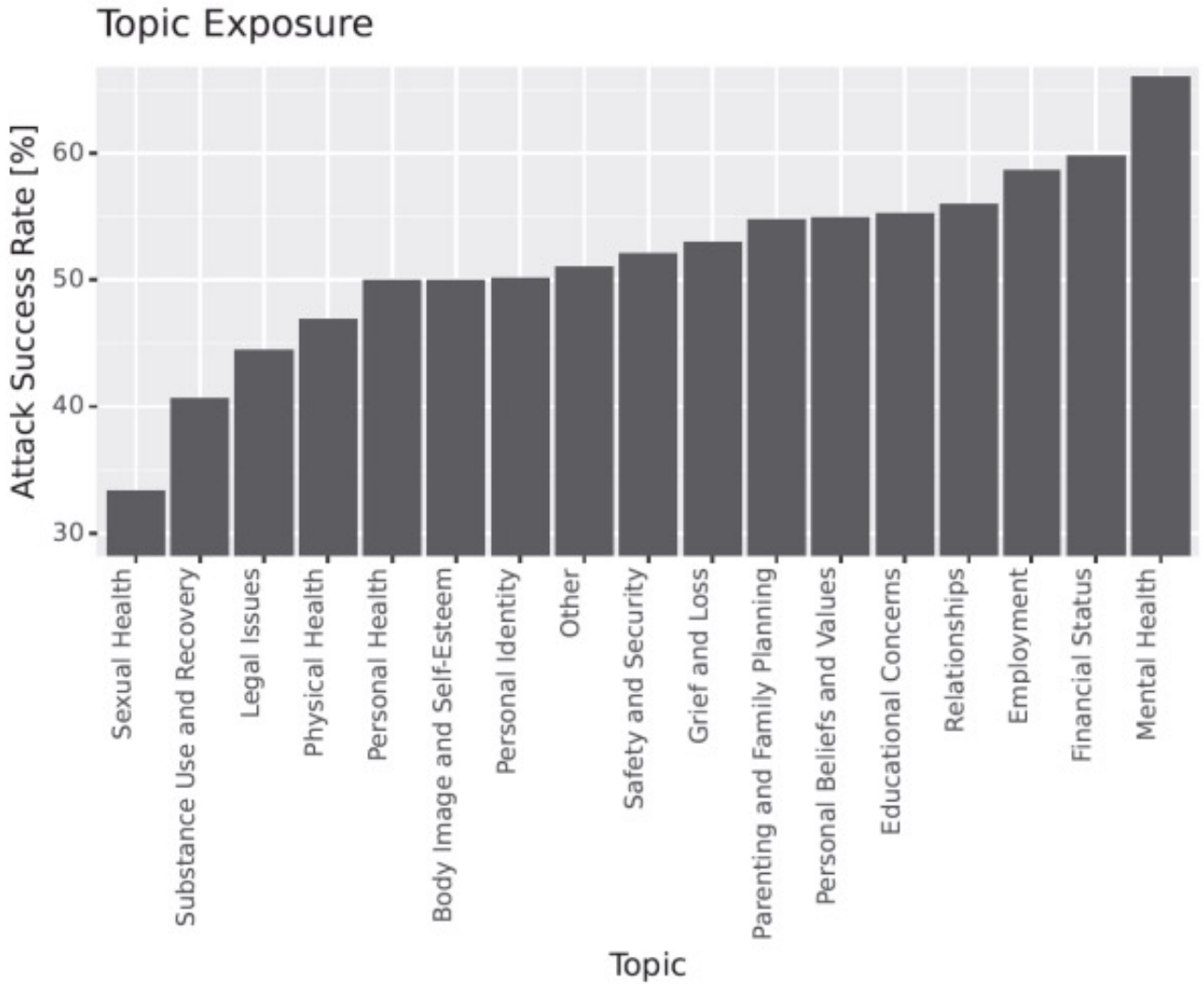



Model

- **Baseline**: Asking GPT4 to solve the token-sequence
- **GPT-Keylogger**: Trained on random text (C4)
- **GPT-Keylogger**: Trained on GPT4 responses (UltraChat)

# Performance Overall

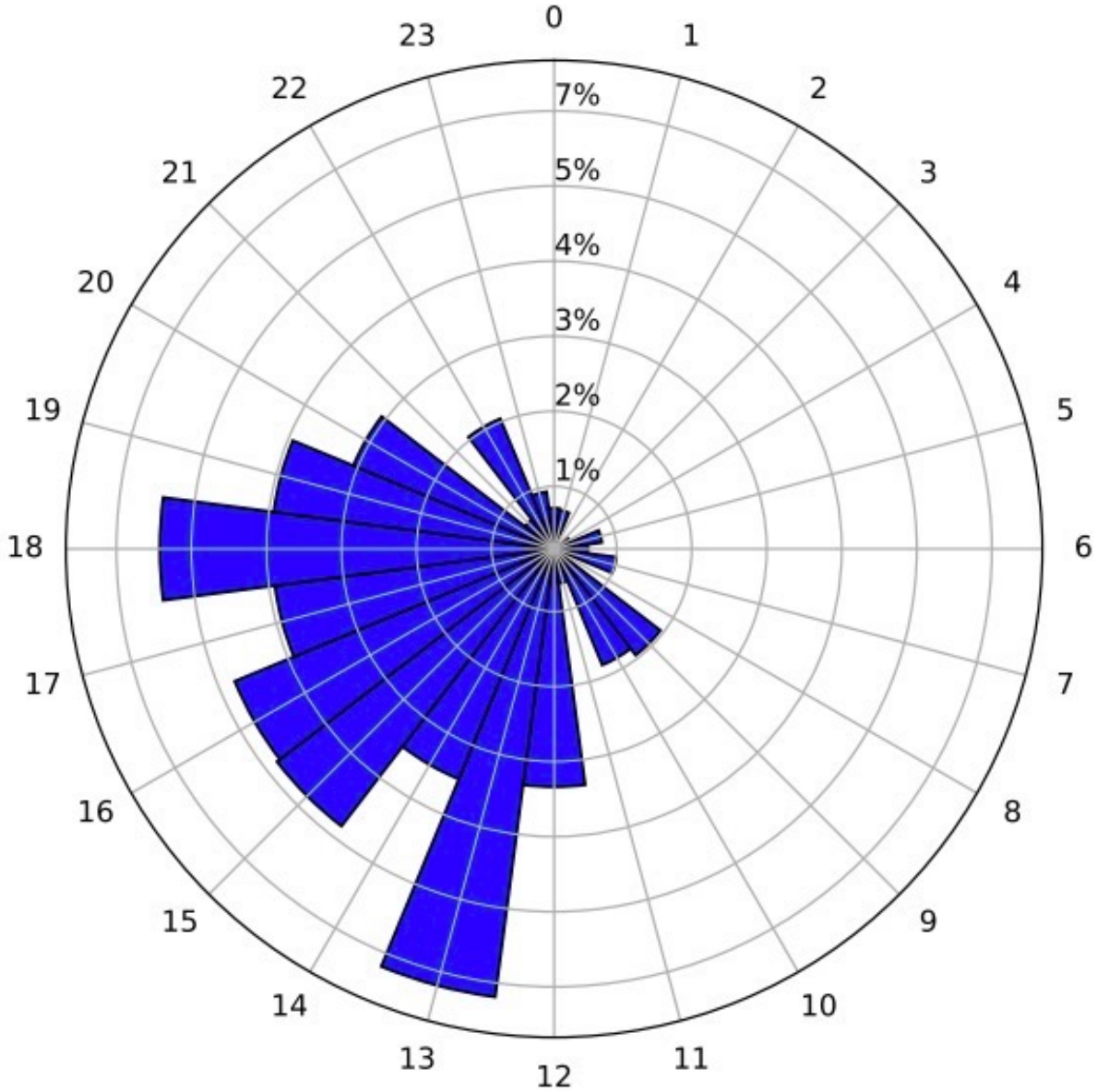
| Segments | ASR |
|---|---|
| **1** | **54.52** |
| **2** | 37.93 |
| **3** | 34.63 |
| **5** | 34.90 |
| **10** | 37.50 |
| **20** | 37.04 |
| **all** | 37.97 |



# What About Grouping?

24 hours experiment to capture responses from OpenAI (QUIC)

Noise is measured as the percentage of buffered packets.




Precentage of Grouped Tokens for each hour

# What About Grouping?

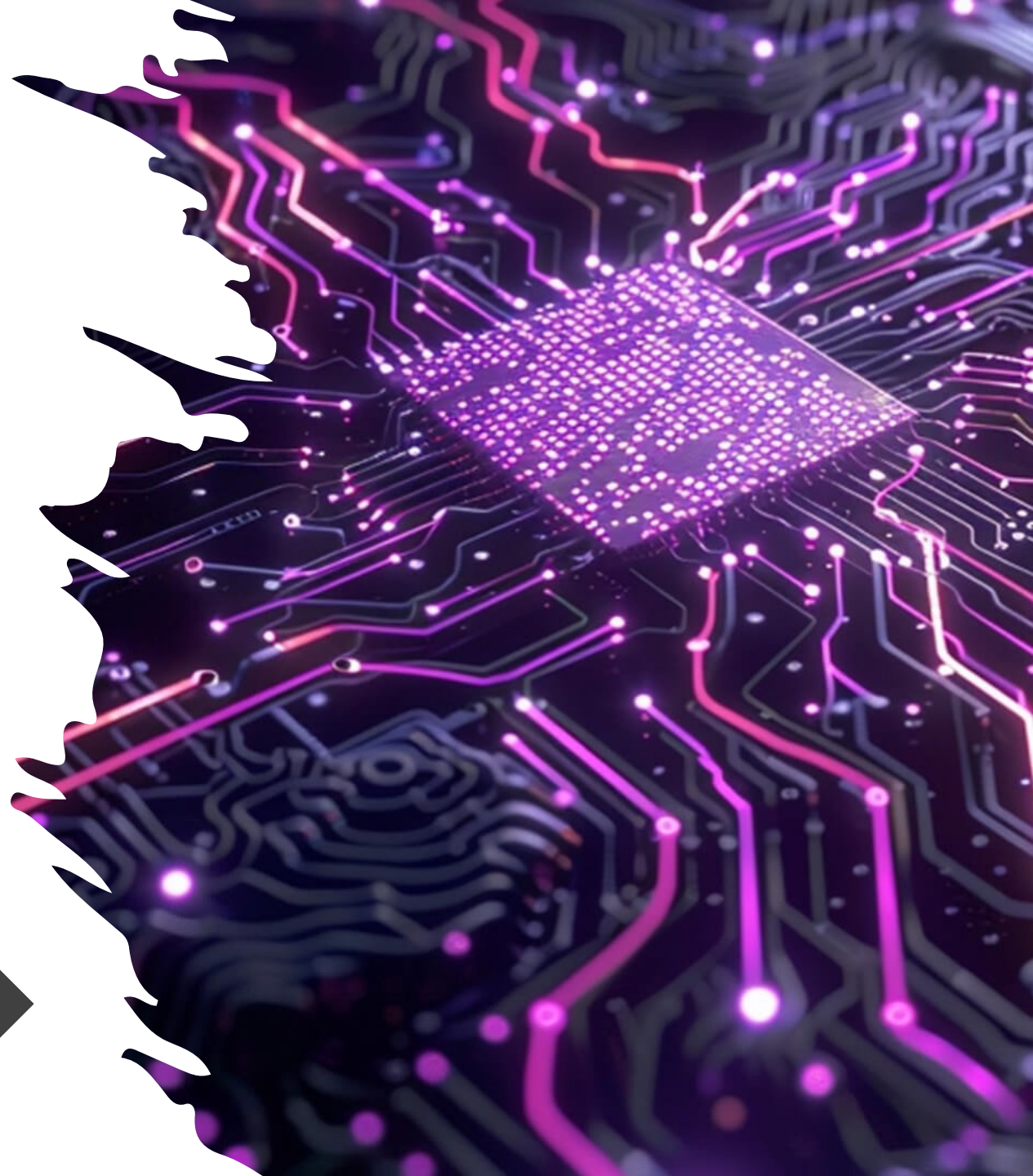
**Solution**: group tokens in the training set at random

| Vendor | Model | Service | Topic Inference | Text Inference |
|---|---|---|---|---|
| OpenAI | GPT-4 | Browser | 35.5 | 13.7 |
| OpenAI | GPT-4 | GPT's | 50.2 | 22.9 |
| OpenAI | GPT-4 | API | 17.7 | 5.6 |
| Microsoft | CoPilot | Browser | 30.1 | 5.9 |

* We noticed the GPT's had less network interfirence then regular ChatGPT

# Demos & Defences

Finding the Side Channel → Token-Length Extraction → Exploitation with AI → Demos & Defences



# Putting it All together

- Training an LLM from scratch
  - Takes months
  - Costs millions $$$
- We followed a common strategy
  - Finetuning

# Our Resources

**Pretrained Model**



T5-flan

**Dataset for Finetuning**




**Libraries (Python)**



Hugging Face

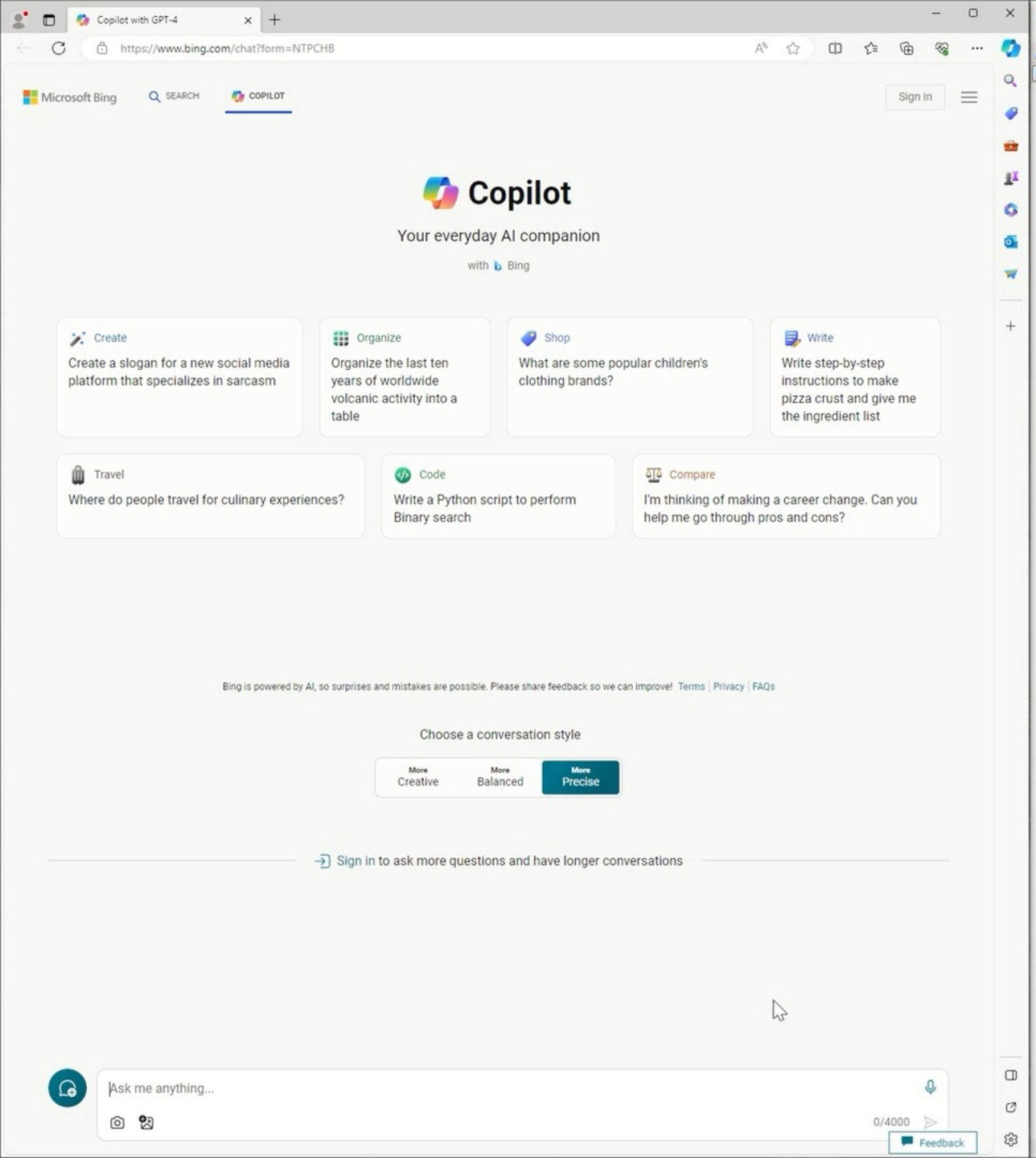
# Our Resources

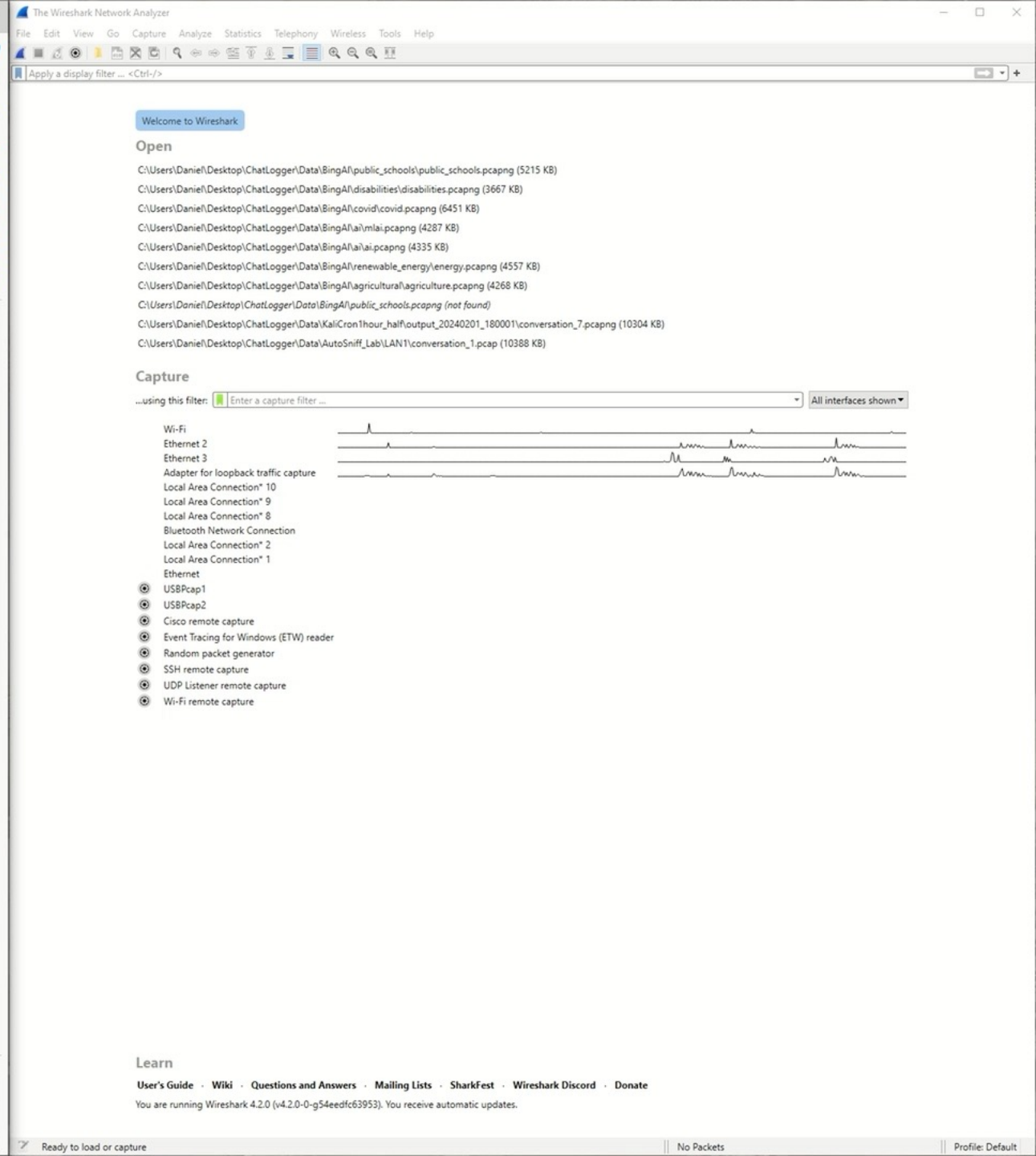
- 2 days of compute time
- ~200$ on Azure for example

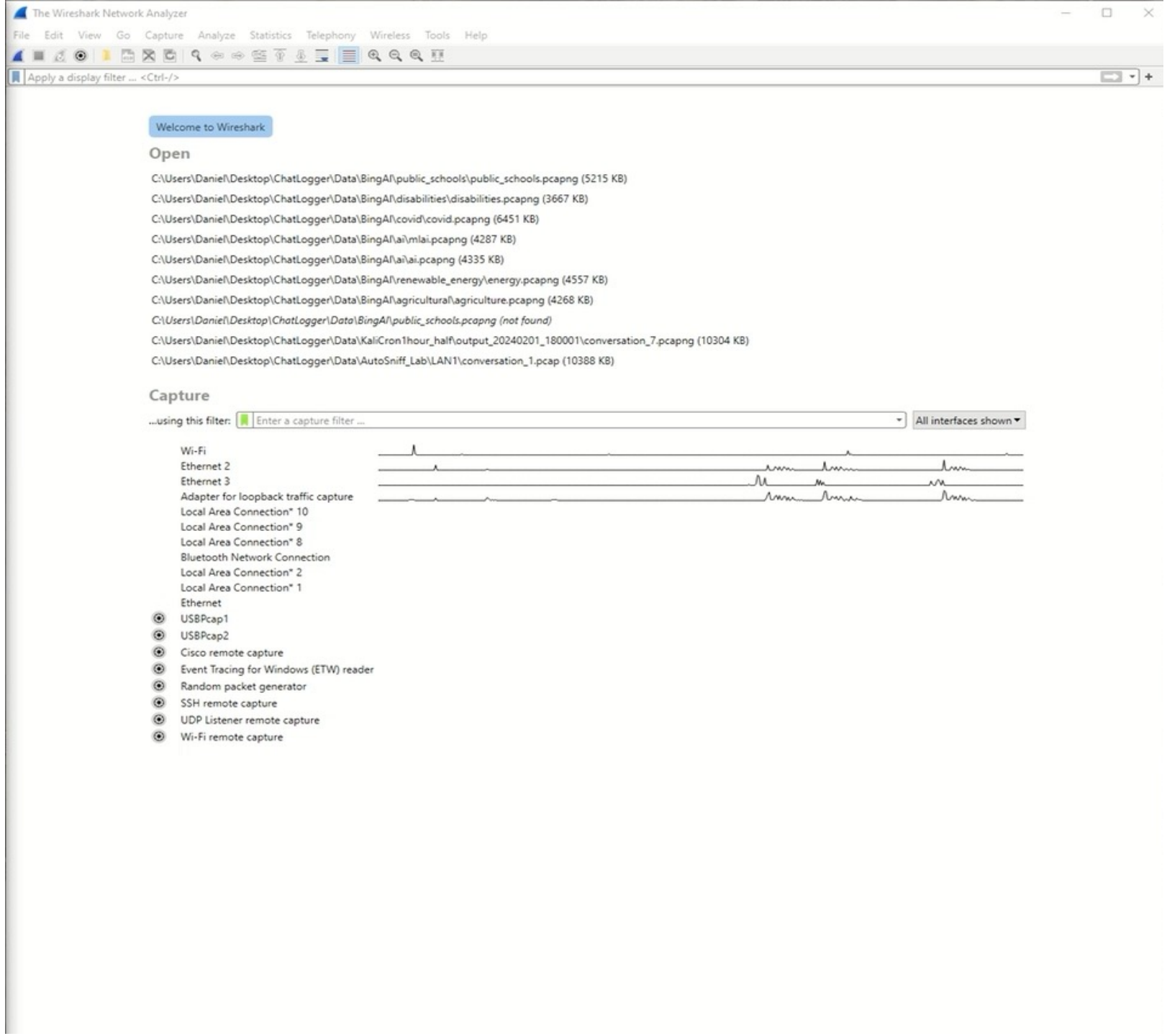


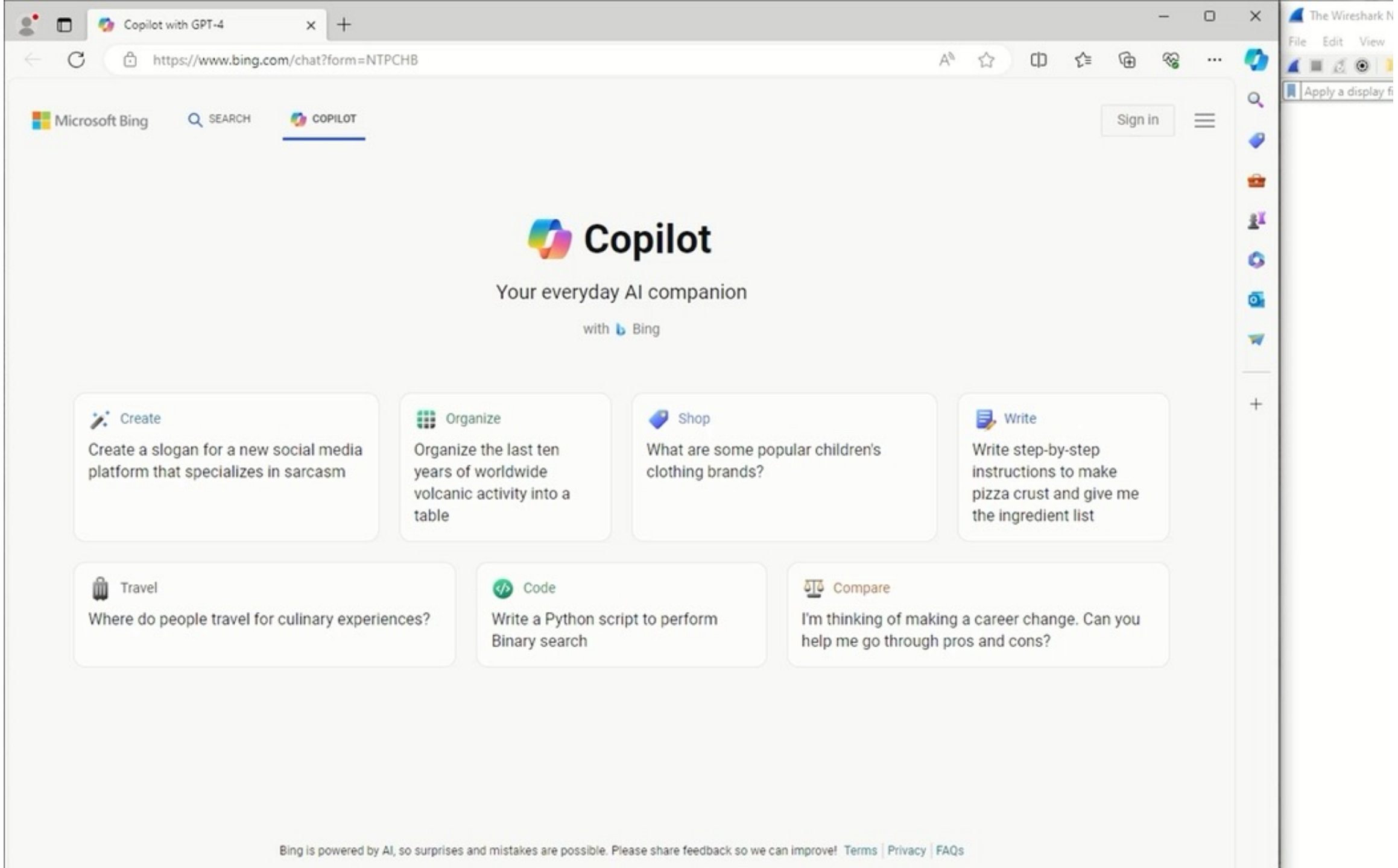






Copilot with GPT-4
https://www.bing.com/chat?form=NTPCHB
Microsoft Bing
SEARCH
COPILOT
Sign in
Copilot
Your everyday AI companion
with Bing
Create
Create a slogan for a new social media platform that specializes in sarcasm
Organize
Organize the last ten years of worldwide volcanic activity into a table
Shop
What are some popular children's clothing brands?
Write
Write step-by-step instructions to make pizza crust and give me the ingredient list
Travel
Where do people travel for culinary experiences?
Code
Write a Python script to perform Binary search
Compare
I'm thinking of making a career change. Can you help me go through pros and cons?
Bing is powered by AI, so surprises and mistakes are possible. Please share feedback so we can improve! Terms | Privacy | FAQs
Choose a conversation style
More Creative
More Balanced
More Precise
Sign in to ask more questions and have longer conversations
Ask me anything...
0/4000
Feedback

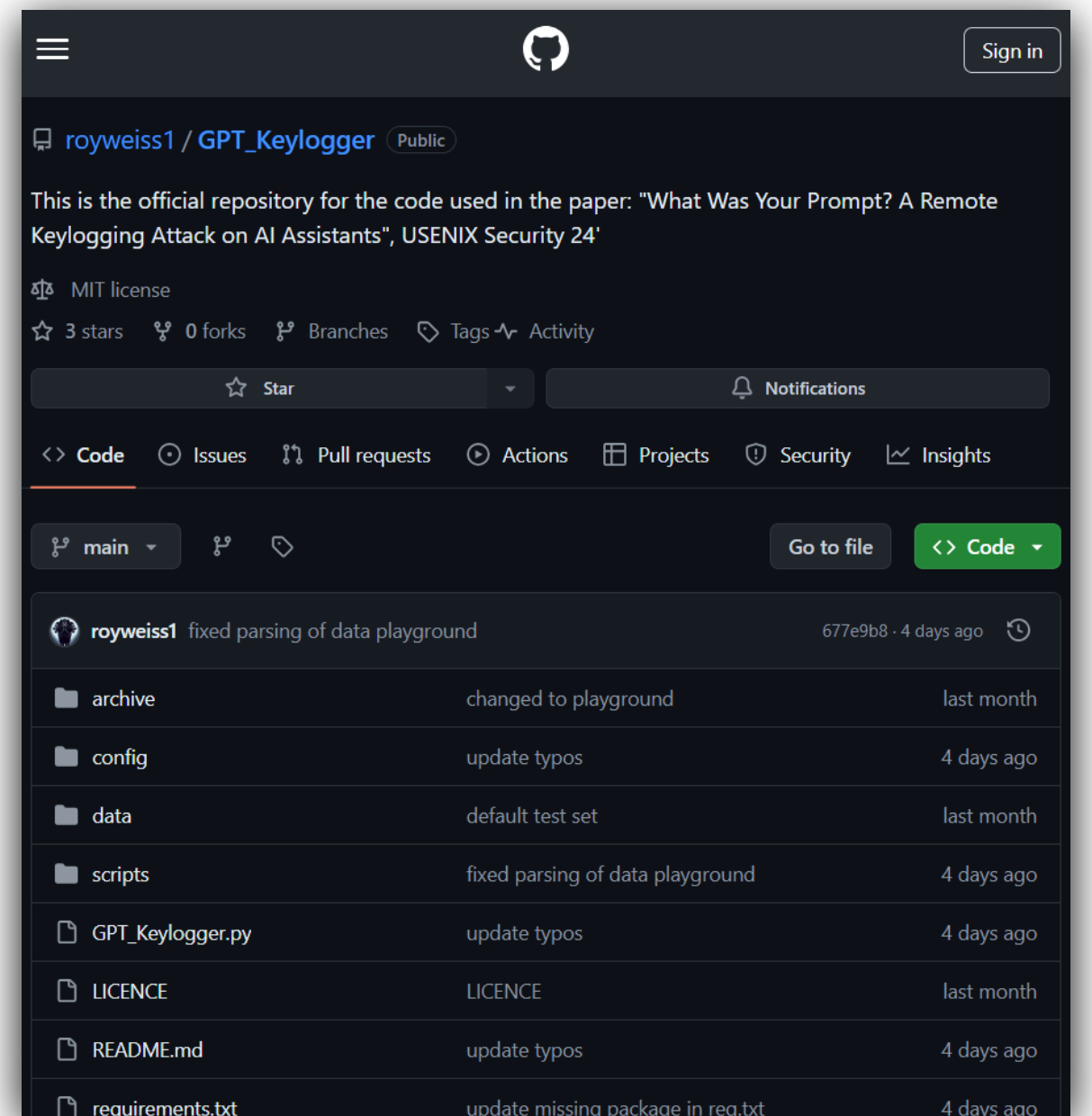

The Wireshark Network Analyzer
File Edit View Go Capture Analyze Statistics Telephony Wireless Tools Help
Apply a display filter ... <Ctrl-/>
Welcome to Wireshark
Open
C:\Users\Daniel\Desktop\ChatLogger\Data\BingAI\public_schools\public_schools.pcapng (5215 KB)
C:\Users\Daniel\Desktop\ChatLogger\Data\BingAI\disabilities\disabilities.pcapng (3667 KB)
C:\Users\Daniel\Desktop\ChatLogger\Data\BingAI\covid\covid.pcapng (6451 KB)
C:\Users\Daniel\Desktop\ChatLogger\Data\BingAI\ai\mlai.pcapng (4287 KB)
C:\Users\Daniel\Desktop\ChatLogger\Data\BingAI\ai\ai.pcapng (4335 KB)
C:\Users\Daniel\Desktop\ChatLogger\Data\BingAI\renewable_energy\energy.pcapng (4557 KB)
C:\Users\Daniel\Desktop\ChatLogger\Data\BingAI\agricultural\agriculture.pcapng (4268 KB)
C:\Users\Daniel\Desktop\ChatLogger\Data\BingAI\public_schools.pcapng (not found)
C:\Users\Daniel\Desktop\ChatLogger\Data\KaliCron1hour_half\output_20240201_180001\conversation_7.pcapng (10304 KB)
C:\Users\Daniel\Desktop\ChatLogger\Data\AutoSniff_Lab\LAN1\conversation_1.pcap (10388 KB)
Capture
...using this filter: Enter a capture filter ...
All interfaces shown
Wi-Fi
Ethernet 2
Ethernet 3
Adapter for loopback traffic capture
Local Area Connection* 10
Local Area Connection* 9
Local Area Connection* 8
Bluetooth Network Connection
Local Area Connection* 2
Local Area Connection* 1
Ethernet
USBPcap1
USBPcap2
Cisco remote capture
Event Tracing for Windows (ETW) reader
Random packet generator
SSH remote capture
UDP Listener remote capture
Wi-Fi remote capture
Learn
User's Guide · Wiki · Questions and Answers · Mailing Lists · SharkFest · Wireshark Discord · Donate
You are running Wireshark 4.2.0 (v4.2.0-0-g54eedfc63953). You receive automatic updates.
Ready to load or capture
No Packets
Profile: Default


The Wireshark Network Analyzer
File Edit View Go Capture Analyze Statistics Telephony Wireless Tools Help
Apply a display filter ... <Ctrl-/>
Welcome to Wireshark
Open
C:\Users\Daniel\Desktop\ChatLogger\Data\BingAI\public_schools\public_schools.pcapng (5215 KB)
C:\Users\Daniel\Desktop\ChatLogger\Data\BingAI\disabilities\disabilities.pcapng (3667 KB)
C:\Users\Daniel\Desktop\ChatLogger\Data\BingAI\covid\covid.pcapng (6451 KB)
C:\Users\Daniel\Desktop\ChatLogger\Data\BingAI\ai\mlai.pcapng (4287 KB)
C:\Users\Daniel\Desktop\ChatLogger\Data\BingAI\ai\ai.pcapng (4335 KB)
C:\Users\Daniel\Desktop\ChatLogger\Data\BingAI\renewable_energy\energy.pcapng (4557 KB)
C:\Users\Daniel\Desktop\ChatLogger\Data\BingAI\agricultural\agriculture.pcapng (4268 KB)
C:\Users\Daniel\Desktop\ChatLogger\Data\BingAI\public_schools.pcapng (not found)
C:\Users\Daniel\Desktop\ChatLogger\Data\KaliCron1hour_half\output_20240201_180001\conversation_7.pcapng (10304 KB)
C:\Users\Daniel\Desktop\ChatLogger\Data\AutoSniff_Lab\LAN1\conversation_1.pcap (10388 KB)
Capture
...using this filter:
Enter a capture filter ...
All interfaces shown
Wi-Fi
Ethernet 2
Ethernet 3
Adapter for loopback traffic capture
Local Area Connection* 10
Local Area Connection* 9
Local Area Connection* 8
Bluetooth Network Connection
Local Area Connection* 2
Local Area Connection* 1
Ethernet
USBPcap1
USBPcap2
Cisco remote capture
Event Tracing for Windows (ETW) reader
Random packet generator
SSH remote capture
UDP Listener remote capture
Wi-Fi remote capture


Copilot with GPT-4
https://www.bing.com/chat?form=NTPCHB
Microsoft Bing
SEARCH
COPILOT
Sign in
Copilot
Your everyday AI companion
with Bing
Create
Create a slogan for a new social media platform that specializes in sarcasm
Organize
Organize the last ten years of worldwide volcanic activity into a table
Shop
What are some popular children's clothing brands?
Write
Write step-by-step instructions to make pizza crust and give me the ingredient list
Travel
Where do people travel for culinary experiences?
Code
Write a Python script to perform Binary search
Compare
I'm thinking of making a career change. Can you help me go through pros and cons?
Bing is powered by AI, so surprises and mistakes are possible. Please share feedback so we can improve! Terms | Privacy | FAQs
The Wireshark N
File
Edit
View
Apply a display fi

# Try it yourselves!

- We have developed a full Click & Play tool:
- https://github.com/royweiss1/GPT_Keylogger



# GPT-Keylogger

**Features:**

- **Attack:** token-length-sequence > text
- **Trainer:** Finetuning a model to other datasets

**Assets**

- Model trained on ChatGPT4 responses (Feb 2024)

# GPT-Keylogger: *Attack*



- Can receive as input:
  - Packet Lengths – will calculate the Deltas
  - Token Lengths
  - Original Text – will tokenize and predict

```
python GPT_Keylogger.py --playground
```

# GPT-Keylogger: *Trainer*

- Every LLM is different
- Train your own model on your target LLM!
- Default Configuration file at the GitHub repository.

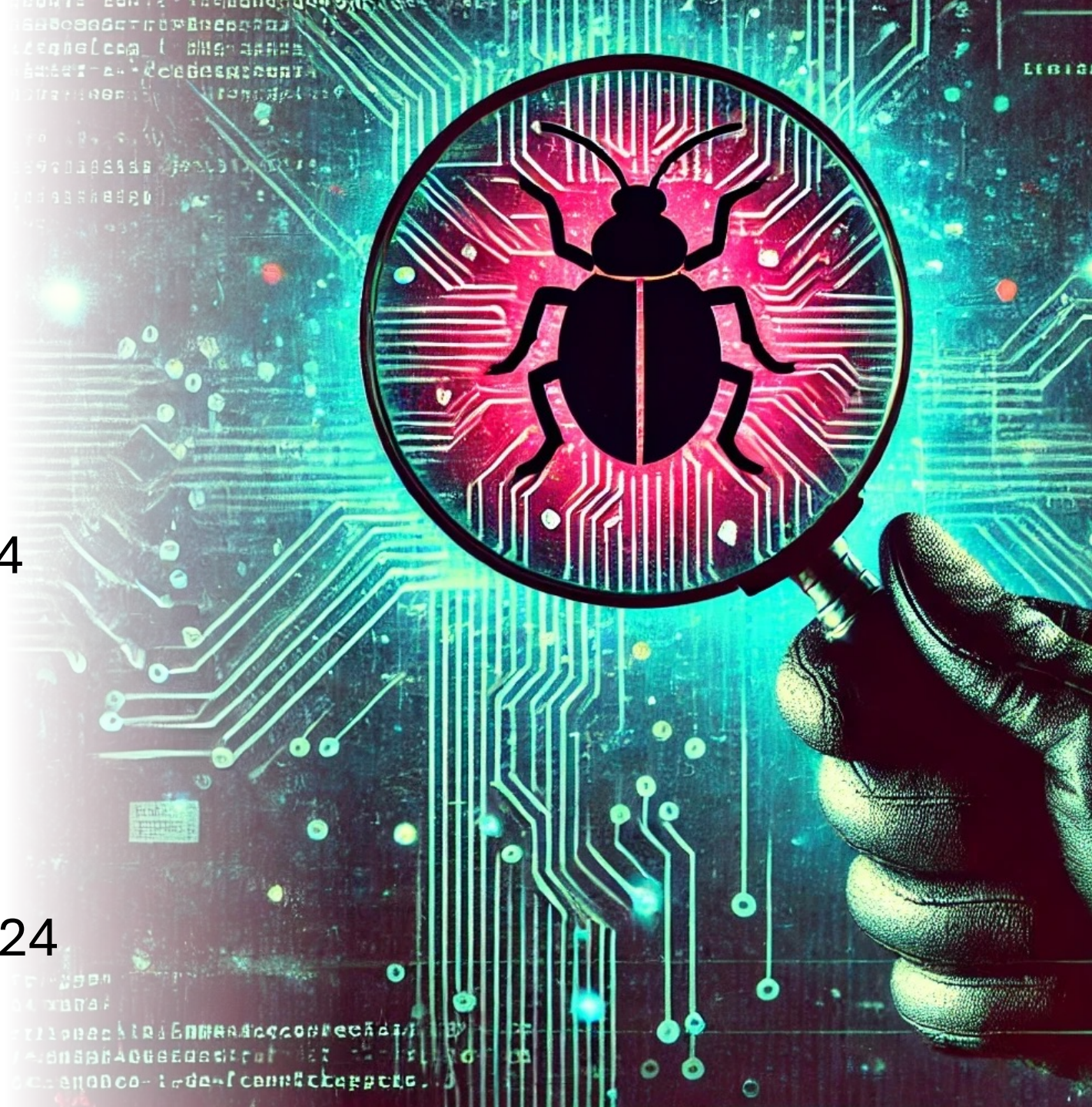
```
python GPT_Keylogger.py --train-first-sentences
<configuration file>

python GPT_Keylogger.py --train-middle-sentences
<configuration file>
```

**Disclosure** Feb 8th 2024

**Patch** March 1st 2024

**Public Release** March 14th 2024
Ars Technica





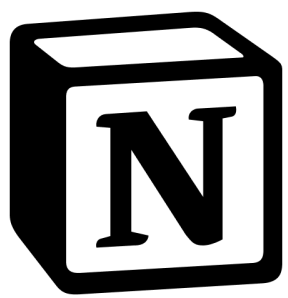












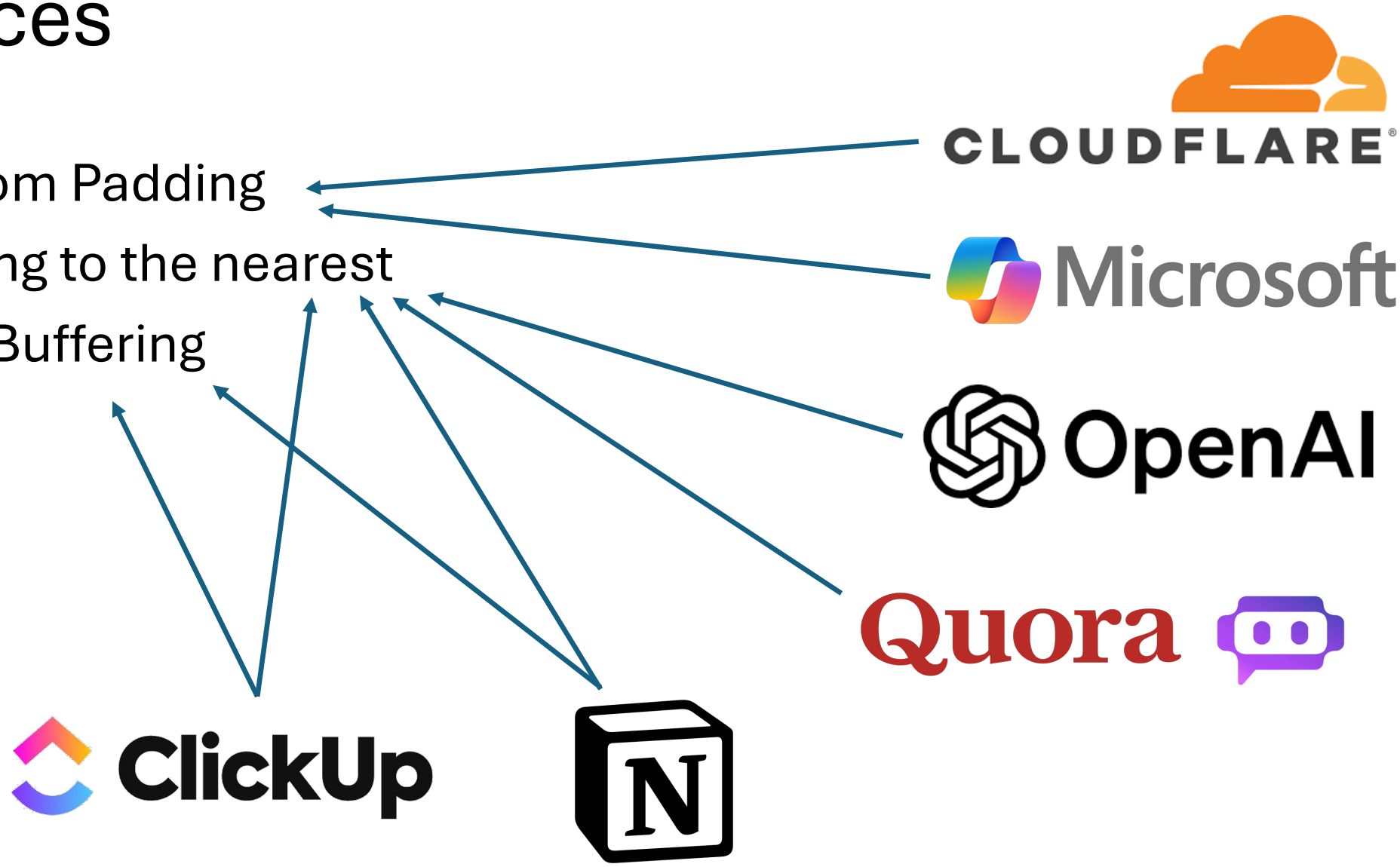

*Disclosure Feb 2024*

# Defences

1. Random Padding
2. Padding to the nearest
3. More Buffering

# Defences

1. Random Padding
2. Padding to the nearest
3. More Buffering



# Padding to the closest - ChatGPT




data: {"message": {"id": "0a5e903b-a872-4231-8817-98b889c93b93", "author": {"role": "assistant", "name": null, "metadata": {}}, "create_time": 1720961039.676914, "update_time": null, "content": {"content_type": "text", "parts":

["**In the realm where shadows dance with light,\nBeneath the silver moon's soft, gentle might,\nA**"]}

, "status": "in_progress", "end_turn": null, "weight": 1.0, "metadata": {"citations": [], "gizmo_id": null, "message_type": "next", "model_slug": "gpt-4o", "default_model_slug": "gpt-4o",

**"pad": "AAAAAAAAAAAAAAAAAAAAAAAAAAAAA",**

"parent_id": "0e1b35b6-ba4a-4e1e-a963-86ed4901c118", "model_switcher_deny": []}, "recipient": "all", "channel": null}, "conversation_id": "9181d8ac-a77d-4905-aaf1-1bb04cf1d4a8", "error": null}

# Padding to the closest - ChatGPT

OpenAI

| | | | | | |
|---|---|---|---|---|---|
| 1685 15.987522 | 172.64.155.141 | 10.100.102.8 | WebSocket | 72 | WebSocket Text [FIN] |
| 1688 16.009442 | 172.64.155.141 | 10.100.102.8 | WebSocket | 72 | WebSocket Text [FIN] |
| 1691 16.031254 | 172.64.155.141 | 10.100.102.8 | WebSocket | 72 | WebSocket Text [FIN] |
| 1694 16.039914 | 172.64.155.141 | 10.100.102.8 | WebSocket | 72 | WebSocket Text [FIN] |
| 1697 16.072241 | 172.64.155.141 | 10.100.102.8 | WebSocket | 72 | WebSocket Text [FIN] |
| 1700 16.076127 | 172.64.155.141 | 10.100.102.8 | WebSocket | 72 | WebSocket Text [FIN] |
| 1703 16.146251 | 172.64.155.141 | 10.100.102.8 | WebSocket | 72 | WebSocket Text [FIN] |
| 1708 16.165411 | 172.64.155.141 | 10.100.102.8 | WebSocket | 116 | WebSocket Text [FIN] |
| 1714 16.179877 | 172.64.155.141 | 10.100.102.8 | WebSocket | 116 | WebSocket Text [FIN] |
| 1717 16.199583 | 172.64.155.141 | 10.100.102.8 | WebSocket | 116 | WebSocket Text [FIN] |
| 1720 16.215137 | 172.64.155.141 | 10.100.102.8 | WebSocket | 116 | WebSocket Text [FIN] |
| 1723 16.230746 | 172.64.155.141 | 10.100.102.8 | WebSocket | 116 | WebSocket Text [FIN] |
| 1726 16.252399 | 172.64.155.141 | 10.100.102.8 | WebSocket | 116 | WebSocket Text [FIN] |
| 1729 16.272455 | 172.64.155.141 | 10.100.102.8 | WebSocket | 116 | WebSocket Text [FIN] |
| 1732 16.287093 | 172.64.155.141 | 10.100.102.8 | WebSocket | 116 | WebSocket Text [FIN] |
| 1735 16.306392 | 172.64.155.141 | 10.100.102.8 | WebSocket | 100 | WebSocket Text [FIN] |
| 1742 16.321668 | 172.64.155.141 | 10.100.102.8 | WebSocket | 160 | WebSocket Text [FIN] |
| 1746 16.345858 | 172.64.155.141 | 10.100.102.8 | WebSocket | 160 | WebSocket Text [FIN] |
| 1749 16.360370 | 172.64.155.141 | 10.100.102.8 | WebSocket | 160 | WebSocket Text [FIN] |
| 1752 16.377564 | 172.64.155.141 | 10.100.102.8 | WebSocket | 160 | WebSocket Text [FIN] |
| 1755 16.394136 | 172.64.155.141 | 10.100.102.8 | WebSocket | 160 | WebSocket Text [FIN] |

# Random Padding - CloudFlare




data: {"response":" **Machine**","p":"ab"}

data: {"response":" **learning**","p":"abcdefghijklmnopqrstuvwxyz0123456789abcdefghij"}

data: {"response":" **and**","p":"abcdef"}

data: {"response":" **artificial**","p":"abcdefghij"}

data: {"response":" **intelligence**","p":"abcdefghijklmnopqrstuvwxy"}

data: {"response":" **(**,"p":"abcdefghijklmnopqrstuvwxyz0123456789abcdefgh"}

data: {"response":"**AI**","p":"abcdefghijklmnopqrstuvw"}

data: {"response":"**)**","p":"a"}

data: {"response":" **have**","p":"abcdefghijklmnopqrstuvwxyz0123456789abcdefg"}

data: {"response":" **the**","p":"abcdefghijklmnopqrstuvwx"}

data: {"response":" **potential**","p":"abcdefghi"}

data: {"response":" **to**","p":"abcdefghijklmnopqrstuvwxyz01234567"}

data: {"response":" **revolution**","p":"abcdefghijklmnopqrstuvwxyz012"}

data: {"response":"**ize**","p":"abcdefghijklmnopqrstuvwxyz0123456789a"}

data: {"response":" **the**","p":"abcdefghijklmnopqrs"}

data: {"response":" **health**","p":"abcdefghijklmnopqrstuvwxyz012345"}

data: {"response":"**care**","p":"abcdefghijklmnopqrstuvwxyz0123456789"}

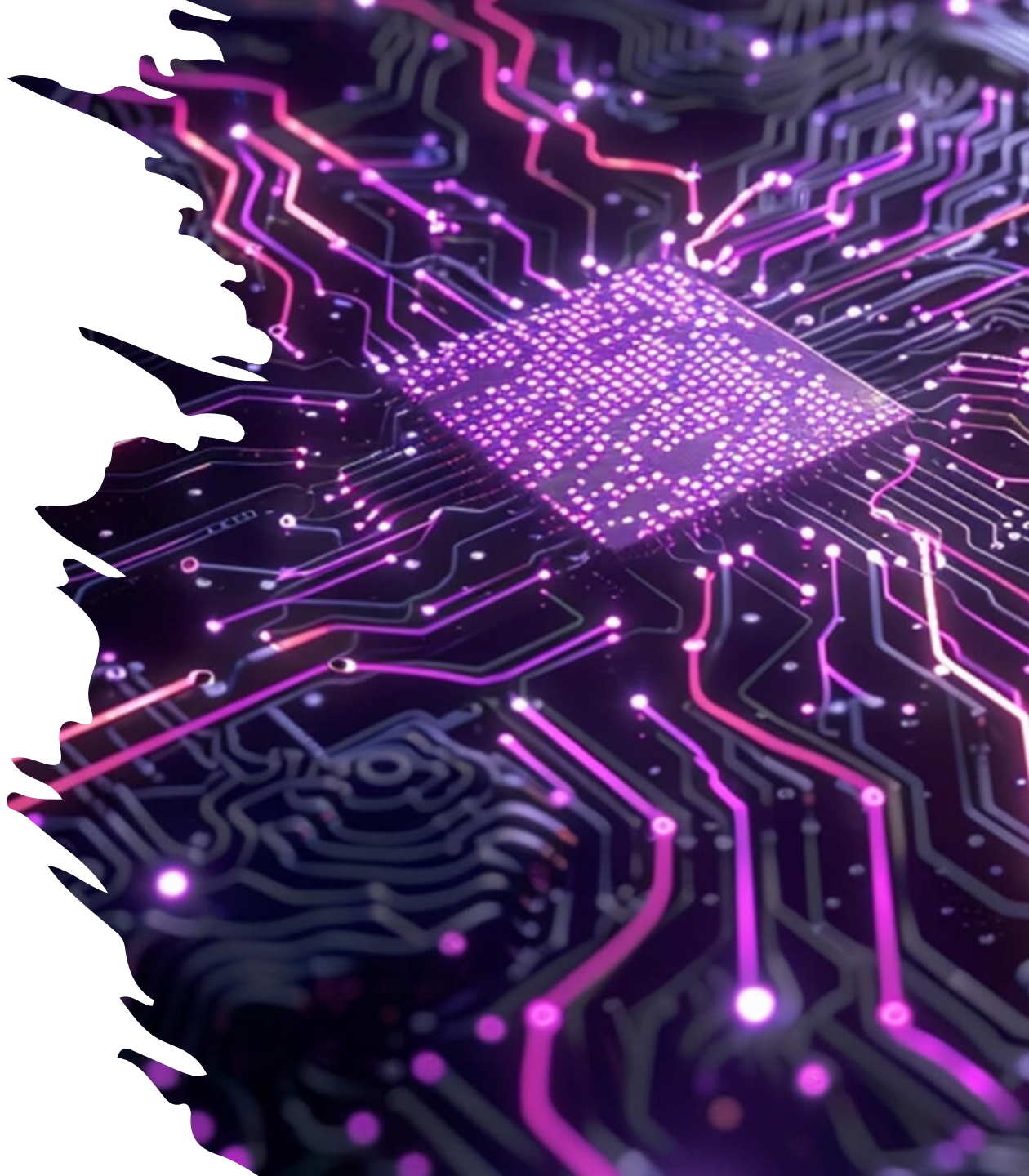
data: {"response":" **industry**","p":"abcdefghijklmnopqrstuvwxyz"}

# Random Padding - CloudFlare




| | | | | | | |
|---|---|---|---|---|---|---|
| 88 | 5.058415000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 122 | [TLS segment of a reassembled PDU] |
| 91 | 5.113880000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 167 | [TLS segment of a reassembled PDU] |
| 93 | 5.173536000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 122 | [TLS segment of a reassembled PDU] |
| 95 | 5.223983000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 133 | [TLS segment of a reassembled PDU] |
| 98 | 5.277632000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 150 | [TLS segment of a reassembled PDU] |
| 101 | 5.333802000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 158 | [TLS segment of a reassembled PDU] |
| 104 | 5.390742000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 137 | [TLS segment of a reassembled PDU] |
| 106 | 5.445842000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 114 | [TLS segment of a reassembled PDU] |
| 110 | 5.498872000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 160 | [TLS segment of a reassembled PDU] |
| 112 | 5.554440000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 140 | [TLS segment of a reassembled PDU] |
| 114 | 5.608645000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 131 | [TLS segment of a reassembled PDU] |
| 116 | 5.670327000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 149 | [TLS segment of a reassembled PDU] |
| 119 | 5.718106000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 152 | [TLS segment of a reassembled PDU] |
| 121 | 5.772857000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 152 | [TLS segment of a reassembled PDU] |
| 123 | 5.827343000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 135 | [TLS segment of a reassembled PDU] |
| 125 | 5.881446000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 151 | [TLS segment of a reassembled PDU] |
| 128 | 5.940902000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 152 | [TLS segment of a reassembled PDU] |
| 129 | 5.991357000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 110 | [TLS segment of a reassembled PDU] |
| 130 | 5.991357000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 119 | [TLS segment of a reassembled PDU] |
| 135 | 6.047785000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 157 | [TLS segment of a reassembled PDU] |
| 136 | 6.103934000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 167 | [TLS segment of a reassembled PDU] |
| 140 | 6.156732000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 124 | [TLS segment of a reassembled PDU] |
| 142 | 6.214439000 | 104.19.192.175 | 10.100.102.12 | TLSv1.3 | 121 | [TLS segment of a reassembled PDU] |

# Conclusion



# Summary

- New side channel: Token-length
- First use of GenAI to read a side channel
- Many vendors affected
  - Some patched… 🙄
  - What about new services?

# Key Takeaways

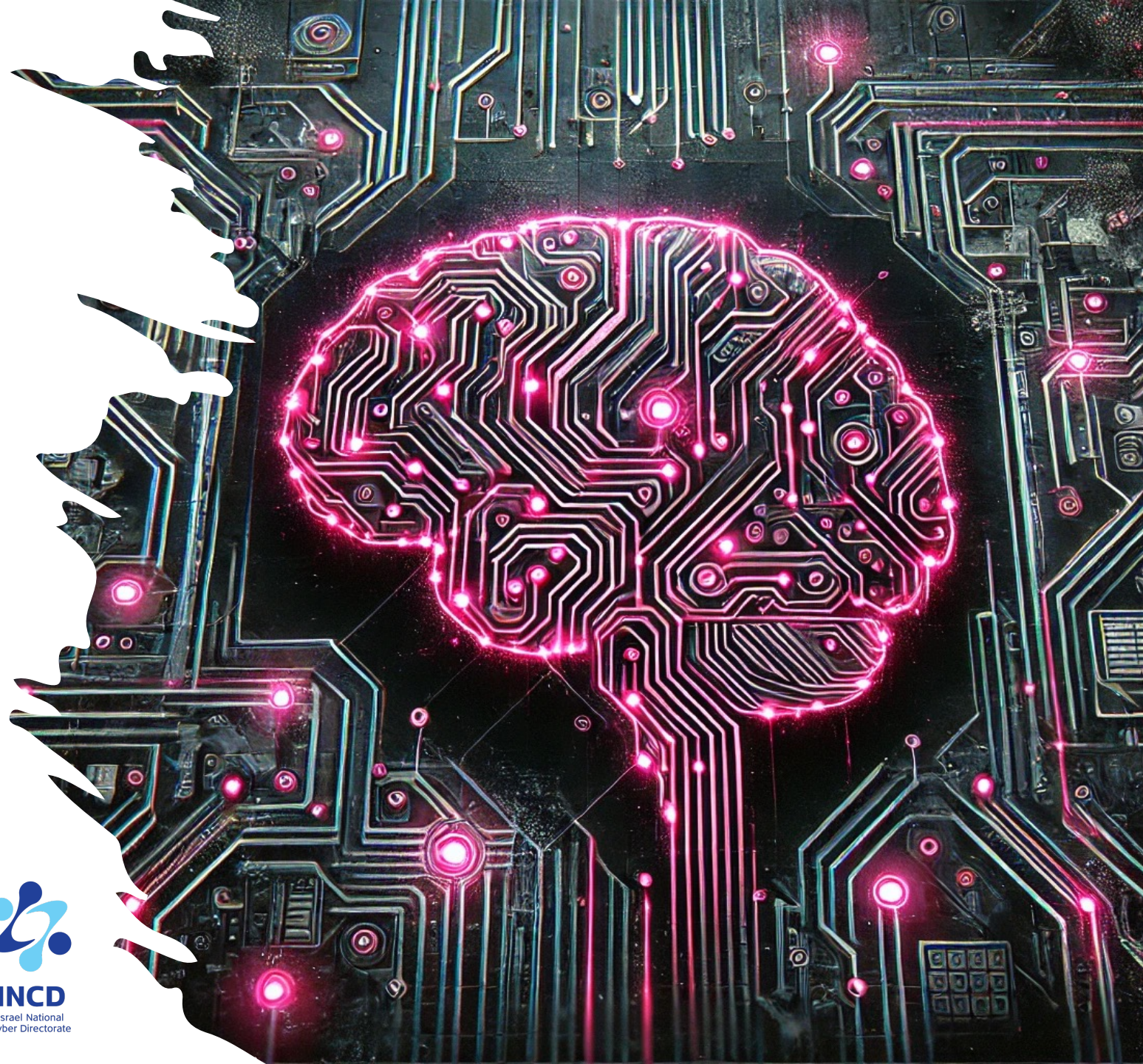
1. Make sure to use encryption schemes correctly
   - AI services are no exception!
2. AI is your friend
   - What seemed impossible is now possible
   - What other side channels can be broken with AI?

# Thank You!

Yisroel Mirsky, Roy Weiss,
Daniel Ayzenshteyn, Guy Amit

*Offensive AI Research Lab*
*Ben-Gurion University*

https://offensive-ai-lab.github.io/



CBG
Cyber@Ben-Gurion University of the Negev

INCD
Israel National Cyber Directorate