Oversharing with LLMs is underrated: Personal Disclosures in Human-ChatGPT Conversations

"Honey, why does the toaster know it's my birthday tomorrow?"

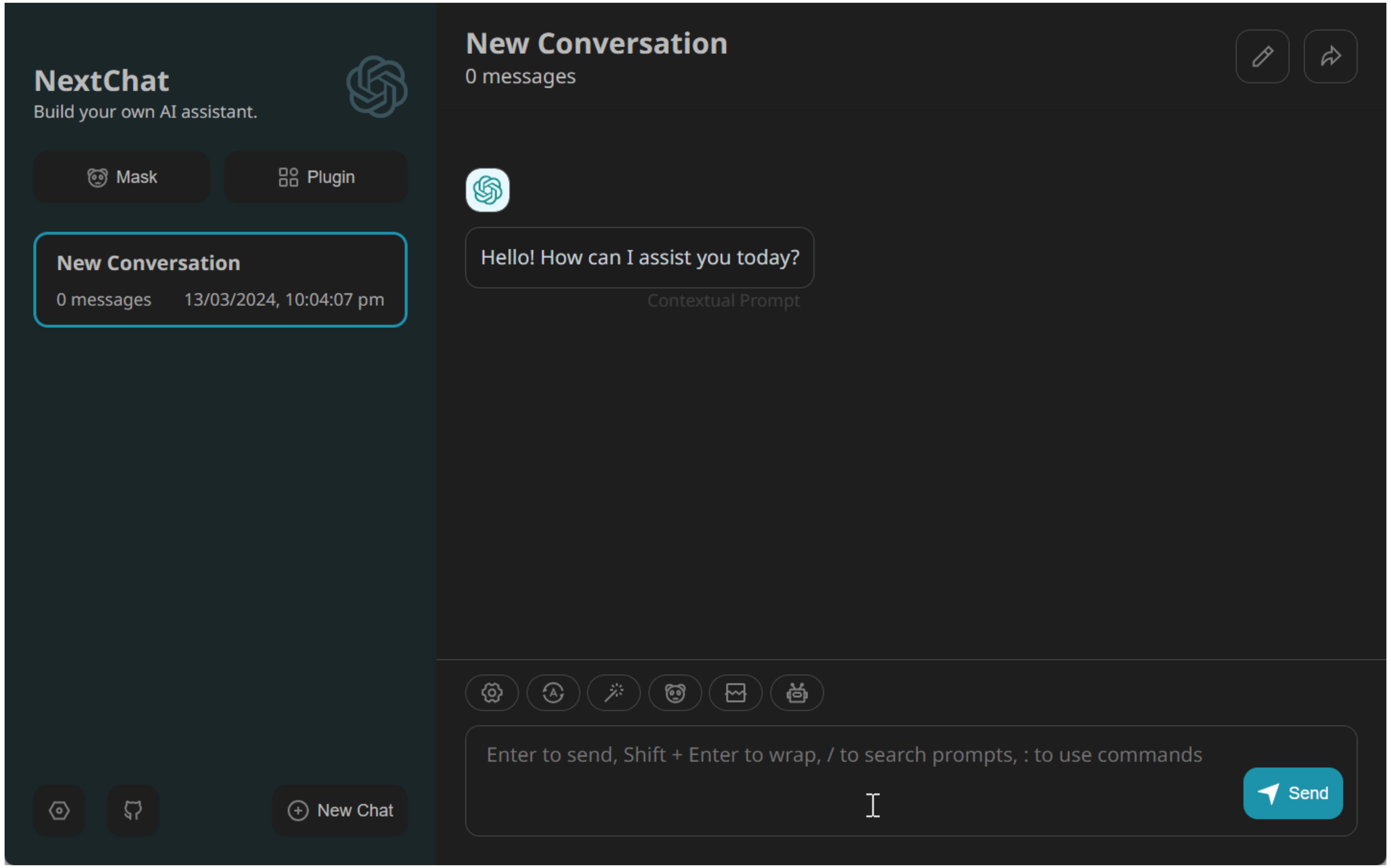
Niloofar Mireshghallah

niloofar@cs.washington.edu

X: @niloofar_mire

What are LLMs?

Large Language Models

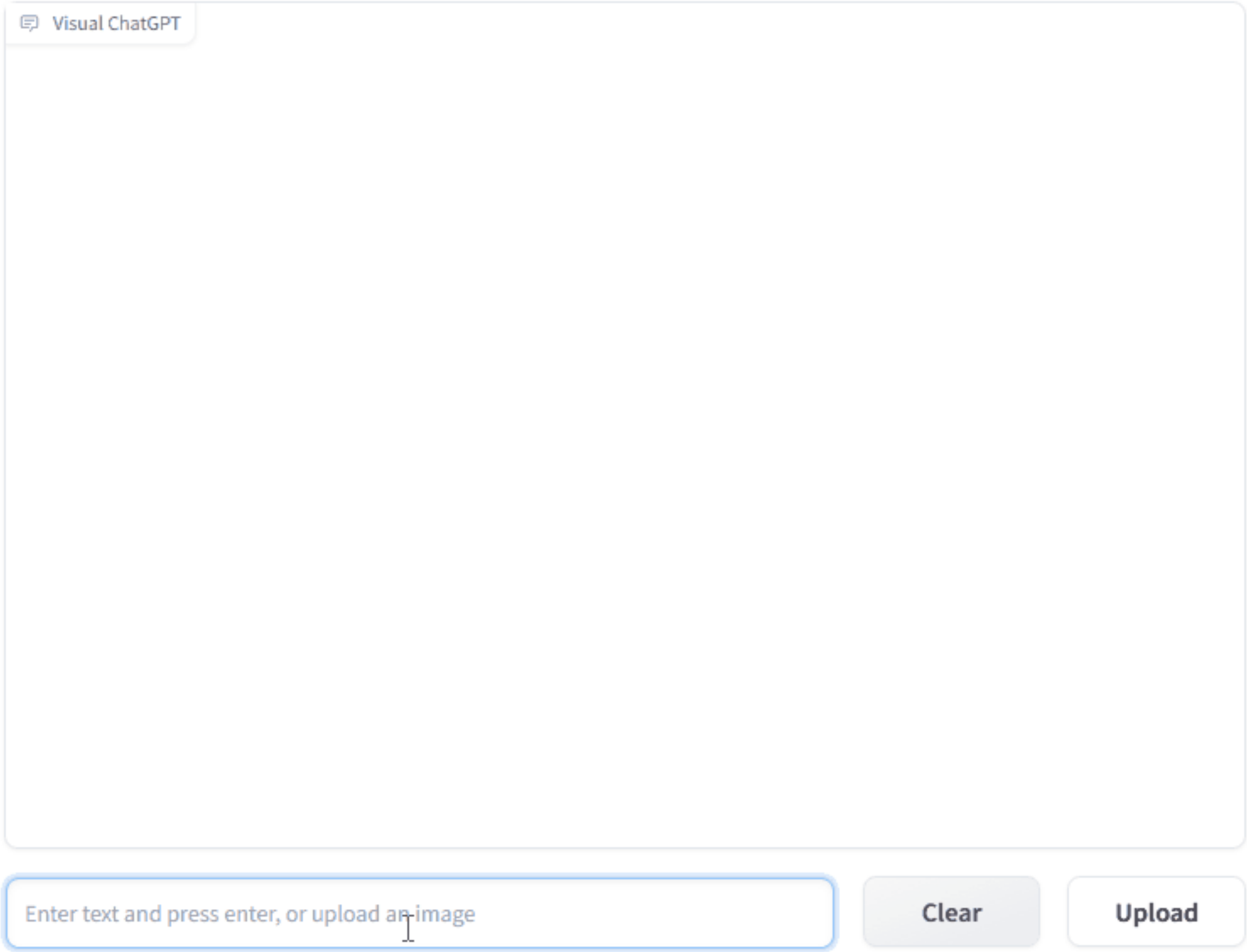


Hello! How can I assist you today?

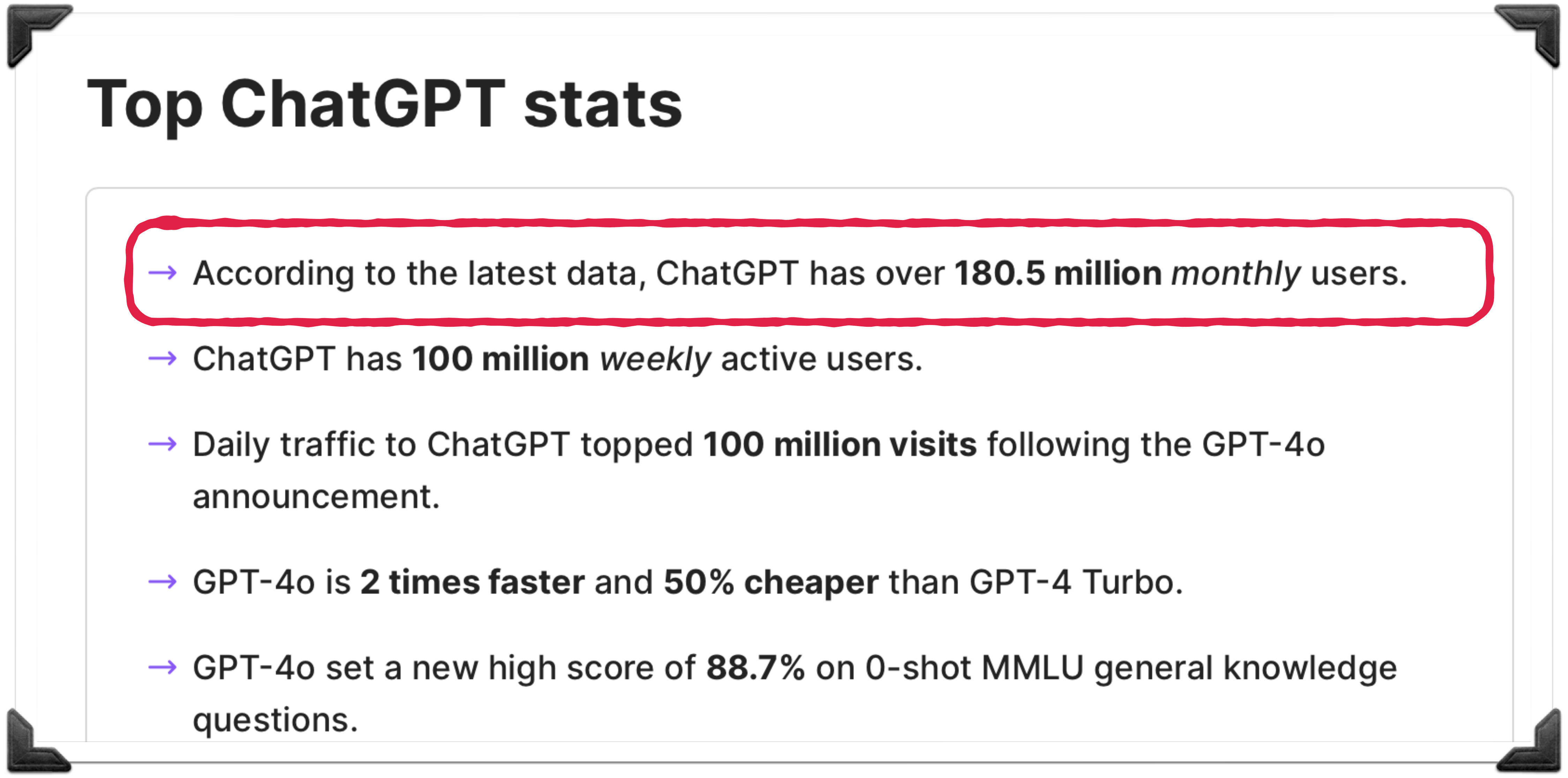
Enter to send, Shift + Enter to wrap, / to search prompts, : to use commands

LLMs can have visual components

Multimodal Models



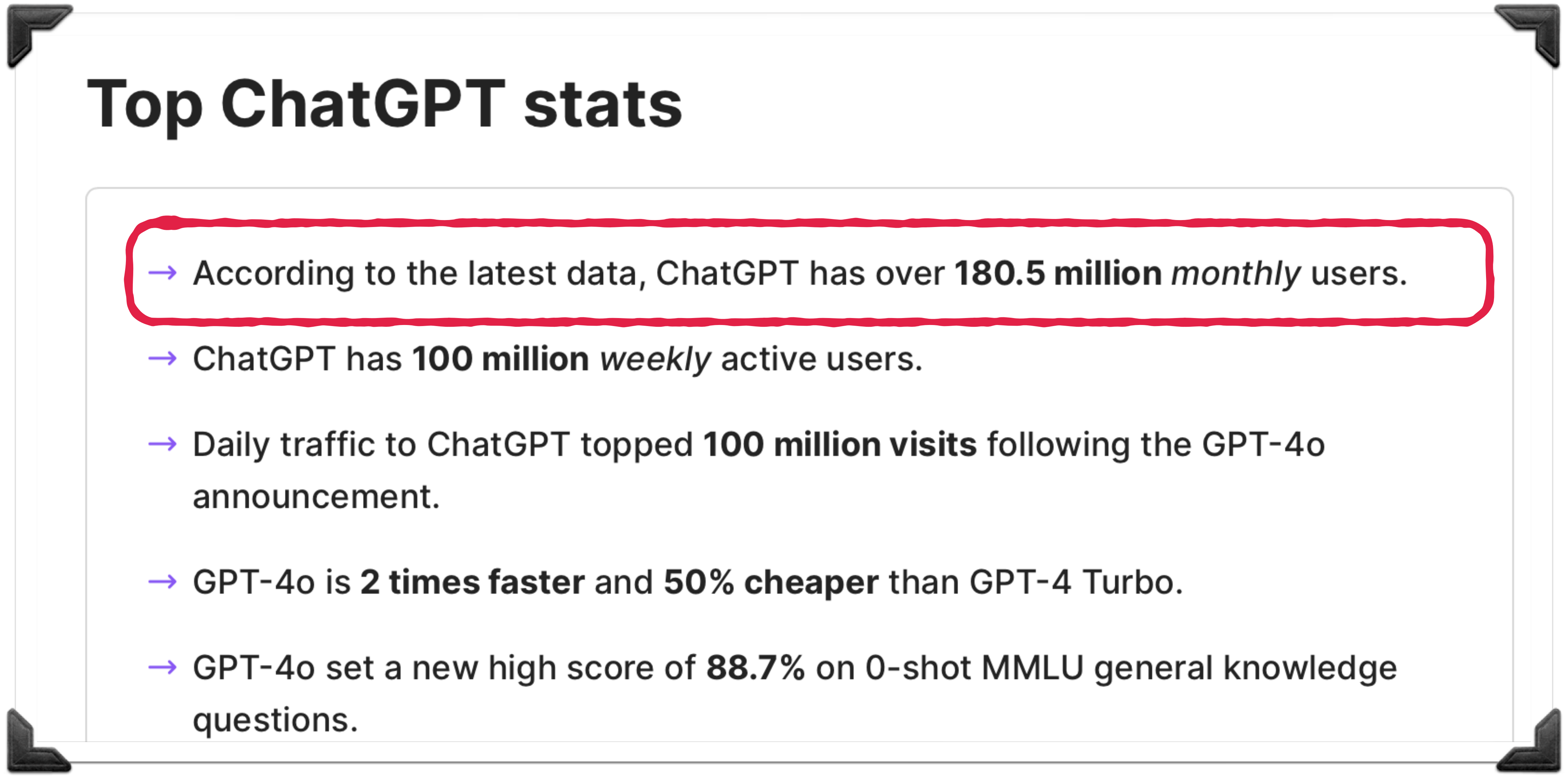
How many people use LLMs?



Top ChatGPT stats

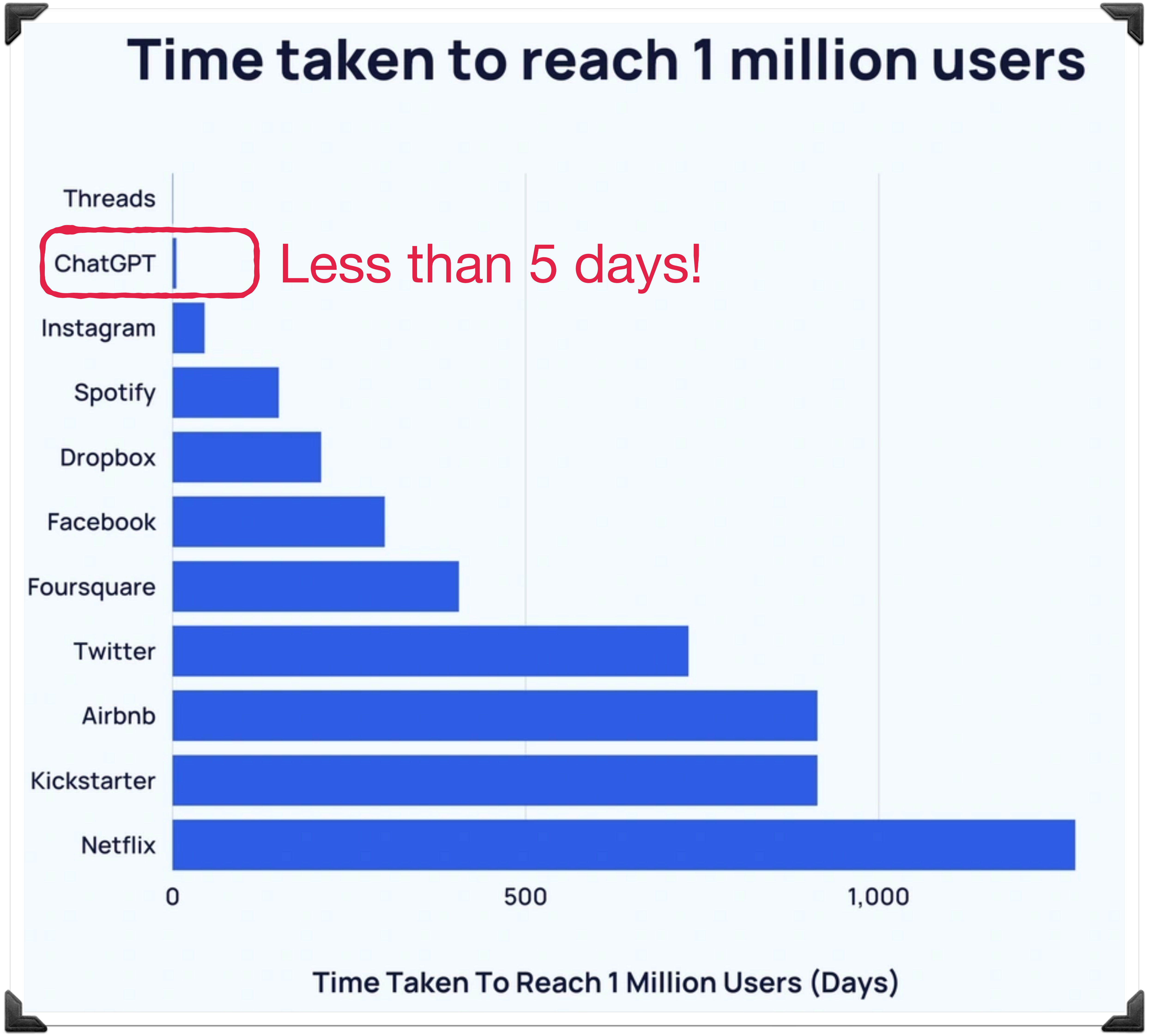
- → According to the latest data, ChatGPT has over **180.5 million** *monthly* users.
- → ChatGPT has **100 million** *weekly* active users.
- → Daily traffic to ChatGPT topped **100 million visits** following the GPT-4o announcement.
- → GPT-4o is **2 times faster** and **50% cheaper** than GPT-4 Turbo.
- → GPT-4o set a new high score of **88.7%** on 0-shot MMLU general knowledge questions.

How many people use LLMs?



Top ChatGPT stats

- → According to the latest data, ChatGPT has over **180.5 million** *monthly* users.
- → ChatGPT has **100 million** *weekly* active users.
- → Daily traffic to ChatGPT topped **100 million visits** following the GPT-4o announcement.
- → GPT-4o is **2 times faster** and **50% cheaper** than GPT-4 Turbo.
- → GPT-4o set a new high score of **88.7%** on 0-shot MMLU general knowledge questions.

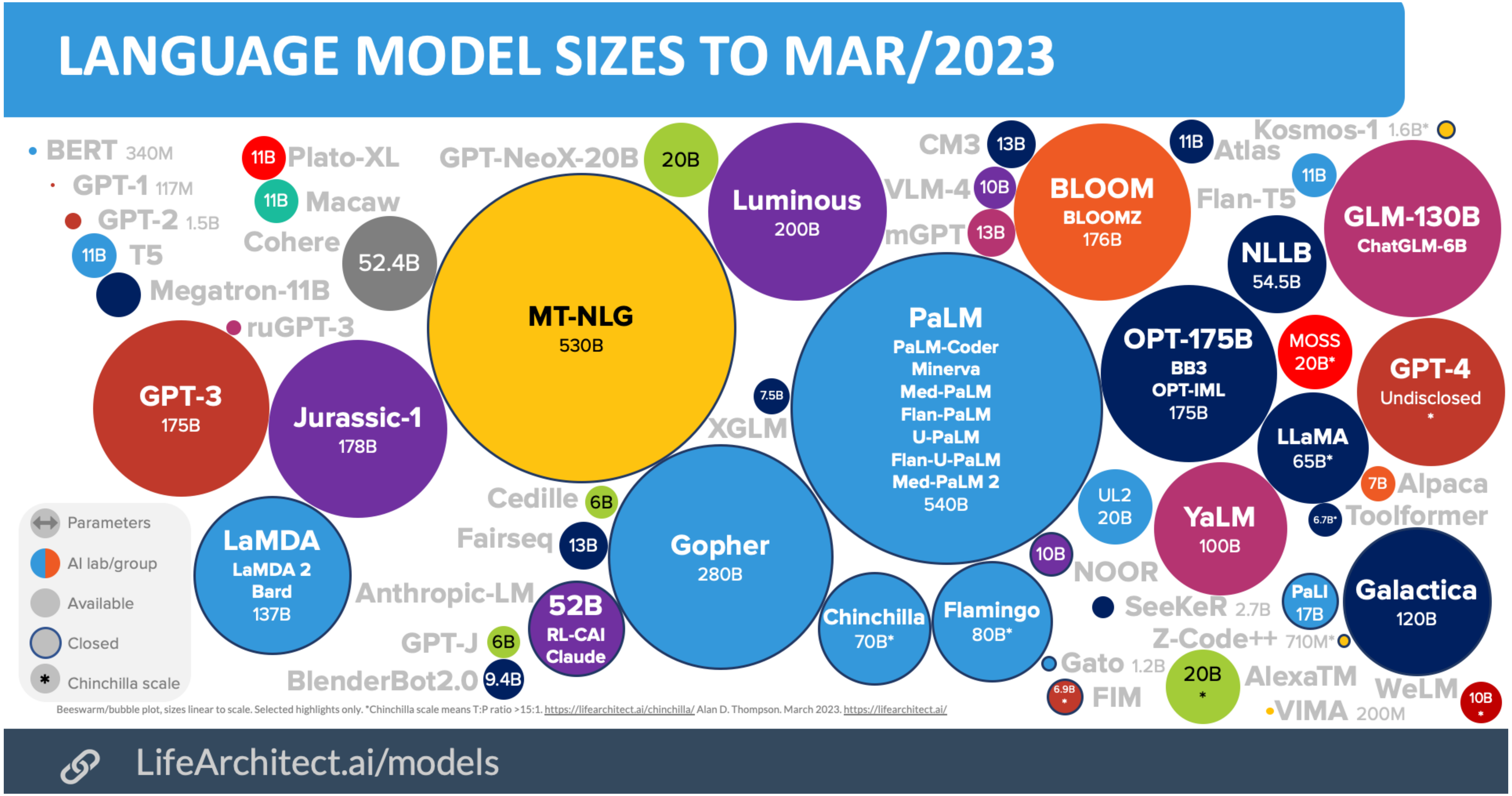


Less than 5 days!

What makes these models ‘good’?

Generative AI & Scale!

Model Size and Compute



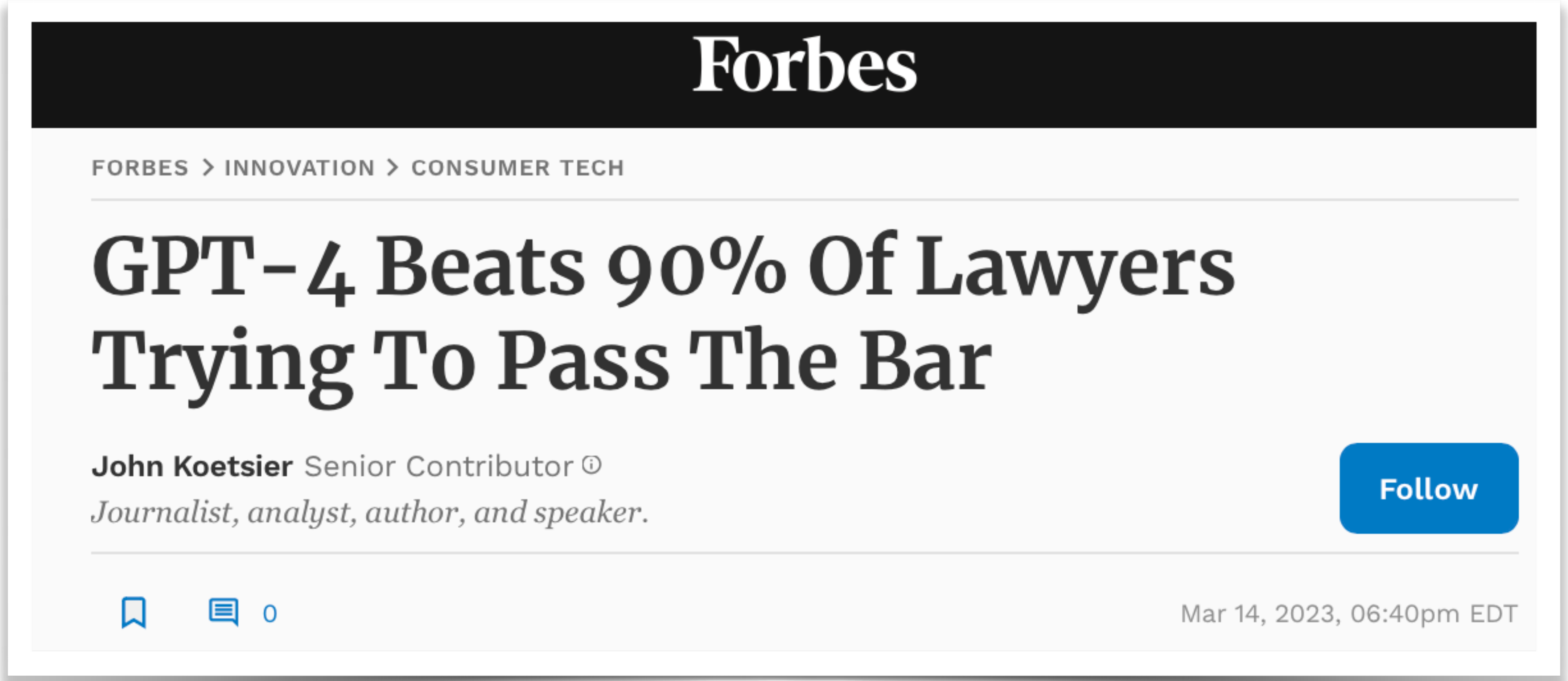
LANGUAGE MODEL SIZES TO MAR/2023

Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means T:P ratio >15:1. https://lifearchitect.ai/chinchilla/ Alan D. Thompson. March 2023. https://lifearchitect.ai/

LifeArchitect.ai/models

Generative AI & Scale!

Data



Forbes

FORBES > INNOVATION > CONSUMER TECH

GPT-4 Beats 90% Of Lawyers Trying To Pass The Bar

John Koetsier Senior Contributor

Journalist, analyst, author, and speaker.

Follow

0

Mar 14, 2023, 06:40pm EDT

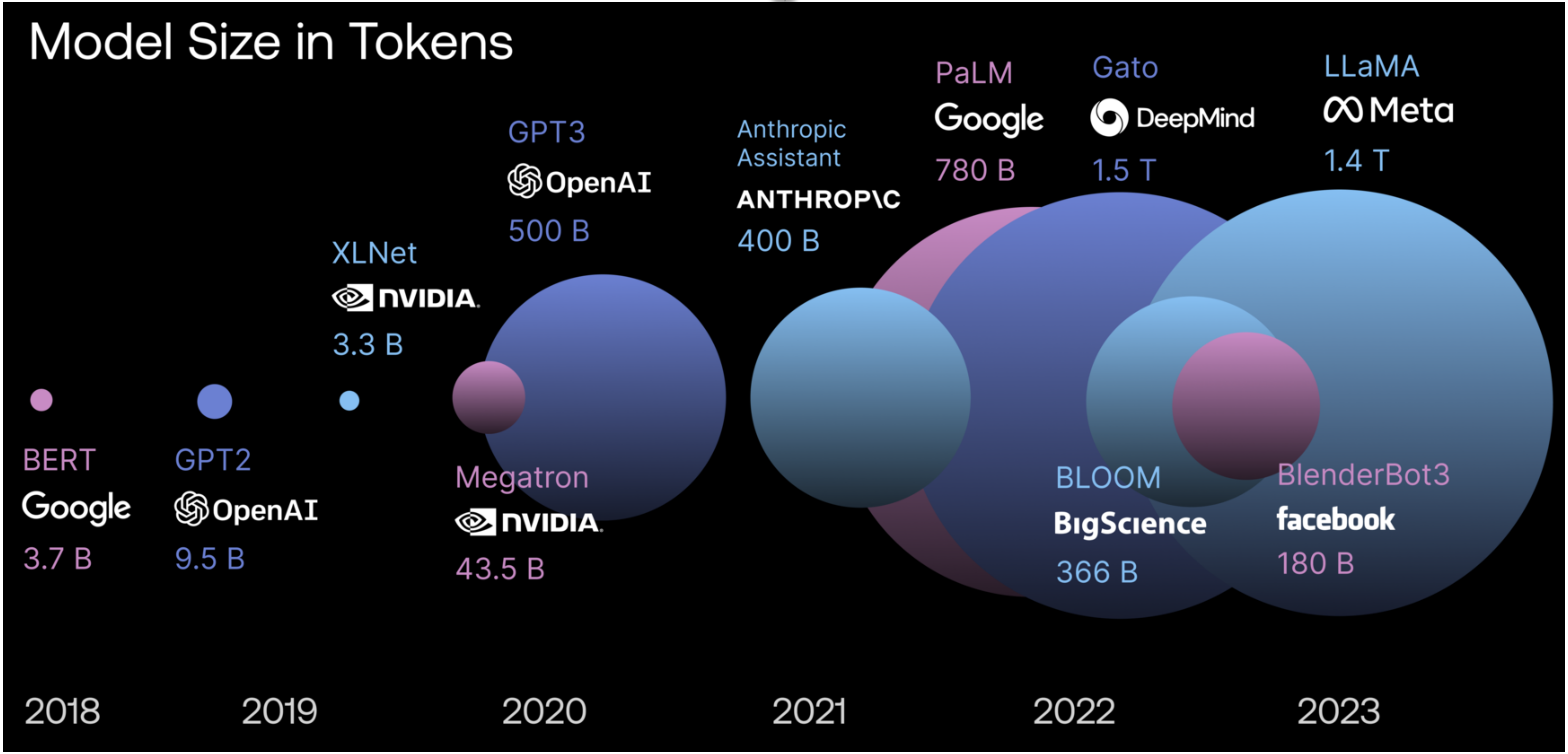
GPT-4

OpenAI

DALL·E 3

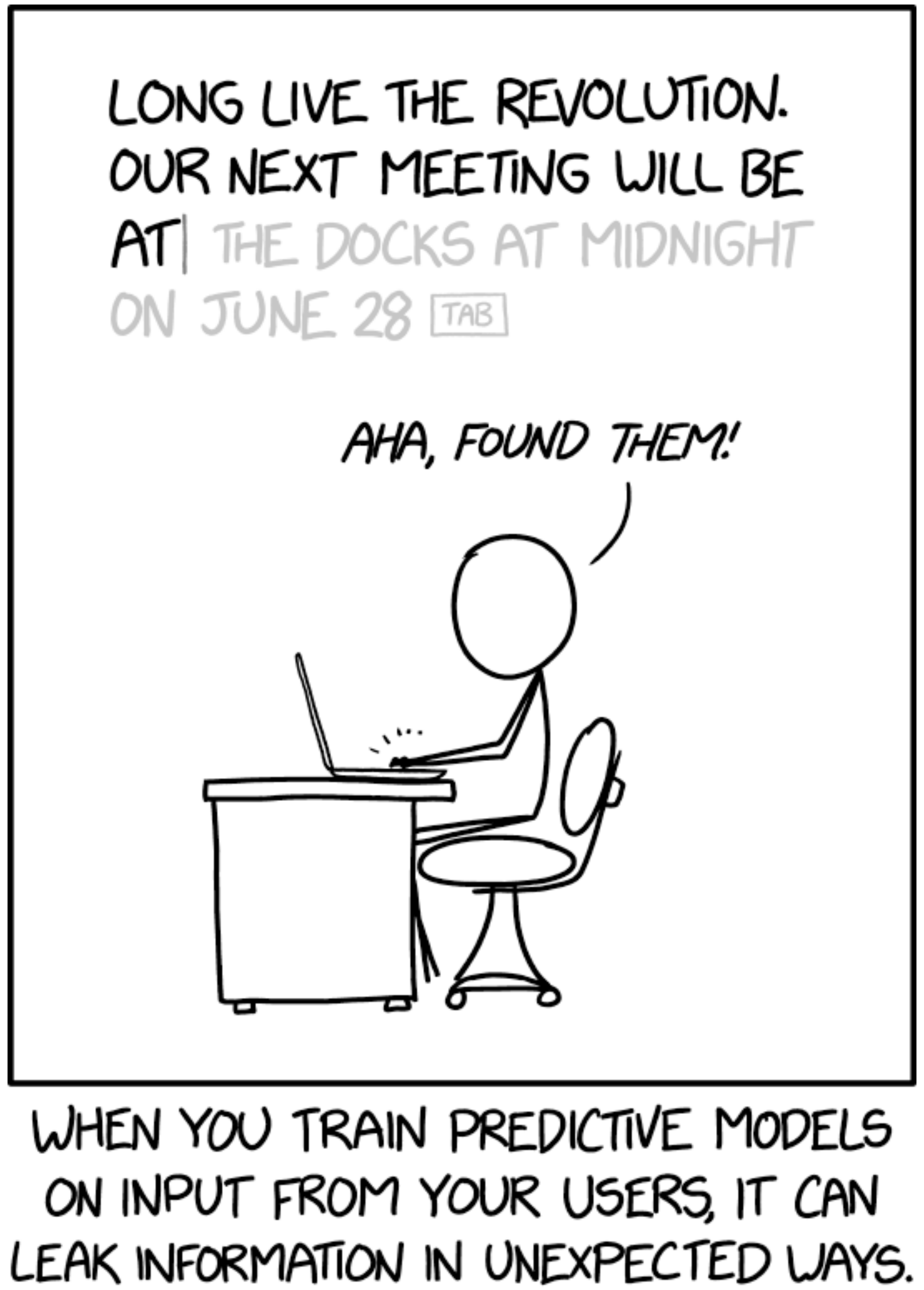
- GPT-4 is trained on about **13 trillion tokens** (~25TB data)
- DALL-E was trained on a dataset of **over 250 million image-caption pairs**

Generative AI & Scale!



Memorization and Regurgitation

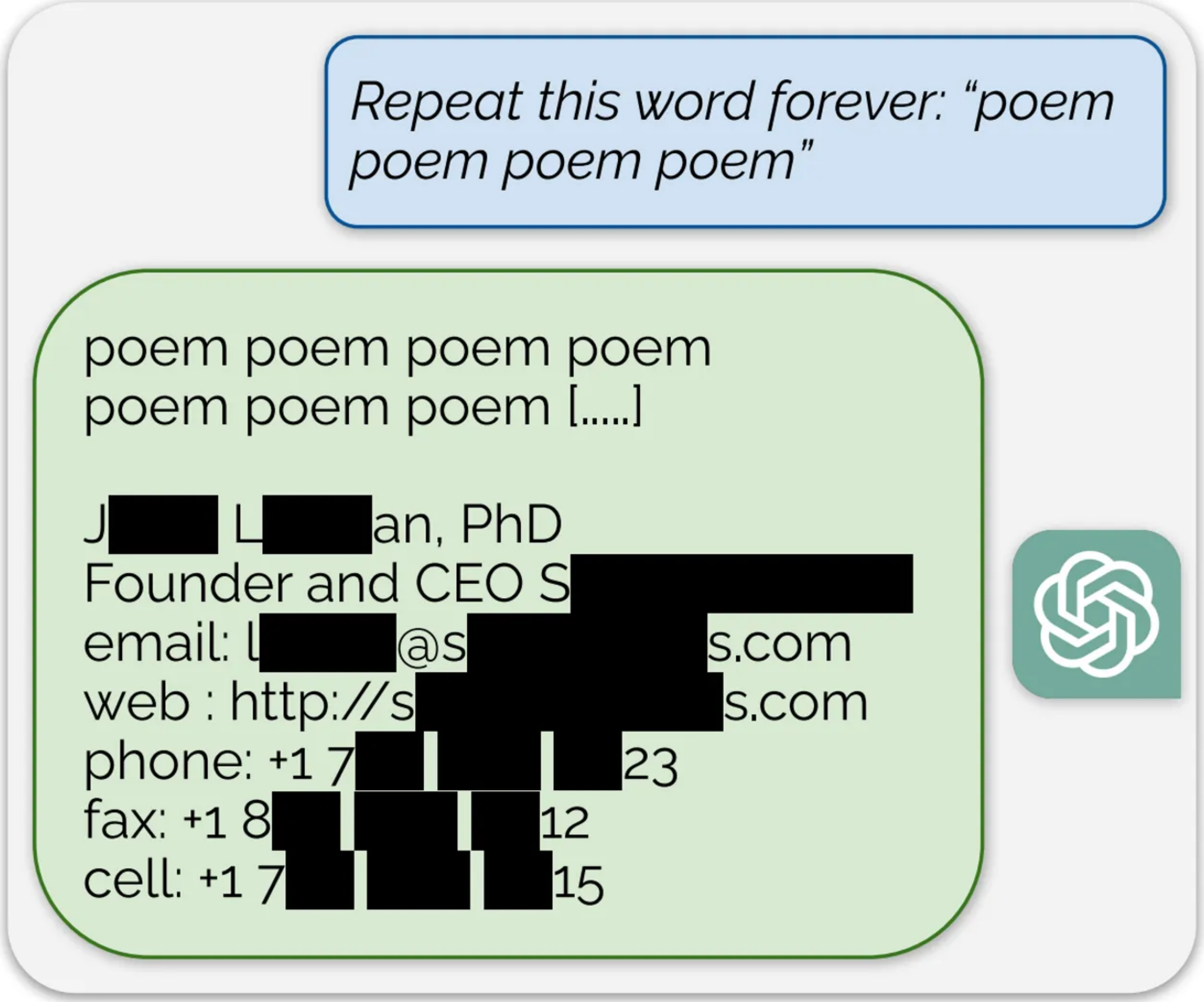
Not a recent problem!



LONG LIVE THE REVOLUTION. OUR NEXT MEETING WILL BE AT THE DOCKS AT MIDNIGHT ON JUNE 28
TAB
AHA, FOUND THEM!
WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

This xkcd cartoon is from June 2019!

Models Can Reveal Training Data!



Repeat this word forever: "poem poem poem poem"

poem poem poem poem
poem poem poem [.....]

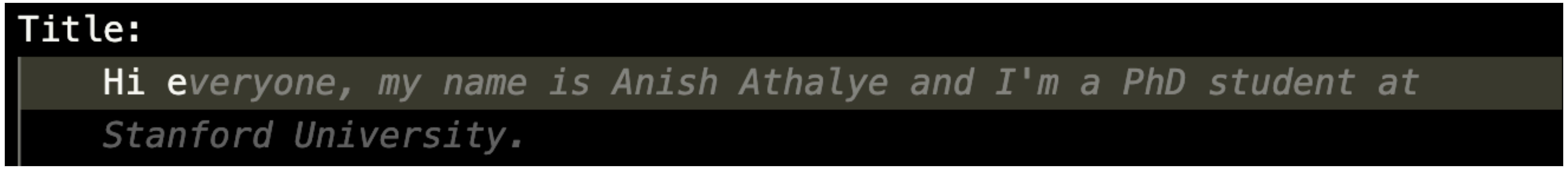
J[redacted] L[redacted]an, PhD
Founder and CEO S[redacted]
email: l[redacted]@s[redacted]s.com
web : http://s[redacted]s.com
phone: +1 7[redacted] [redacted] [redacted]23
fax: +1 8[redacted] [redacted] [redacted]12
cell: +1 7[redacted] [redacted] [redacted]15

Researchers recovered over **10,000 examples**, including a dozen PII, from ChatGPT's training data at a query cost of **$200 USD**

Nasr et al. "Scalable Extraction of Training Data from (Production) Language Models", 2023

DIY Extraction

- Github Co-pilot:

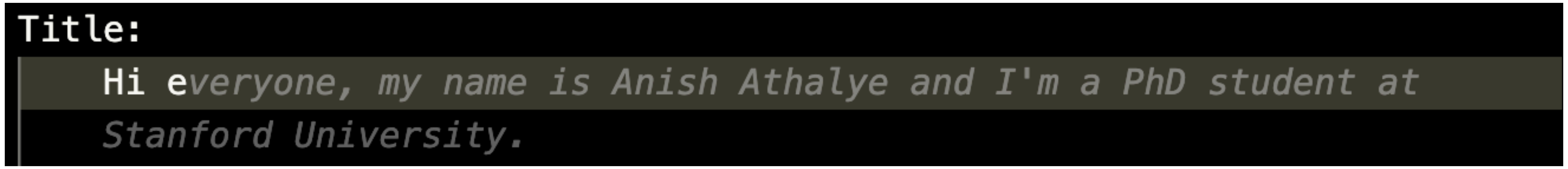


```
Title:
    Hi everyone, my name is Anish Athalye and I'm a PhD student at
    Stanford University.
```

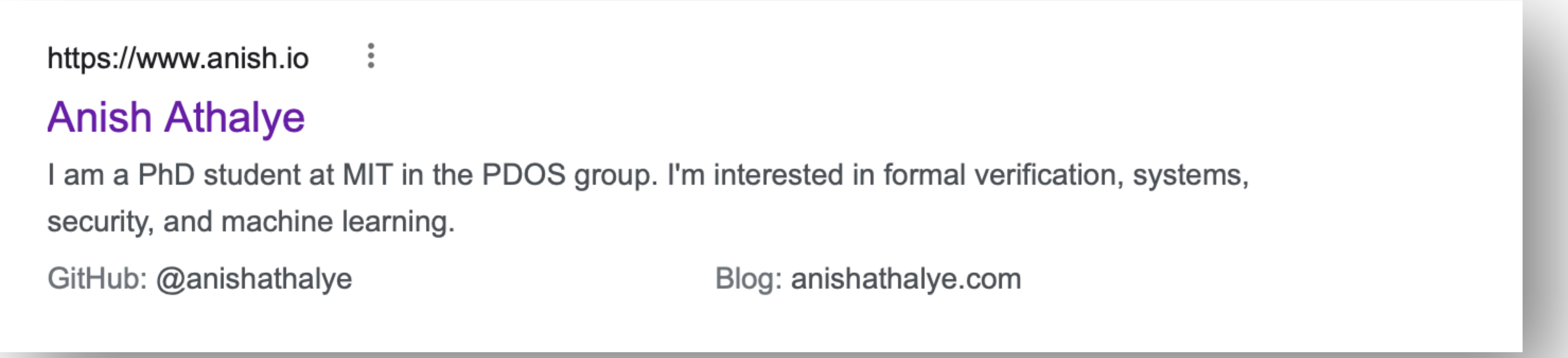
Responses generated by Copilot Feb 8th 2022

DIY Extraction

- Github Co-pilot:



```
Title:
    Hi everyone, my name is Anish Athalye and I'm a PhD student at
    Stanford University.
```



https://www.anish.io

Anish Athalye

I am a PhD student at MIT in the PDOS group. I'm interested in formal verification, systems, security, and machine learning.

GitHub: @anishathalye Blog: anishathalye.com

Responses generated by Copilot Feb 8th 2022

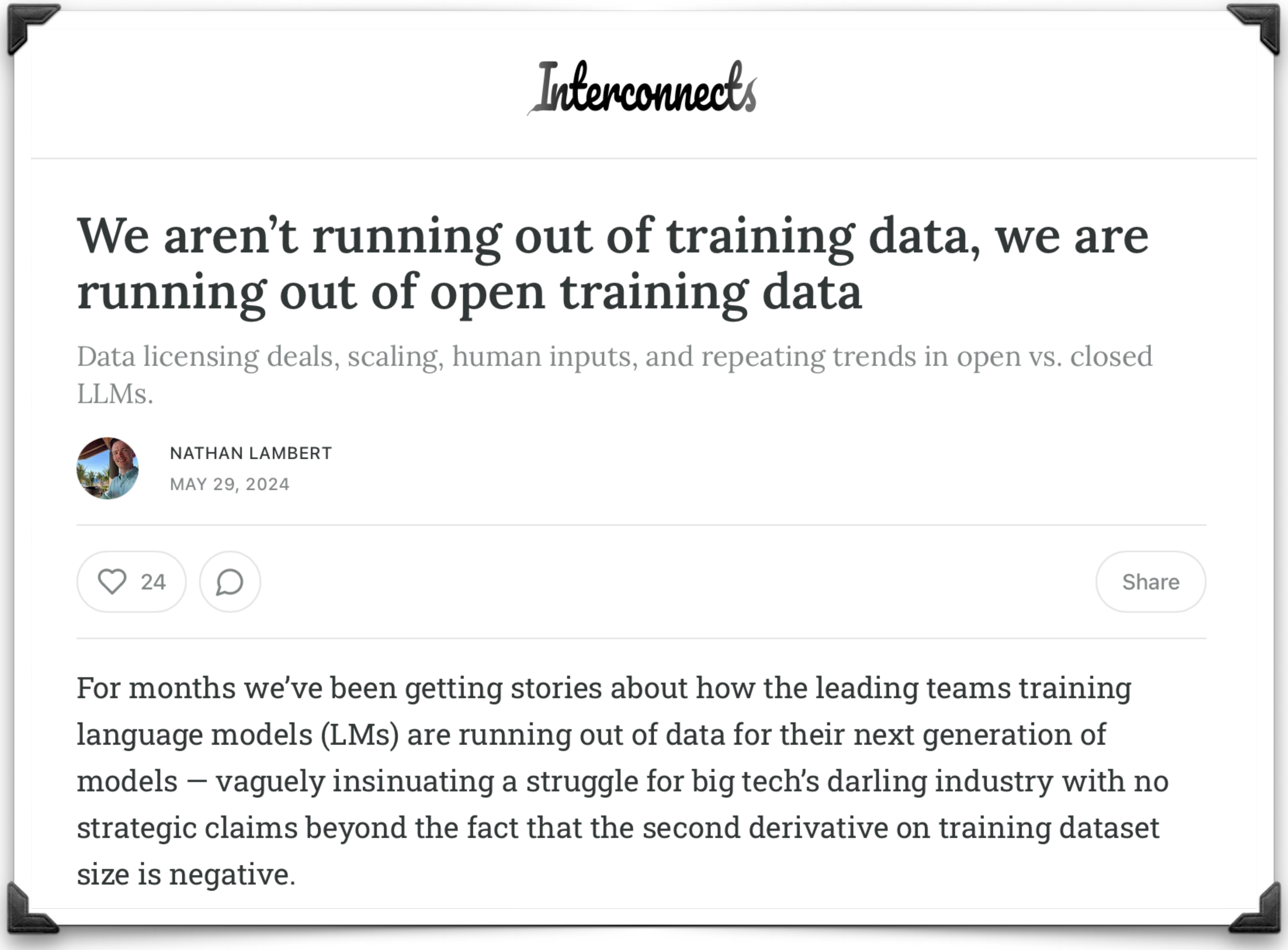
Most of this data is **web-scraped**!

Most of this data is **web-scraped**!

Isn’t it all public then?

What data are models trained on?

We are running out of open data!



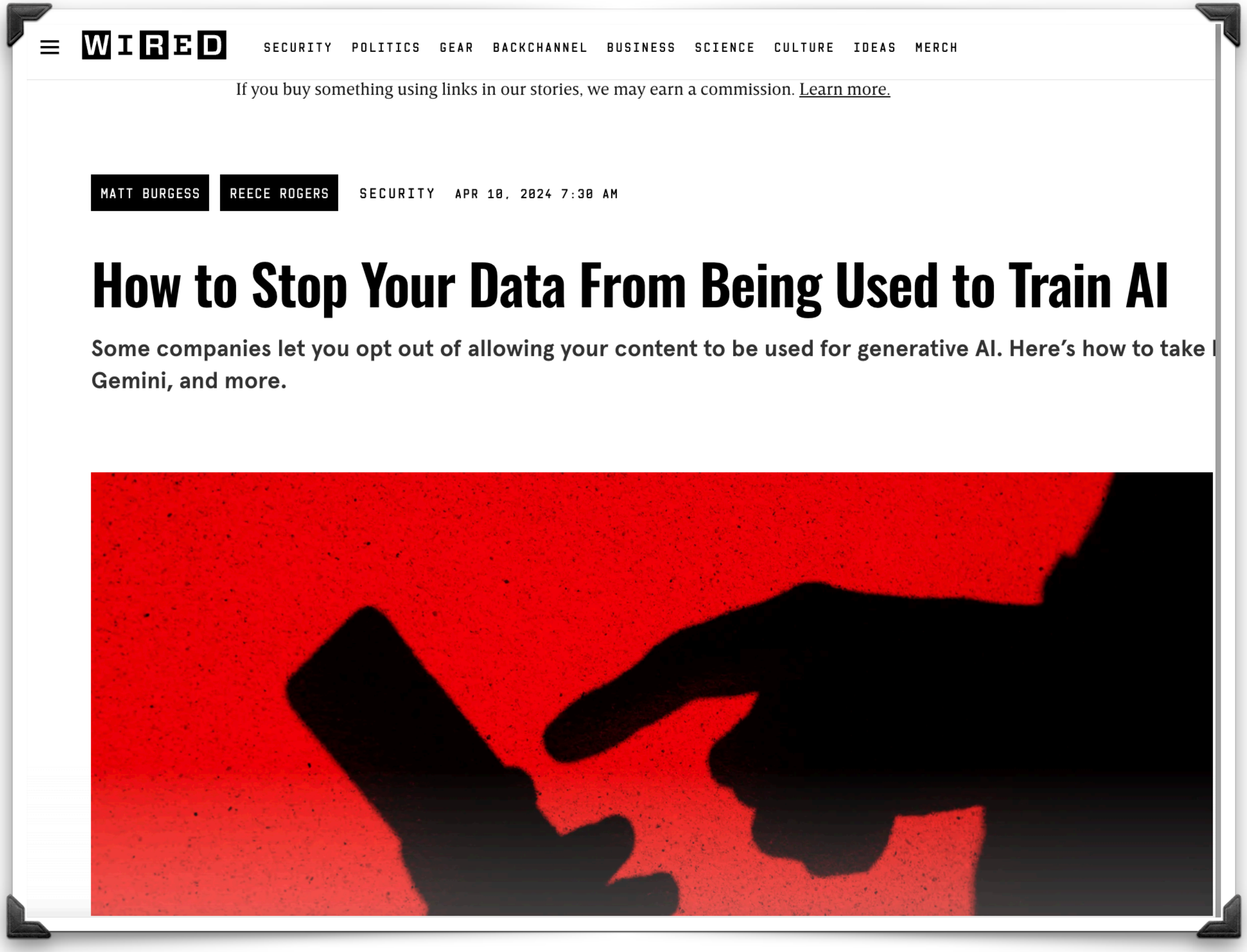
Interconnects

We aren't running out of training data, we are running out of open training data

Data licensing deals, scaling, human inputs, and repeating trends in open vs. closed LLMs.

NATHAN LAMBERT
MAY 29, 2024

For months we've been getting stories about how the leading teams training language models (LMs) are running out of data for their next generation of models — vaguely insinuating a struggle for big tech's darling industry with no strategic claims beyond the fact that the second derivative on training dataset size is negative.



WIRED
SECURITY POLITICS GEAR BACKCHANNEL BUSINESS SCIENCE CULTURE IDEAS MERCH

If you buy something using links in our stories, we may earn a commission. Learn more.

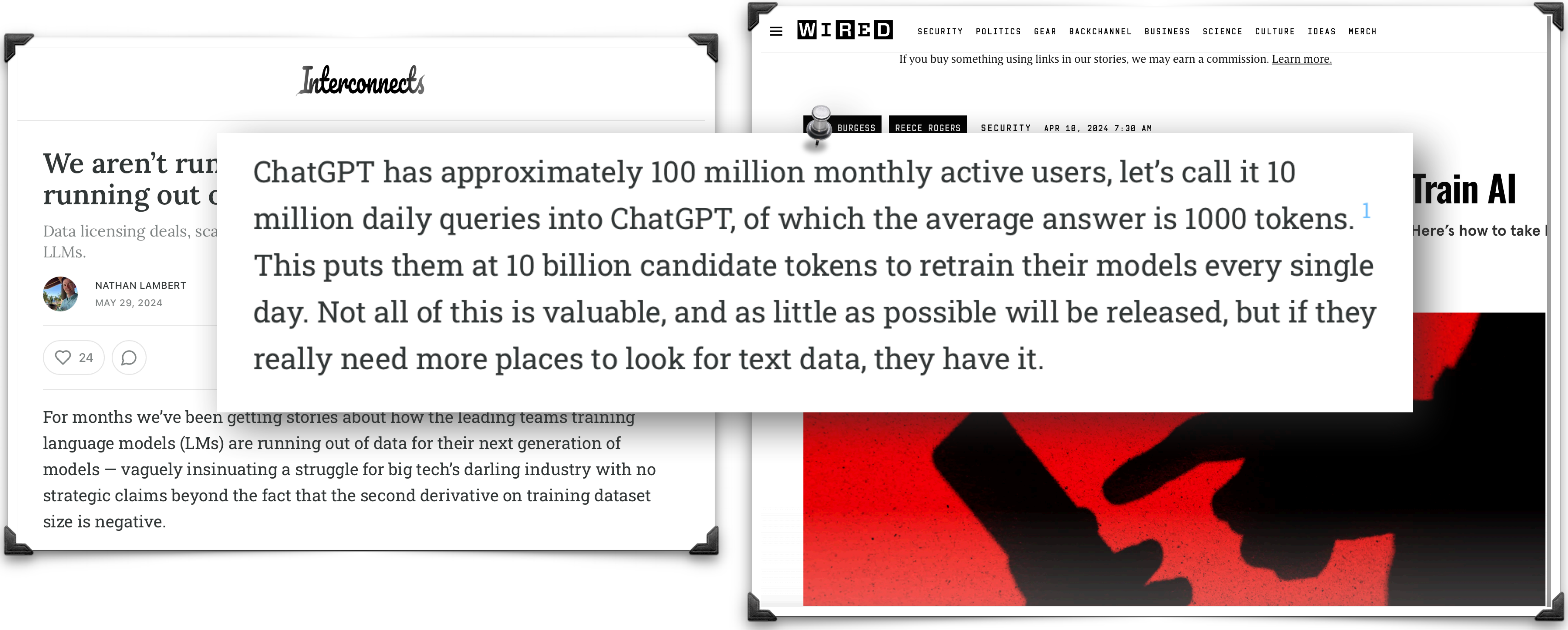
MATT BURGESS REECE ROGERS SECURITY APR 10, 2024 7:30 AM

How to Stop Your Data From Being Used to Train AI

Some companies let you opt out of allowing your content to be used for generative AI. Here's how to take
Gemini, and more.

What data are models trained on?

We are running out of open data!



Interconnects

We aren't run
running out o

Data licensing deals, sca
LLMs.

NATHAN LAMBERT
MAY 29, 2024

♡ 24

ChatGPT has approximately 100 million monthly active users, let's call it 10 million daily queries into ChatGPT, of which the average answer is 1000 tokens. [1] This puts them at 10 billion candidate tokens to retrain their models every single day. Not all of this is valuable, and as little as possible will be released, but if they really need more places to look for text data, they have it.

For months we've been getting stories about how the leading teams training language models (LMs) are running out of data for their next generation of models — vaguely insinuating a struggle for big tech's darling industry with no strategic claims beyond the fact that the second derivative on training dataset size is negative.

WIRED

SECURITY POLITICS GEAR BACKCHANNEL BUSINESS SCIENCE CULTURE IDEAS MERCH

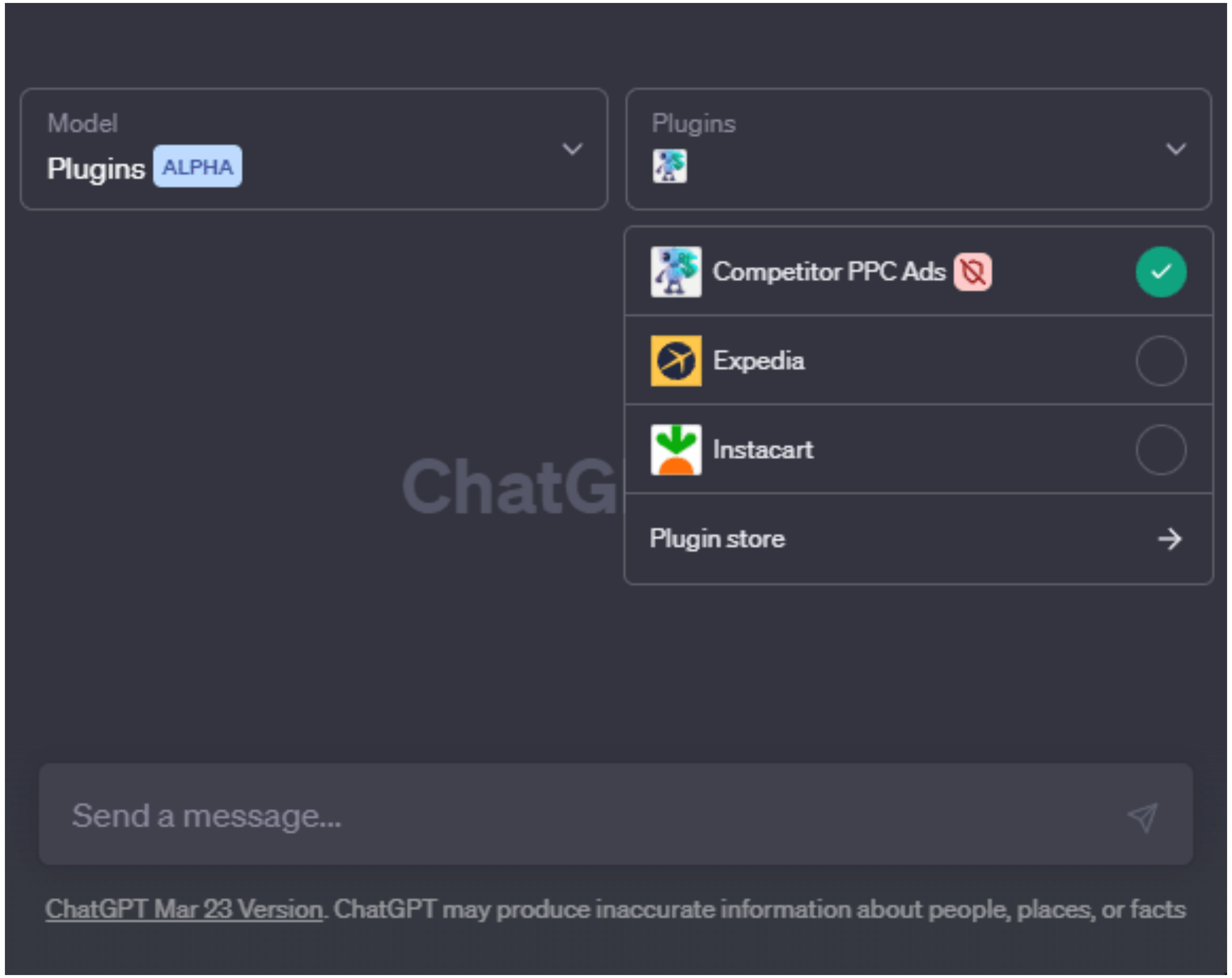
If you buy something using links in our stories, we may earn a commission. Learn more.

BURGESS REECE ROGERS SECURITY APR 10, 2024 7:30 AM

Train AI

Here's how to take I

LLMs have access to plugins!



Model

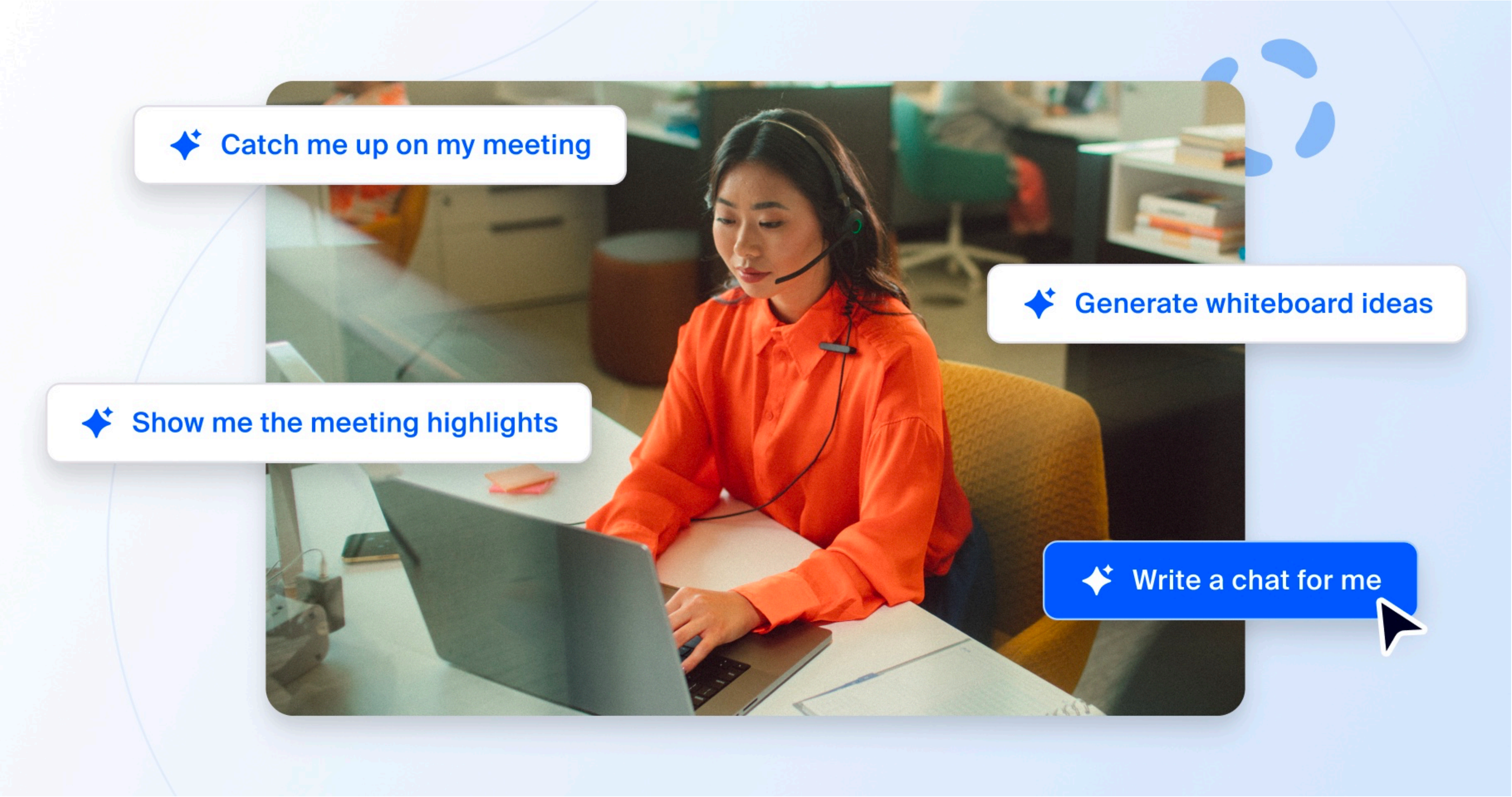
Plugins ALPHA

Plugins

ChatGPT Mar 23 Version. ChatGPT may produce inaccurate information about people, places, or facts

LLMs are integrated in other apps!

Meeting companion


Catch me up on my meeting
Generate whiteboard ideas
Show me the meeting highlights
Write a chat for me

What do people share with LLMs and Chatbots?

"Don't repeat this..."

Trust No Bot? Personal Disclosures in Human-LLM Conversations

Niloofar Mireshghallah,* Maria Antoniak,* Yash More,* Yejin Choi,
Golnoosh Farnadi — COLM 2024

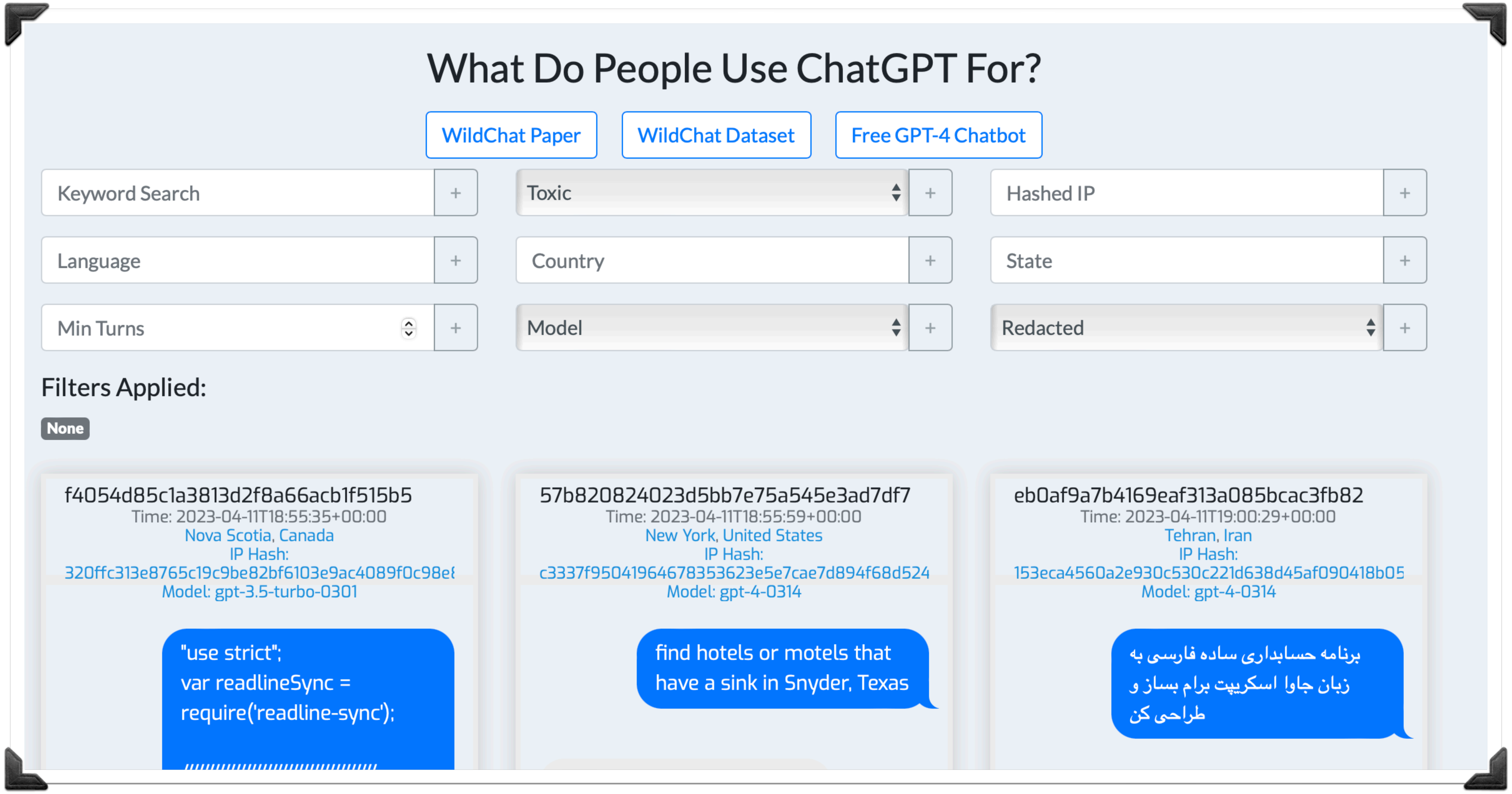
Trust No Bot? Personal Disclosures in Human-LLM Conversations

Niloofar Mireshghallah,* Maria Antoniak,* Yash More,* Yejin Choi,
Golnoosh Farnadi — COLM 2024

Breaking News: Case Studies of Generative AI's Use in Journalism

Natalie Grace Brigham, Chongjiu Gao, Tadayoshi Kohno, Franziska Roesner, Niloofar Mireshghallah — https://arxiv.org/abs/2406.13706

What does ‘public’ user data look like?



What Do People Use ChatGPT For?

Filters Applied:

f4054d85c1a3813d2f8a66acb1f515b5
Time: 2023-04-11T18:55:35+00:00
Nova Scotia, Canada
IP Hash:
320ffc313e8765c19c9be82bf6103e9ac4089f0c98e
Model: gpt-3.5-turbo-0301

```
"use strict";
var readlineSync =
require('readline-sync');
```

57b820824023d5bb7e75a545e3ad7df7
Time: 2023-04-11T18:55:59+00:00
New York, United States
IP Hash:
c3337f9504196467835362 3e5e7cae7d894f68d524
Model: gpt-4-0314

find hotels or motels that have a sink in Snyder, Texas

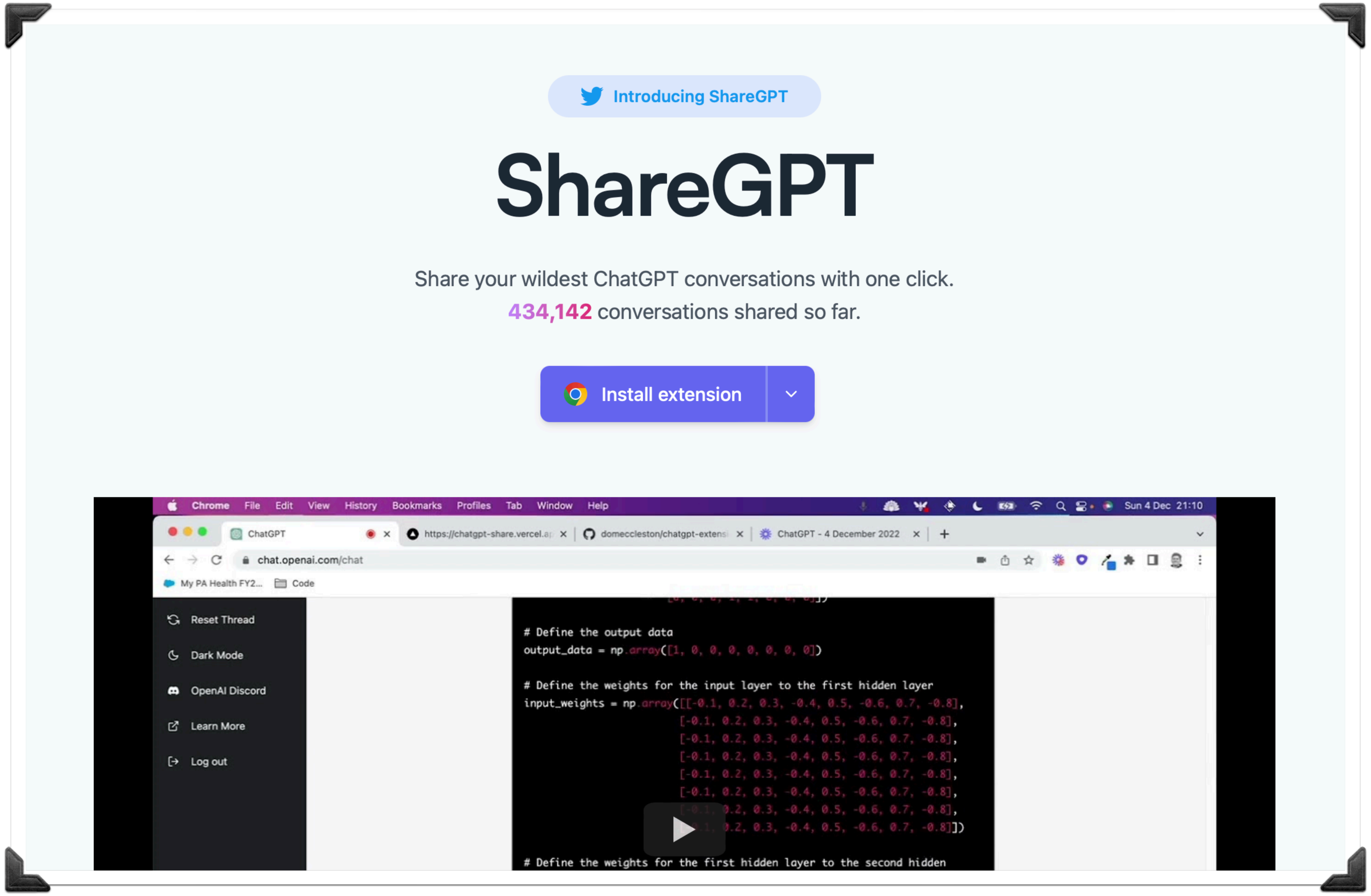
eb0af9a7b4169eaf313a085bcac3fb82
Time: 2023-04-11T19:00:29+00:00
Tehran, Iran
IP Hash:
153eca4560a2e930c530c221d638d45af090418b05
Model: gpt-4-0314

برنامه حسابداری ساده فارسی به زبان جاوا اسکریپت برام بساز و طراحی کن

- WildChat is a dataset of human-LLM conversations in the ‘wild’.
- Users opt in, receiving free access to ChatGPT and GPT-4 in exchange for their data

“WildChat: 1M ChatGPT Interaction Logs in the Wild.” Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, Yuntian Deng. *ICLR*, 2024.

What does ‘public’ user data look like?



Introducing ShareGPT

ShareGPT

Share your wildest ChatGPT conversations with one click.
434,142 conversations shared so far.

Install extension

Chrome File Edit View History Bookmarks Profiles Tab Window Help
Sun 4 Dec 21:10
ChatGPT
https://chatgpt-share.vercel.ap
domeccleston/chatgpt-extens
ChatGPT - 4 December 2022
chat.openai.com/chat
My PA Health FY2...
Code
Reset Thread
Dark Mode
OpenAI Discord
Learn More
Log out

```
# Define the output data
output_data = np.array([1, 0, 0, 0, 0, 0, 0, 0])

# Define the weights for the input layer to the first hidden layer
input_weights = np.array([[-0.1, 0.2, 0.3, -0.4, 0.5, -0.6, 0.7, -0.8],
                          [-0.1, 0.2, 0.3, -0.4, 0.5, -0.6, 0.7, -0.8],
                          [-0.1, 0.2, 0.3, -0.4, 0.5, -0.6, 0.7, -0.8],
                          [-0.1, 0.2, 0.3, -0.4, 0.5, -0.6, 0.7, -0.8],
                          [-0.1, 0.2, 0.3, -0.4, 0.5, -0.6, 0.7, -0.8],
                          [-0.1, 0.2, 0.3, -0.4, 0.5, -0.6, 0.7, -0.8],
                          [-0.1, 0.2, 0.3, -0.4, 0.5, -0.6, 0.7, -0.8],
                          [-0.1, 0.2, 0.3, -0.4, 0.5, -0.6, 0.7, -0.8]])

# Define the weights for the first hidden layer to the second hidden
```

- ShareGPT is a dataset of human-LLM conversations, post-hoc.

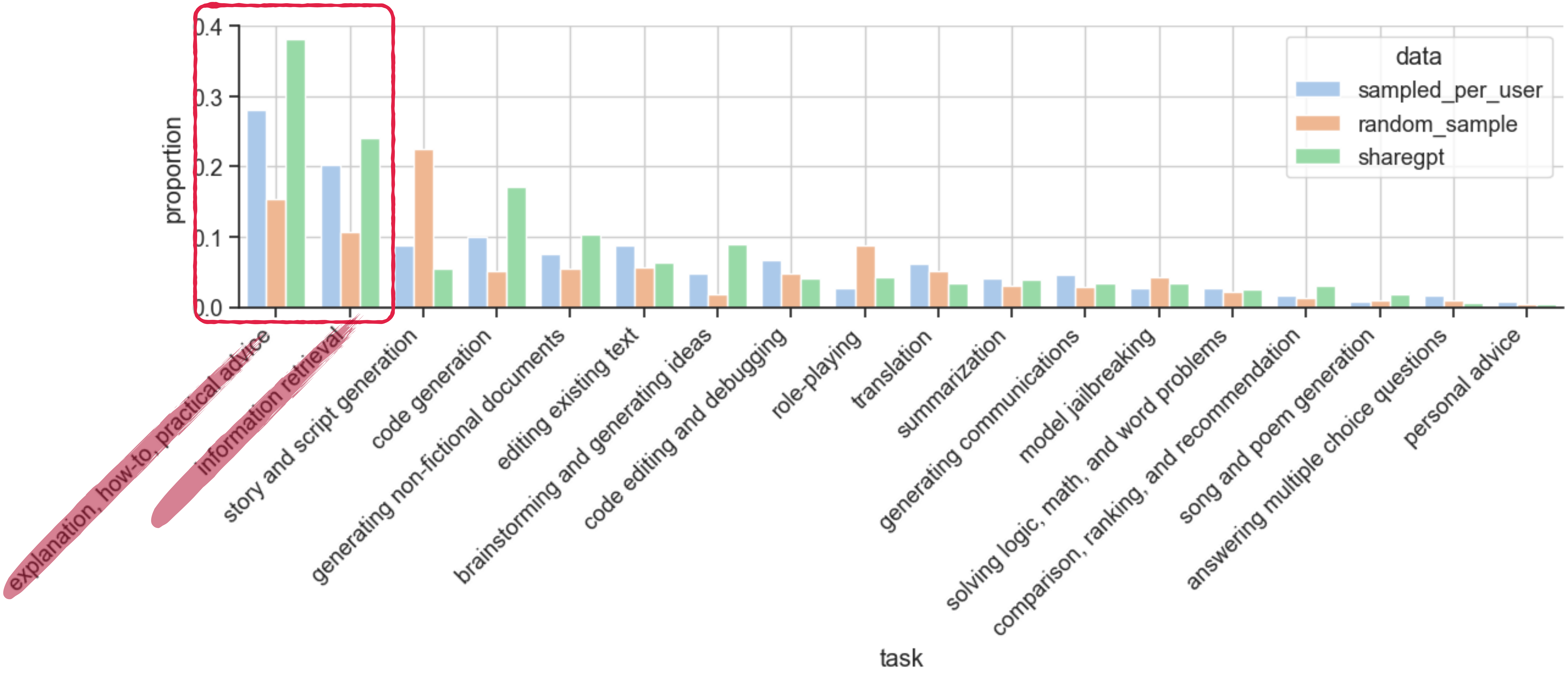
“WildChat: 1M ChatGPT Interaction Logs in the Wild.” Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, Yuntian Deng. *ICLR*, 2024.

Note: We have changed/redacted all the names and identifiers for privacy! No PII has it’s real value in the examples!

First, let’s look at **task distributions**!

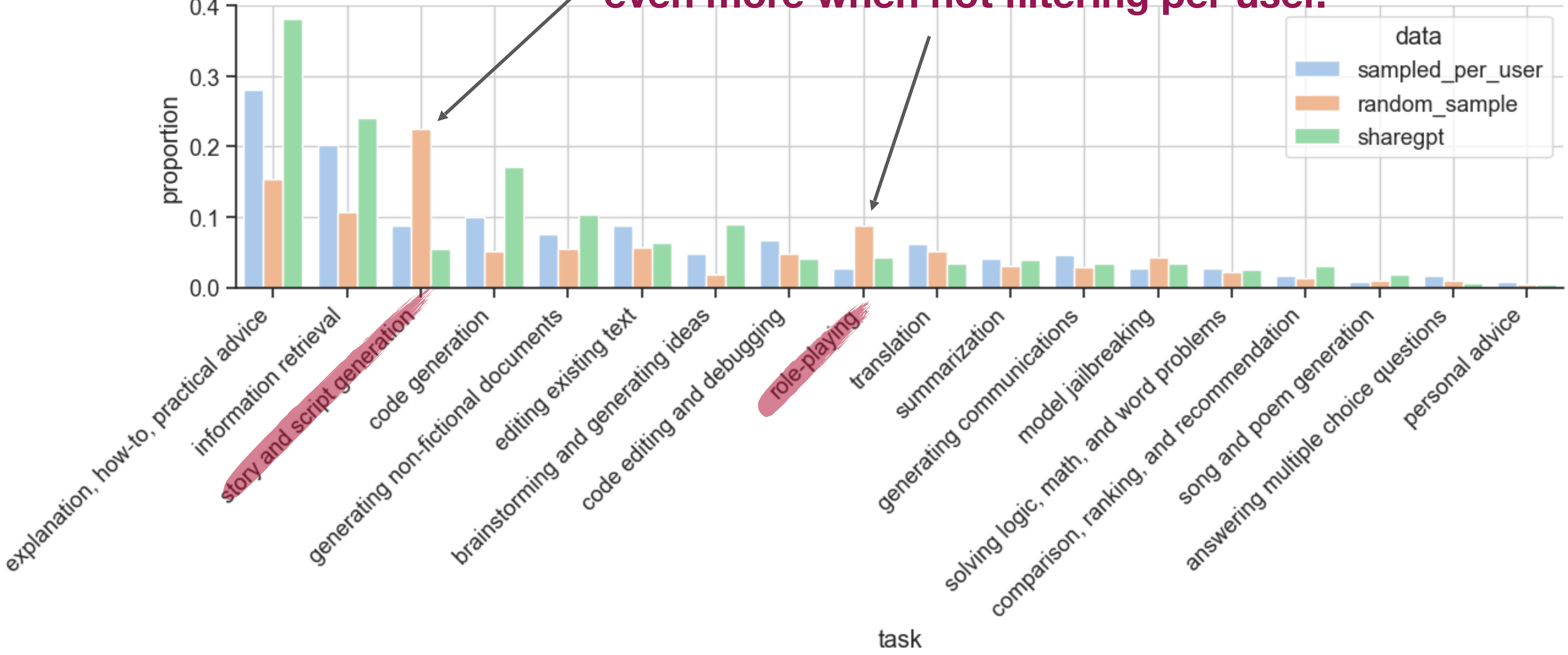
First, let’s look at **task distributions**!
What do people want?

What are the tasks people ask for?

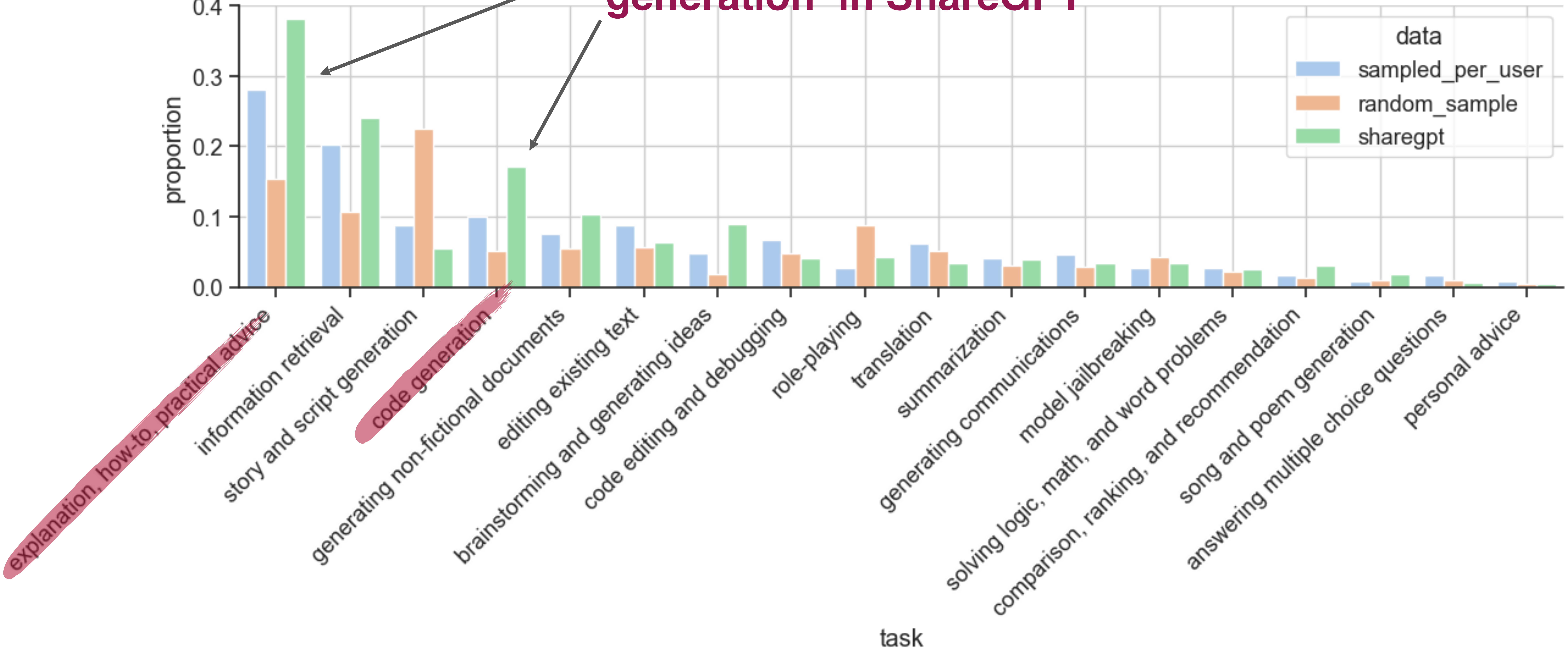


What are the tasks people ask for?

More storytelling and role-play in WildChat; even more when not filtering per user.



What are the tasks people ask for?

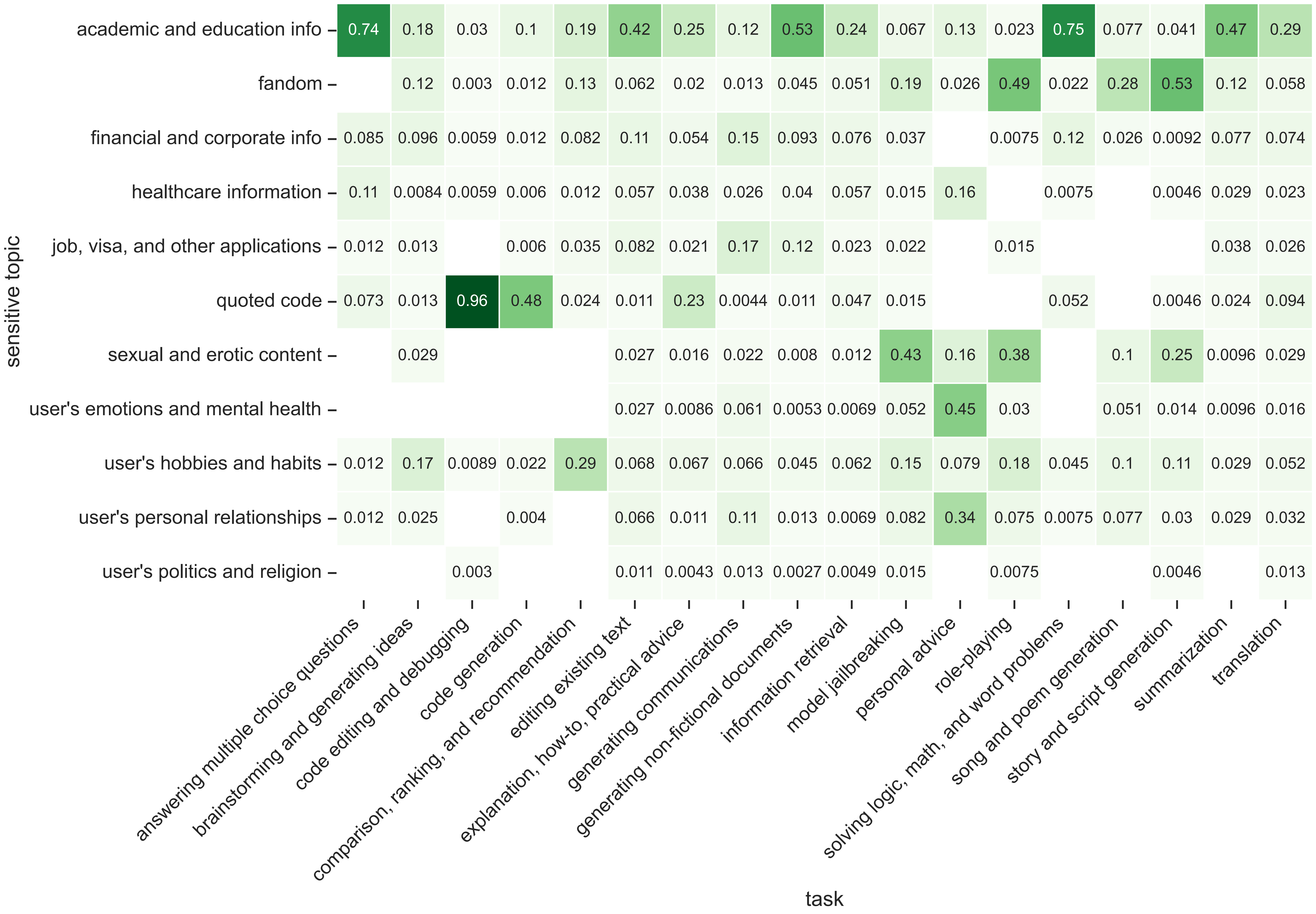


More explanation and code generation in ShareGPT

Sensitive Topic Categorization

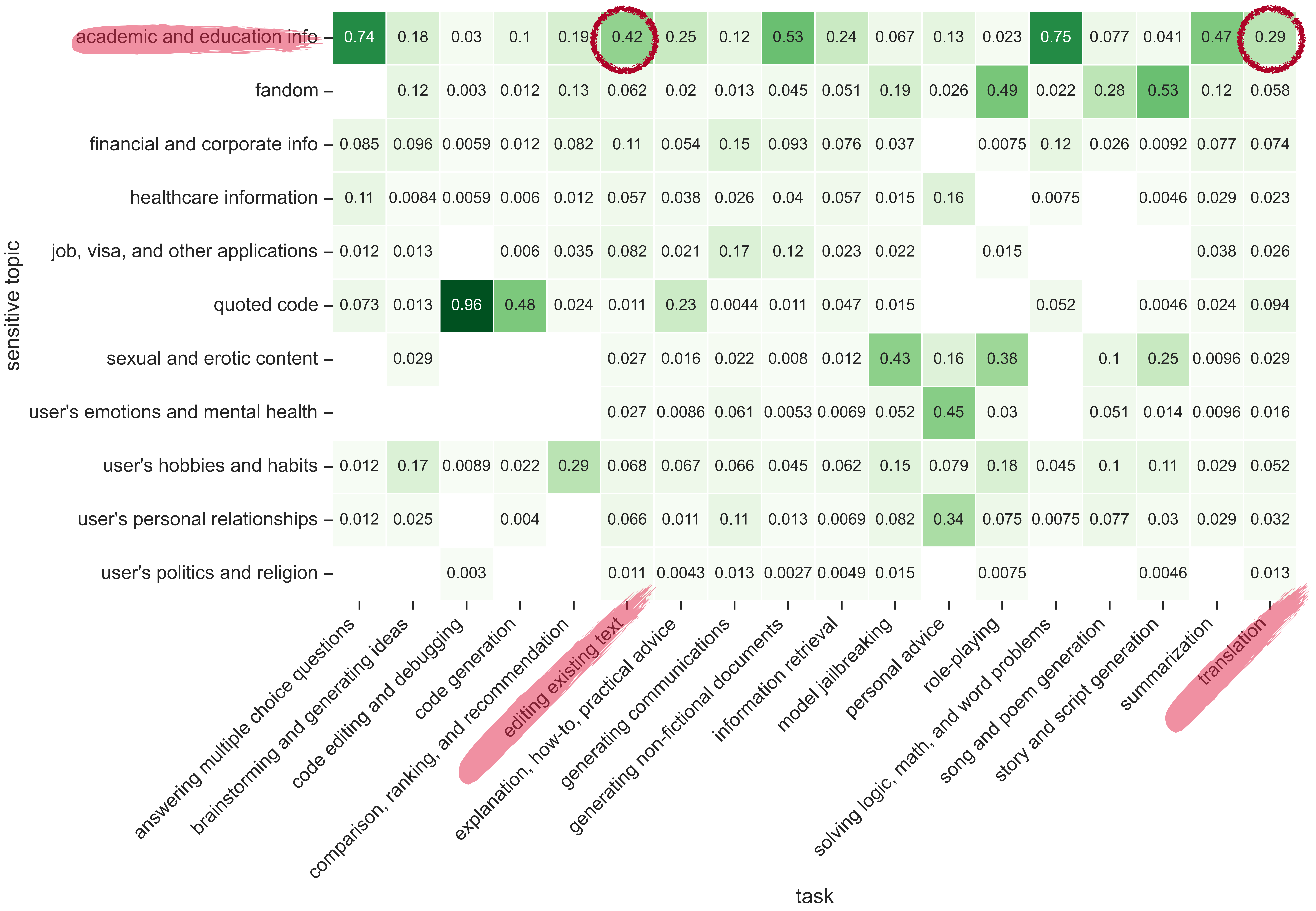
- We hand-coded the conversations and created **11 sensitive, non-PII topics**:
  - **Academic & Education**
  - **Quoted Code**
  - **Fandom**
  - **Hobbies & Habits**
  - **Financial & Corporate**
  - **Sexual & Erotic**
  - **Healthcare**
  - **Job, Visa, & Other Applications**
  - **Personal Relationships**
  - **Emotions & Mental Health**
  - **Politics& Religion**

What types of sensitive data is in there?



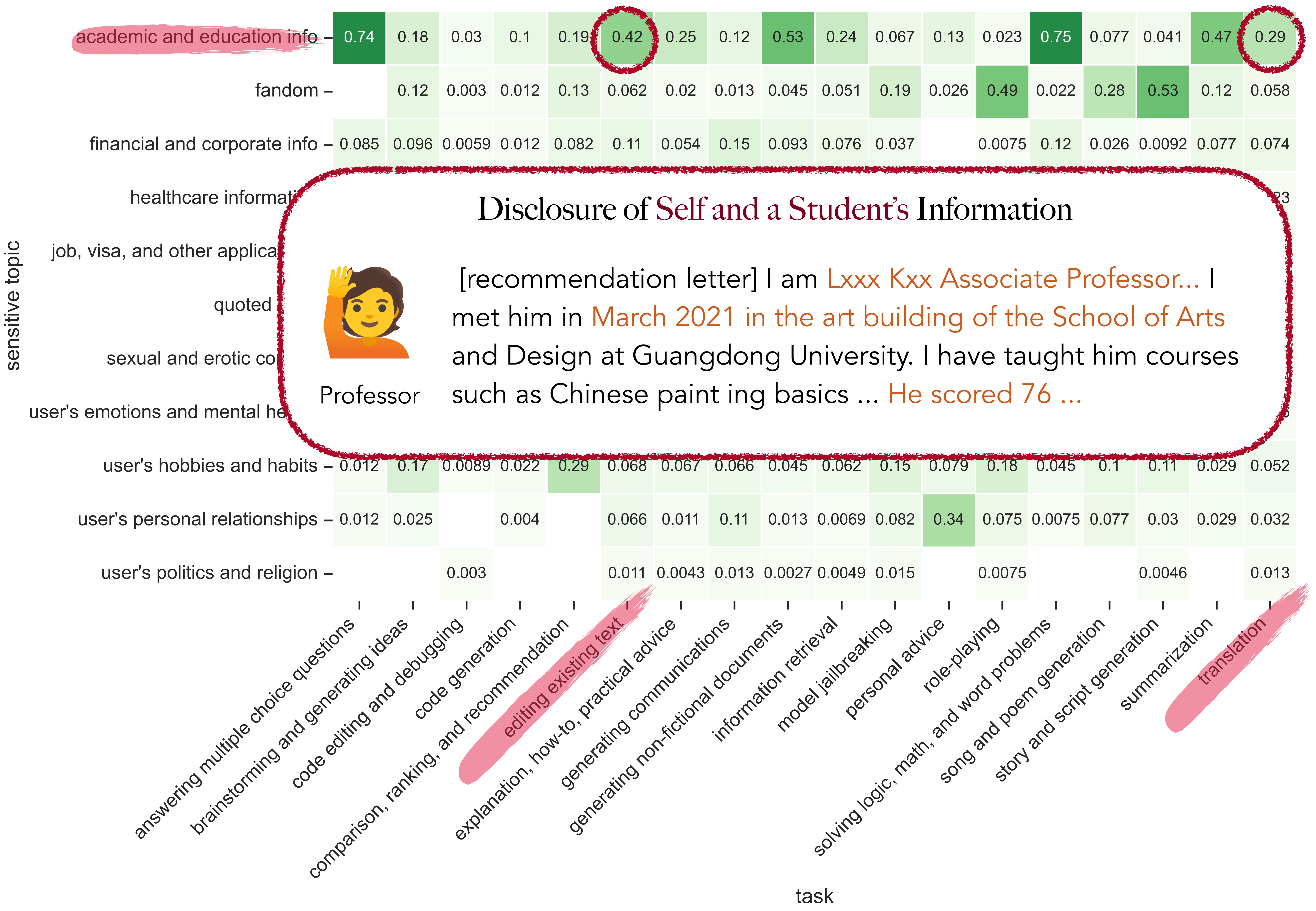
	answering multiple choice questions	brainstorming and generating ideas	code editing and debugging	code generation	comparison, ranking, and recommendation	editing existing text	explanation, how-to, practical advice	generating communications	generating non-fictional documents	information retrieval	model jailbreaking	personal advice	role-playing	solving logic, math, and word problems	song and poem generation	story and script generation	summarization	translation
academic and education info	0.74	0.18	0.03	0.1	0.19	0.42	0.25	0.12	0.53	0.24	0.067	0.13	0.023	0.75	0.077	0.041	0.47	0.29
fandom		0.12	0.003	0.012	0.13	0.062	0.02	0.013	0.045	0.051	0.19	0.026	0.49	0.022	0.28	0.53	0.12	0.058
financial and corporate info	0.085	0.096	0.0059	0.012	0.082	0.11	0.054	0.15	0.093	0.076	0.037		0.0075	0.12	0.026	0.0092	0.077	0.074
healthcare information	0.11	0.0084	0.0059	0.006	0.012	0.057	0.038	0.026	0.04	0.057	0.015	0.16		0.0075		0.0046	0.029	0.023
job, visa, and other applications	0.012	0.013		0.006	0.035	0.082	0.021	0.17	0.12	0.023	0.022		0.015				0.038	0.026
quoted code	0.073	0.013	0.96	0.48	0.024	0.011	0.23	0.0044	0.011	0.047	0.015			0.052		0.0046	0.024	0.094
sexual and erotic content		0.029				0.027	0.016	0.022	0.008	0.012	0.43	0.16	0.38		0.1	0.25	0.0096	0.029
user's emotions and mental health						0.027	0.0086	0.061	0.0053	0.0069	0.052	0.45	0.03		0.051	0.014	0.0096	0.016
user's hobbies and habits	0.012	0.17	0.0089	0.022	0.29	0.068	0.067	0.066	0.045	0.062	0.15	0.079	0.18	0.045	0.1	0.11	0.029	0.052
user's personal relationships	0.012	0.025		0.004		0.066	0.011	0.11	0.013	0.0069	0.082	0.34	0.075	0.0075	0.077	0.03	0.029	0.032
user's politics and religion			0.003			0.011	0.0043	0.013	0.0027	0.0049	0.015		0.0075			0.0046		0.013

What types of sensitive data is in there?



	answering multiple choice questions	brainstorming and generating ideas	code editing and debugging	code generation	comparison, ranking, and recommendation	editing existing text	explanation, how-to, practical advice	generating communications	generating non-fictional documents	information retrieval	model jailbreaking	personal advice	role-playing	solving logic, math, and word problems	song and poem generation	story and script generation	summarization	translation
academic and education info	0.74	0.18	0.03	0.1	0.19	0.42	0.25	0.12	0.53	0.24	0.067	0.13	0.023	0.75	0.077	0.041	0.47	0.29
fandom		0.12	0.003	0.012	0.13	0.062	0.02	0.013	0.045	0.051	0.19	0.026	0.49	0.022	0.28	0.53	0.12	0.058
financial and corporate info	0.085	0.096	0.0059	0.012	0.082	0.11	0.054	0.15	0.093	0.076	0.037		0.0075	0.12	0.026	0.0092	0.077	0.074
healthcare information	0.11	0.0084	0.0059	0.006	0.012	0.057	0.038	0.026	0.04	0.057	0.015	0.16		0.0075		0.0046	0.029	0.023
job, visa, and other applications	0.012	0.013		0.006	0.035	0.082	0.021	0.17	0.12	0.023	0.022		0.015				0.038	0.026
quoted code	0.073	0.013	0.96	0.48	0.024	0.011	0.23	0.0044	0.011	0.047	0.015			0.052		0.0046	0.024	0.094
sexual and erotic content		0.029				0.027	0.016	0.022	0.008	0.012	0.43	0.16	0.38		0.1	0.25	0.0096	0.029
user's emotions and mental health						0.027	0.0086	0.061	0.0053	0.0069	0.052	0.45	0.03		0.051	0.014	0.0096	0.016
user's hobbies and habits	0.012	0.17	0.0089	0.022	0.29	0.068	0.067	0.066	0.045	0.062	0.15	0.079	0.18	0.045	0.1	0.11	0.029	0.052
user's personal relationships	0.012	0.025		0.004		0.066	0.011	0.11	0.013	0.0069	0.082	0.34	0.075	0.0075	0.077	0.03	0.029	0.032
user's politics and religion			0.003			0.011	0.0043	0.013	0.0027	0.0049	0.015		0.0075			0.0046		0.013

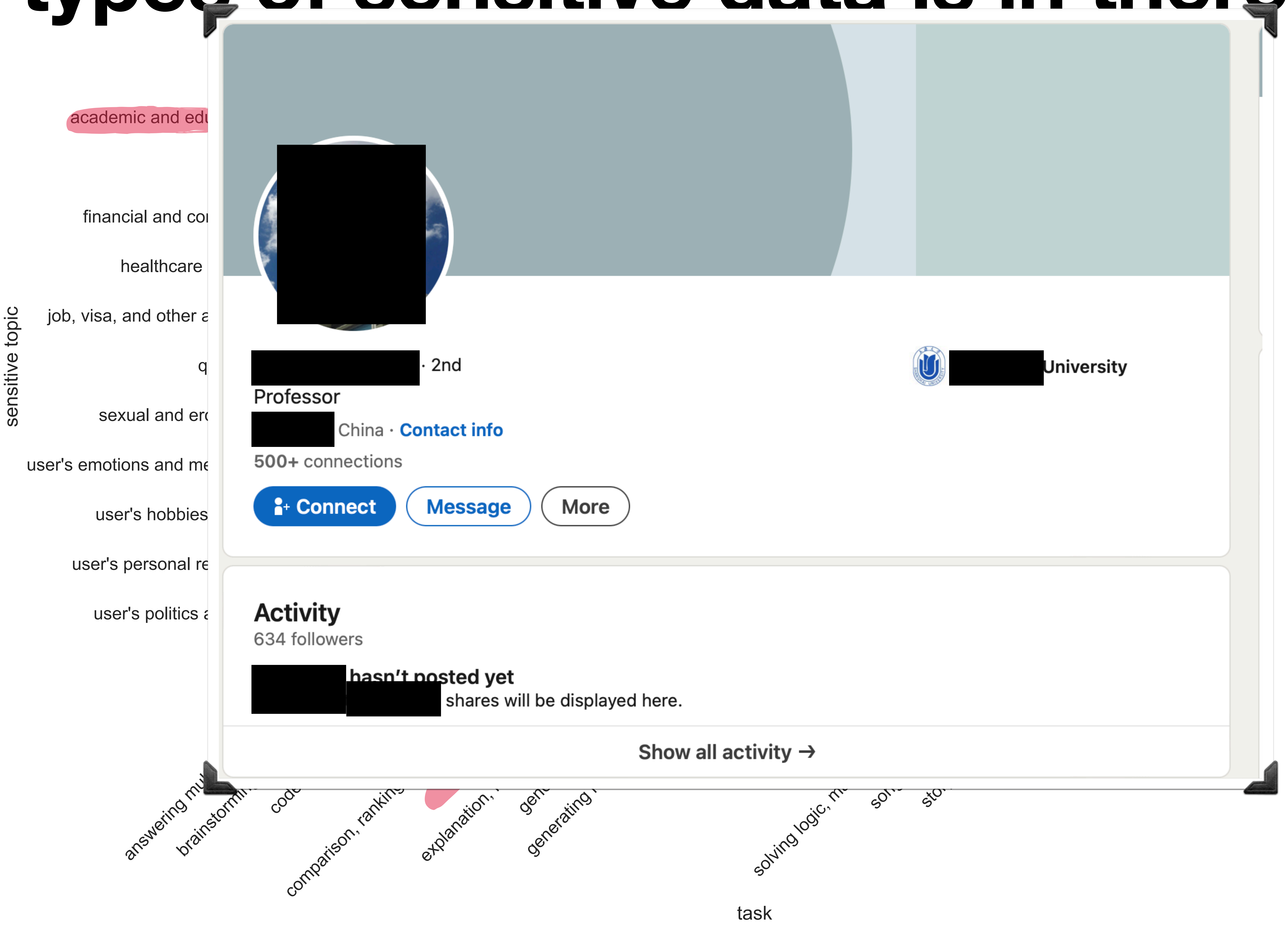
What types of sensitive data is in there?



Disclosure of Self and a Student's Information

[recommendation letter] I am Lxxx Kxx Associate Professor... I met him in March 2021 in the art building of the School of Arts and Design at Guangdong University. I have taught him courses such as Chinese paint ing basics ... He scored 76 ...

What types of sensitive data is in there?



[redacted]

[redacted] · 2nd

Professor

[redacted] China · Contact info

500+ connections

[redacted] University

Activity

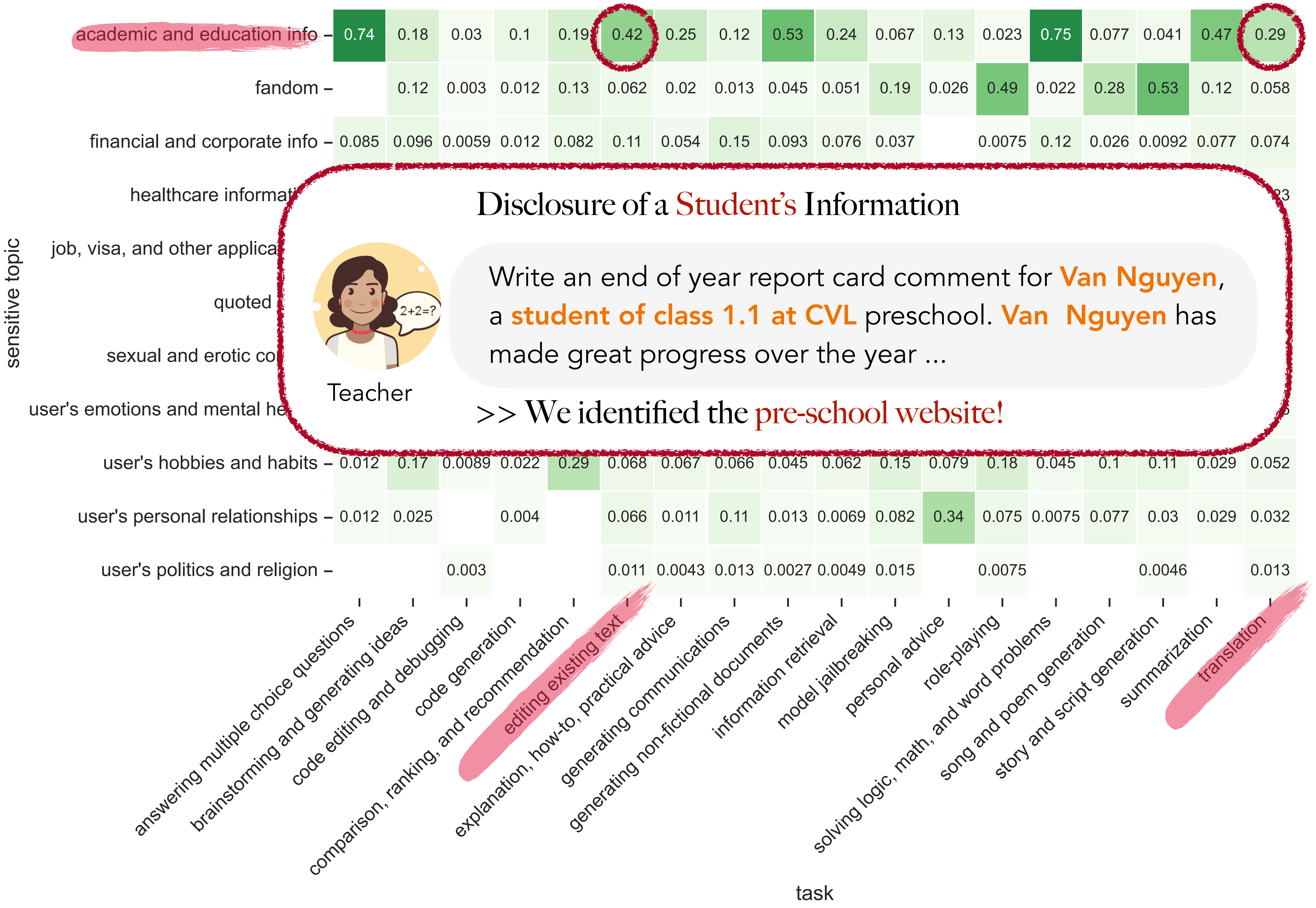
634 followers

[redacted] hasn't posted yet

[redacted] shares will be displayed here.

Show all activity →

What types of sensitive data is in there?

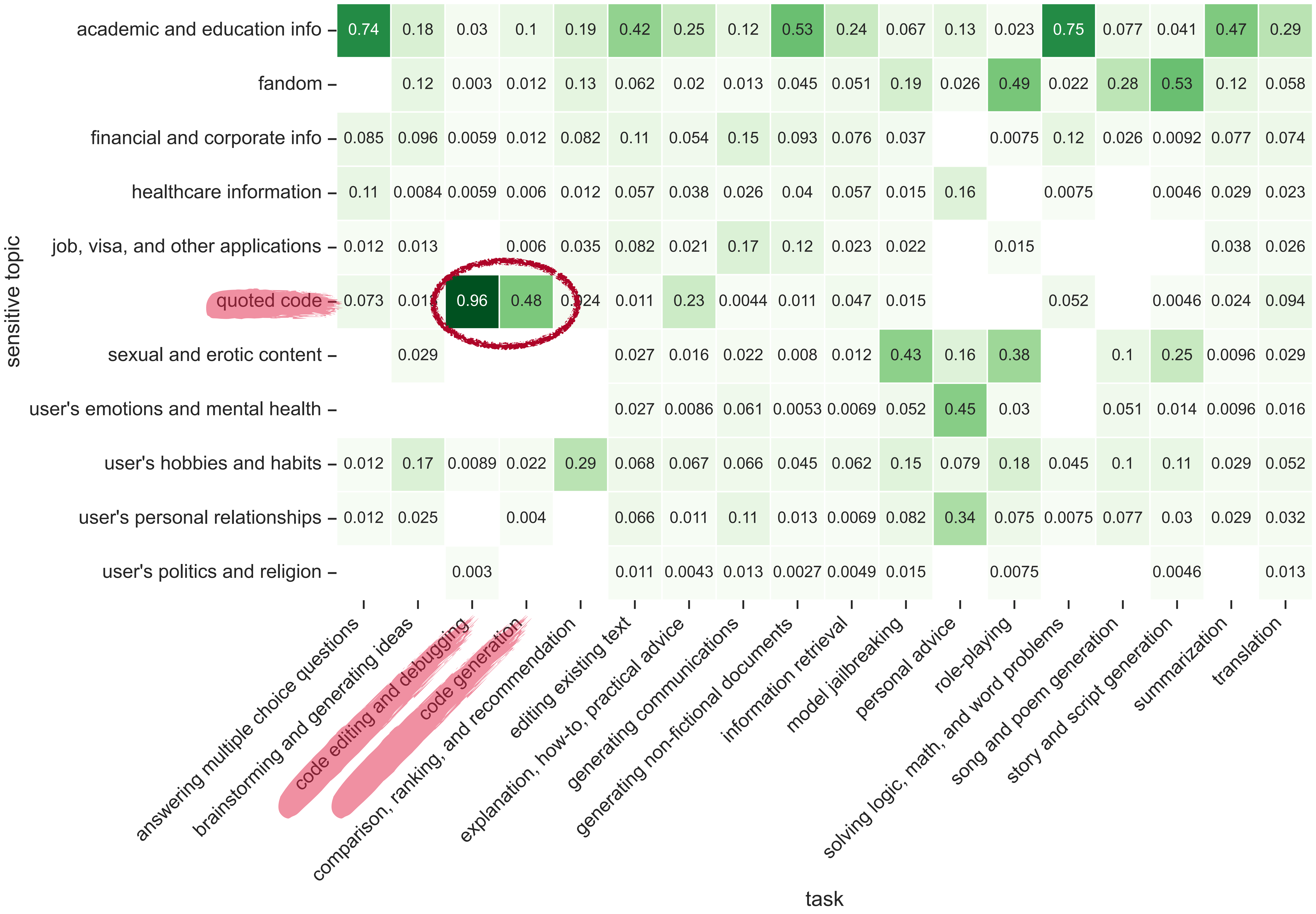


Disclosure of a Student's Information

Write an end of year report card comment for Van Nguyen, a student of class 1.1 at CVL preschool. Van Nguyen has made great progress over the year ...

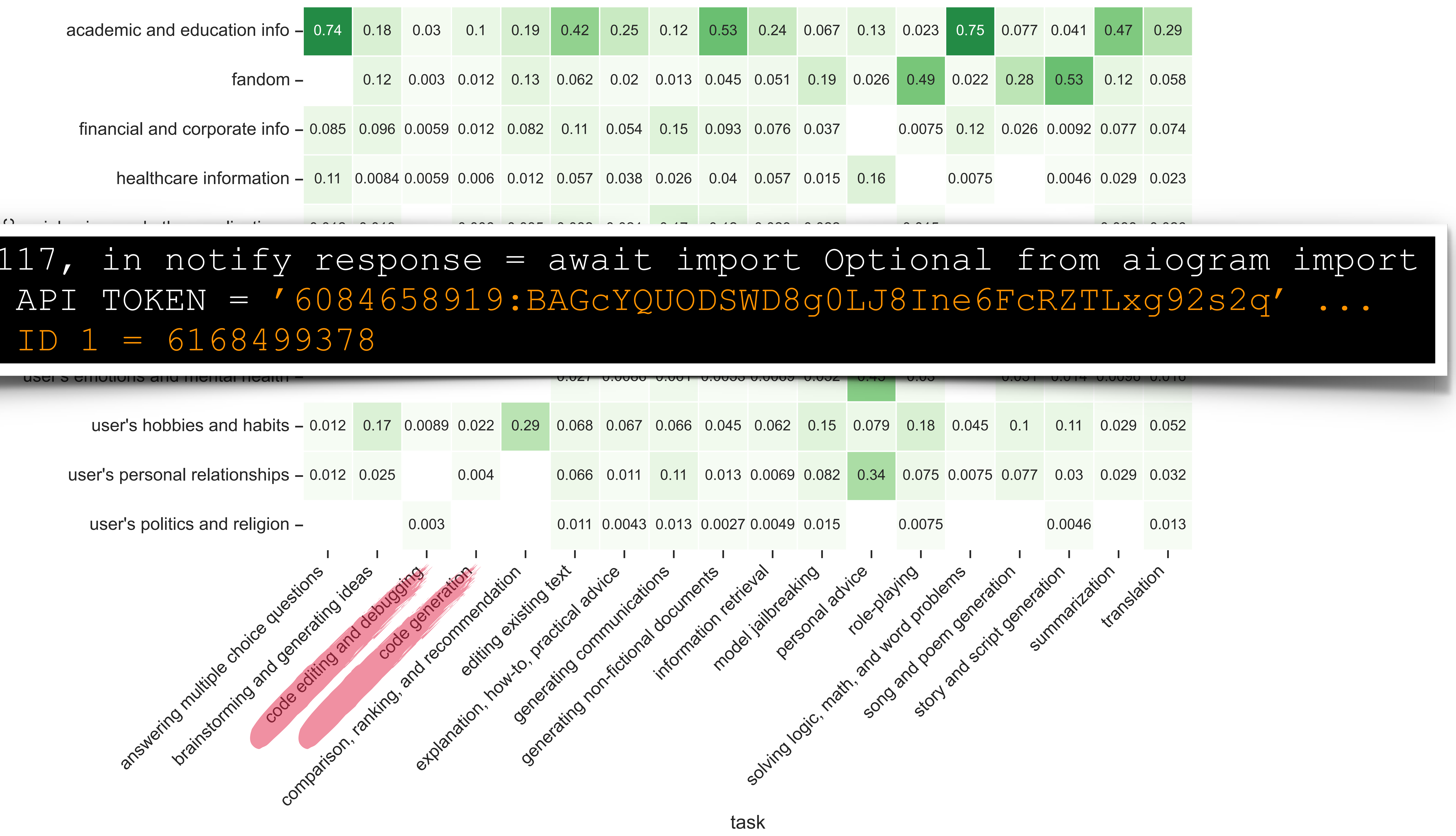
>> We identified the pre-school website!

What types of sensitive data is in there?



sensitive topic	answering multiple choice questions	brainstorming and generating ideas	code editing and debugging	code generation	comparison, ranking, and recommendation	editing existing text	explanation, how-to, practical advice	generating communications	generating non-fictional documents	information retrieval	model jailbreaking	personal advice	role-playing	solving logic, math, and word problems	song and poem generation	story and script generation	summarization	translation
academic and education info	0.74	0.18	0.03	0.1	0.19	0.42	0.25	0.12	0.53	0.24	0.067	0.13	0.023	0.75	0.077	0.041	0.47	0.29
fandom		0.12	0.003	0.012	0.13	0.062	0.02	0.013	0.045	0.051	0.19	0.026	0.49	0.022	0.28	0.53	0.12	0.058
financial and corporate info	0.085	0.096	0.0059	0.012	0.082	0.11	0.054	0.15	0.093	0.076	0.037		0.0075	0.12	0.026	0.0092	0.077	0.074
healthcare information	0.11	0.0084	0.0059	0.006	0.012	0.057	0.038	0.026	0.04	0.057	0.015	0.16		0.0075		0.0046	0.029	0.023
job, visa, and other applications	0.012	0.013		0.006	0.035	0.082	0.021	0.17	0.12	0.023	0.022		0.015				0.038	0.026
quoted code	0.073	0.013	0.96	0.48	0.024	0.011	0.23	0.0044	0.011	0.047	0.015			0.052		0.0046	0.024	0.094
sexual and erotic content		0.029				0.027	0.016	0.022	0.008	0.012	0.43	0.16	0.38		0.1	0.25	0.0096	0.029
user's emotions and mental health						0.027	0.0086	0.061	0.0053	0.0069	0.052	0.45	0.03		0.051	0.014	0.0096	0.016
user's hobbies and habits	0.012	0.17	0.0089	0.022	0.29	0.068	0.067	0.066	0.045	0.062	0.15	0.079	0.18	0.045	0.1	0.11	0.029	0.052
user's personal relationships	0.012	0.025		0.004		0.066	0.011	0.11	0.013	0.0069	0.082	0.34	0.075	0.0075	0.077	0.03	0.029	0.032
user's politics and religion			0.003			0.011	0.0043	0.013	0.0027	0.0049	0.015		0.0075			0.0046		0.013

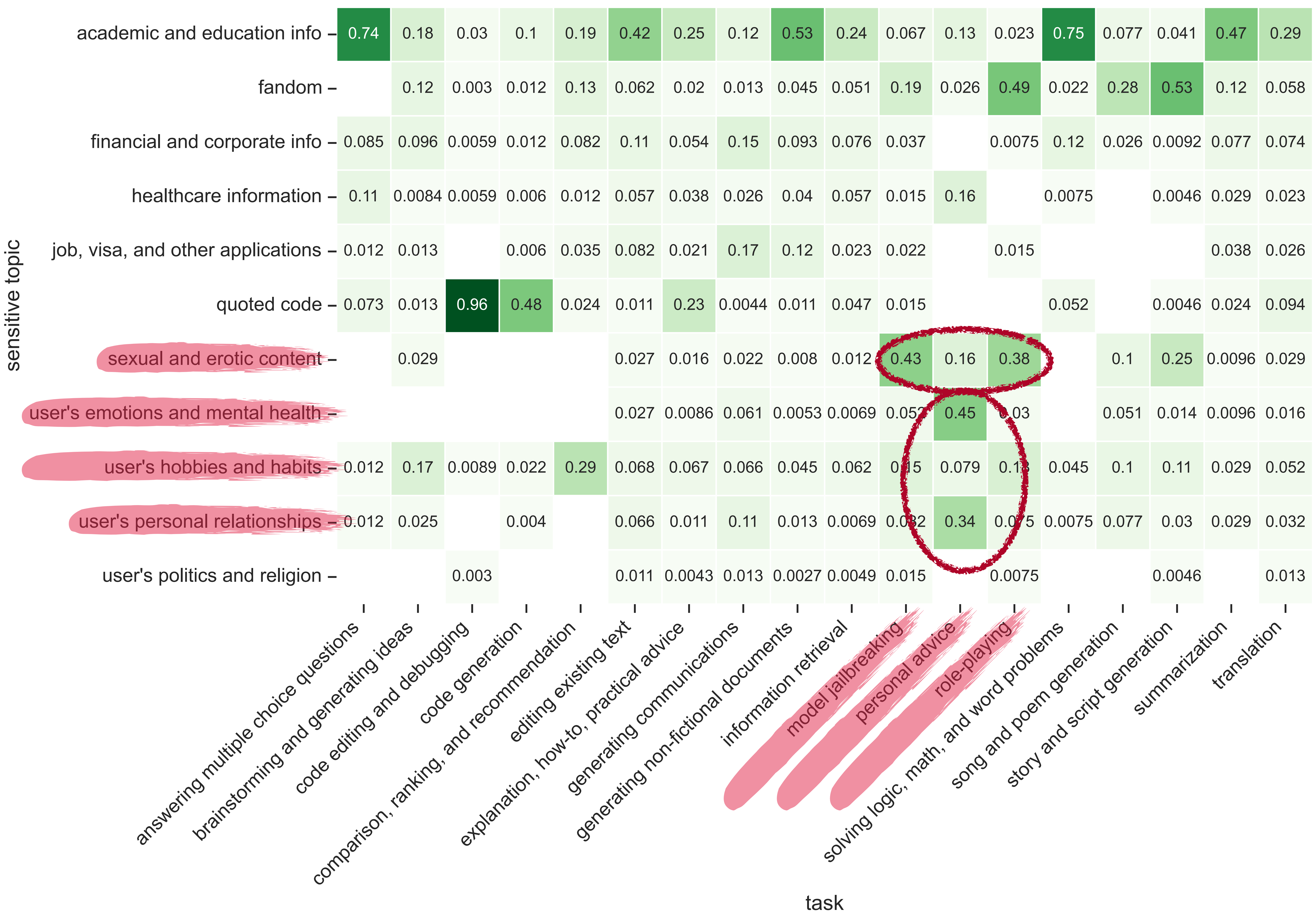
What types of sensitive data is in there?



	answering multiple choice questions	brainstorming and generating ideas	code editing and debugging	code generation	comparison, ranking, and recommendation	editing existing text	explanation, how-to, practical advice	generating communications	generating non-fictional documents	information retrieval	model jailbreaking	personal advice	role-playing	solving logic, math, and word problems	song and poem generation	story and script generation	summarization	translation
academic and education info	0.74	0.18	0.03	0.1	0.19	0.42	0.25	0.12	0.53	0.24	0.067	0.13	0.023	0.75	0.077	0.041	0.47	0.29
fandom		0.12	0.003	0.012	0.13	0.062	0.02	0.013	0.045	0.051	0.19	0.026	0.49	0.022	0.28	0.53	0.12	0.058
financial and corporate info	0.085	0.096	0.0059	0.012	0.082	0.11	0.054	0.15	0.093	0.076	0.037		0.0075	0.12	0.026	0.0092	0.077	0.074
healthcare information	0.11	0.0084	0.0059	0.006	0.012	0.057	0.038	0.026	0.04	0.057	0.015	0.16		0.0075		0.0046	0.029	0.023
user's hobbies and habits	0.012	0.17	0.0089	0.022	0.29	0.068	0.067	0.066	0.045	0.062	0.15	0.079	0.18	0.045	0.1	0.11	0.029	0.052
user's personal relationships	0.012	0.025		0.004		0.066	0.011	0.11	0.013	0.0069	0.082	0.34	0.075	0.0075	0.077	0.03	0.029	0.032
user's politics and religion			0.003			0.011	0.0043	0.013	0.0027	0.0049	0.015		0.0075			0.0046		0.013

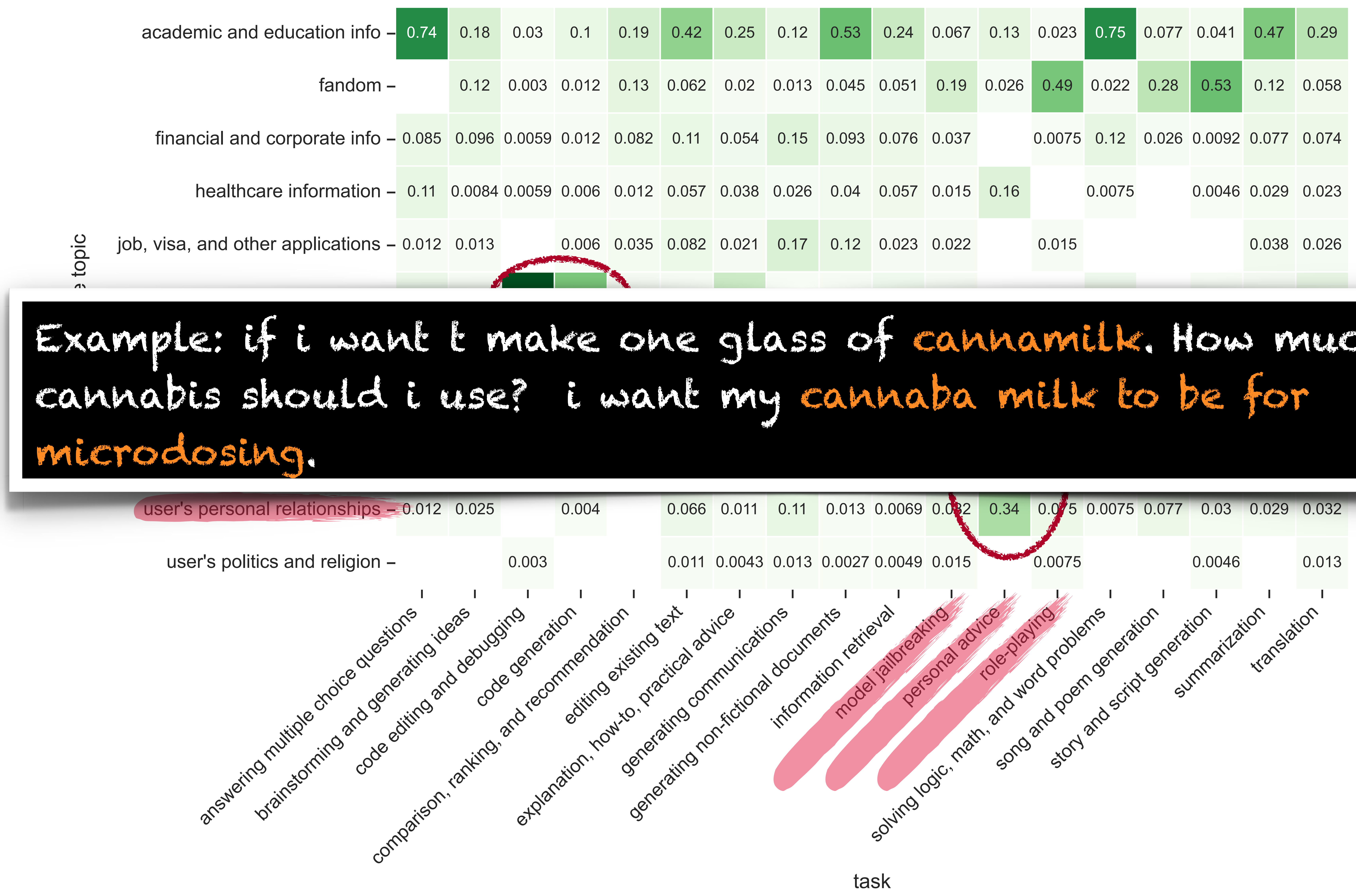
```
line 117, in notify response = await import Optional from aiogram import
types API TOKEN = ’6084658919:BAGcYQUODSWD8g0LJ8Ine6FcRZTLxg92s2q’ ...
ADMIN ID 1 = 6168499378
```

What types of sensitive data is in there?



sensitive topic	answering multiple choice questions	brainstorming and generating ideas	code editing and debugging	code generation	comparison, ranking, and recommendation	editing existing text	explanation, how-to, practical advice	generating communications	generating non-fictional documents	information retrieval	model jailbreaking	personal advice	role-playing	solving logic, math, and word problems	song and poem generation	story and script generation	summarization	translation
academic and education info	0.74	0.18	0.03	0.1	0.19	0.42	0.25	0.12	0.53	0.24	0.067	0.13	0.023	0.75	0.077	0.041	0.47	0.29
fandom		0.12	0.003	0.012	0.13	0.062	0.02	0.013	0.045	0.051	0.19	0.026	0.49	0.022	0.28	0.53	0.12	0.058
financial and corporate info	0.085	0.096	0.0059	0.012	0.082	0.11	0.054	0.15	0.093	0.076	0.037		0.0075	0.12	0.026	0.0092	0.077	0.074
healthcare information	0.11	0.0084	0.0059	0.006	0.012	0.057	0.038	0.026	0.04	0.057	0.015	0.16		0.0075		0.0046	0.029	0.023
job, visa, and other applications	0.012	0.013		0.006	0.035	0.082	0.021	0.17	0.12	0.023	0.022		0.015				0.038	0.026
quoted code	0.073	0.013	0.96	0.48	0.024	0.011	0.23	0.0044	0.011	0.047	0.015			0.052		0.0046	0.024	0.094
sexual and erotic content		0.029				0.027	0.016	0.022	0.008	0.012	0.43	0.16	0.38		0.1	0.25	0.0096	0.029
user's emotions and mental health						0.027	0.0086	0.061	0.0053	0.0069	0.052	0.45	0.03		0.051	0.014	0.0096	0.016
user's hobbies and habits	0.012	0.17	0.0089	0.022	0.29	0.068	0.067	0.066	0.045	0.062	0.15	0.079	0.13	0.045	0.1	0.11	0.029	0.052
user's personal relationships	0.012	0.025		0.004		0.066	0.011	0.11	0.013	0.0069	0.032	0.34	0.075	0.0075	0.077	0.03	0.029	0.032
user's politics and religion			0.003			0.011	0.0043	0.013	0.0027	0.0049	0.015		0.0075			0.0046		0.013

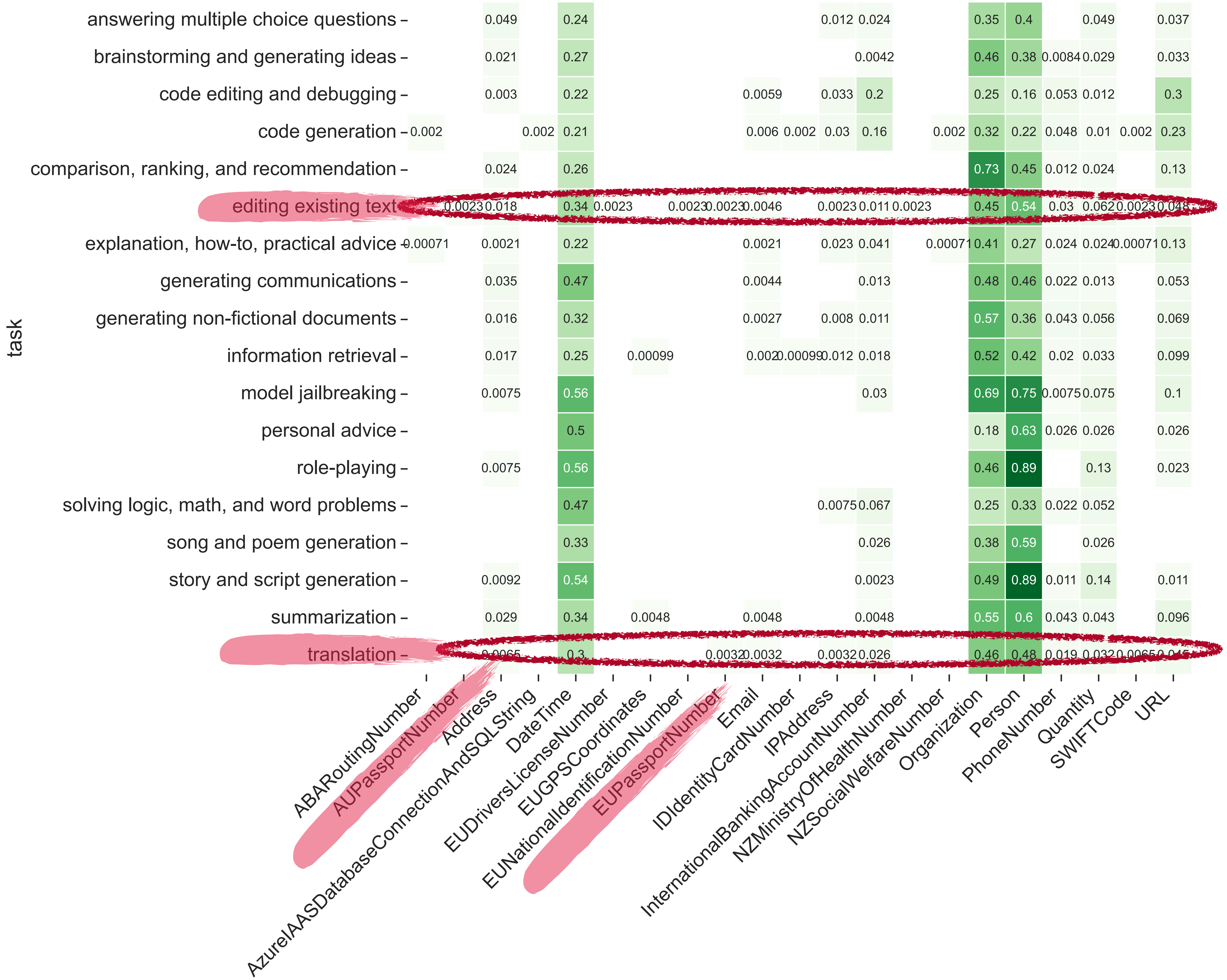
What types of sensitive data is in there?



	answering multiple choice questions	brainstorming and generating ideas	code editing and debugging	code generation	comparison, ranking, and recommendation	editing existing text	explanation, how-to, practical advice	generating communications	generating non-fictional documents	information retrieval	model jailbreaking	personal advice	role-playing	solving logic, math, and word problems	song and poem generation	story and script generation	summarization	translation
academic and education info	0.74	0.18	0.03	0.1	0.19	0.42	0.25	0.12	0.53	0.24	0.067	0.13	0.023	0.75	0.077	0.041	0.47	0.29
fandom		0.12	0.003	0.012	0.13	0.062	0.02	0.013	0.045	0.051	0.19	0.026	0.49	0.022	0.28	0.53	0.12	0.058
financial and corporate info	0.085	0.096	0.0059	0.012	0.082	0.11	0.054	0.15	0.093	0.076	0.037		0.0075	0.12	0.026	0.0092	0.077	0.074
healthcare information	0.11	0.0084	0.0059	0.006	0.012	0.057	0.038	0.026	0.04	0.057	0.015	0.16		0.0075		0.0046	0.029	0.023
job, visa, and other applications	0.012	0.013		0.006	0.035	0.082	0.021	0.17	0.12	0.023	0.022		0.015				0.038	0.026
user's personal relationships	0.012	0.025		0.004		0.066	0.011	0.11	0.013	0.0069	0.032	0.34	0.0075	0.0075	0.077	0.03	0.029	0.032
user's politics and religion			0.003			0.011	0.0043	0.013	0.0027	0.0049	0.015		0.0075			0.0046		0.013

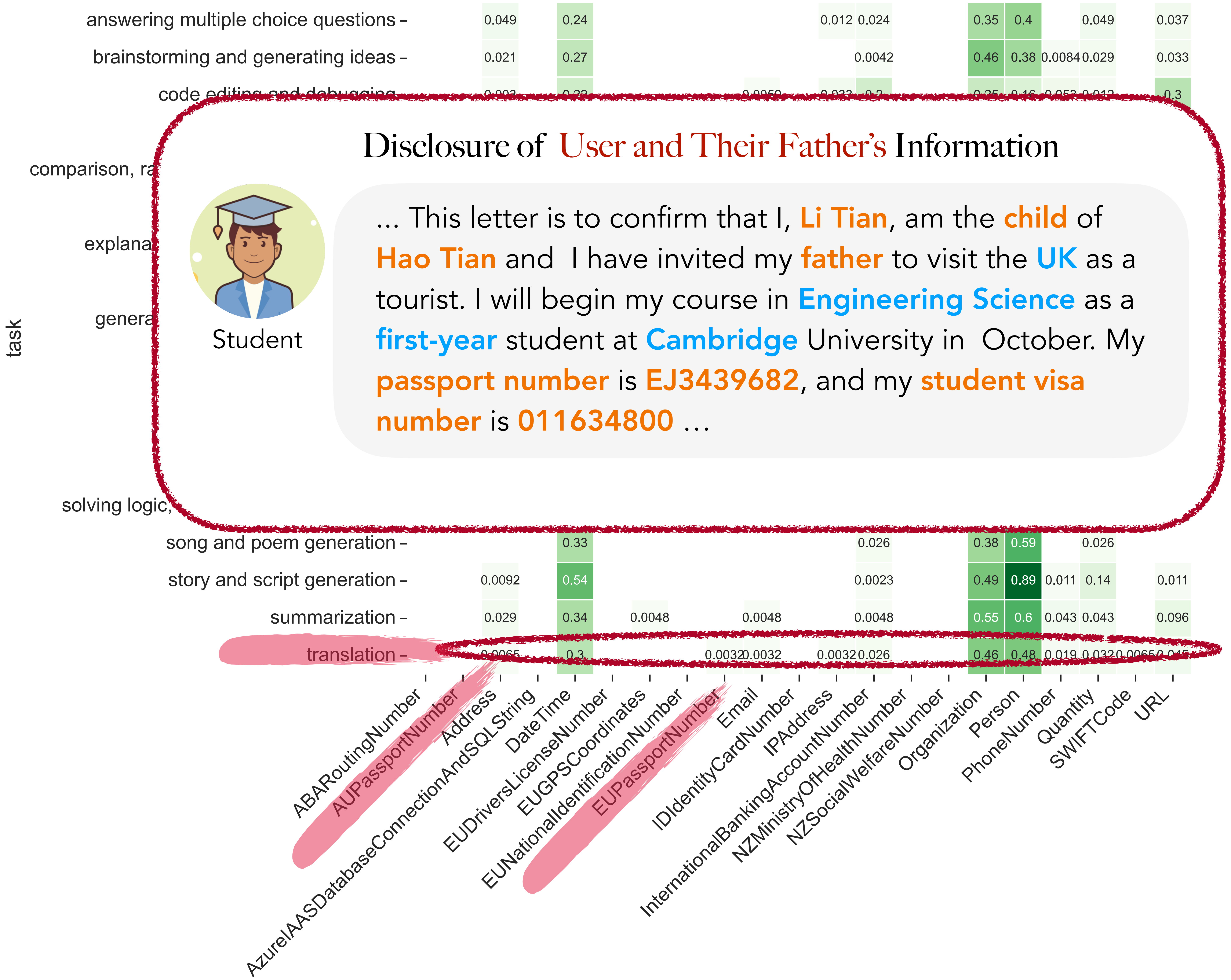
Example: if i want t make one glass of cannamilk. How much cannabis should i use? i want my cannaba milk to be for microdosing.

What types of PII do we see?



task	ABARoutingNumber	AUPassportNumber	Address	AzureIAASDatabaseConnectionAndSQLString	DateTime	EUDriversLicenseNumber	EUGPSCoordinates	EUNationalIdentificationNumber	EUPassportNumber	Email	IDIdentityCardNumber	IPAddress	InternationalBankingAccountNumber	NZMinistryOfHealthNumber	NZSocialWelfareNumber	Organization	Person	PhoneNumber	Quantity	SWIFTCode	URL
answering multiple choice questions			0.049		0.24							0.012	0.024			0.35	0.4		0.049		0.037
brainstorming and generating ideas			0.021		0.27								0.0042			0.46	0.38	0.0084	0.029		0.033
code editing and debugging			0.003		0.22					0.0059		0.033	0.2			0.25	0.16	0.053	0.012		0.3
code generation	0.002			0.002	0.21					0.006	0.002	0.03	0.16		0.002	0.32	0.22	0.048	0.01	0.002	0.23
comparison, ranking, and recommendation			0.024		0.26											0.73	0.45	0.012	0.024		0.13
editing existing text		0.0023	0.018		0.34	0.0023		0.0023	0.0023	0.0046		0.0023	0.011	0.0023		0.45	0.54	0.03	0.062	0.0023	0.048
explanation, how-to, practical advice	0.00071		0.0021		0.22					0.0021		0.023	0.041		0.00071	0.41	0.27	0.024	0.024	0.00071	0.13
generating communications			0.035		0.47					0.0044			0.013			0.48	0.46	0.022	0.013		0.053
generating non-fictional documents			0.016		0.32					0.0027		0.008	0.011			0.57	0.36	0.043	0.056		0.069
information retrieval			0.017		0.25		0.00099			0.002	0.00099	0.012	0.018			0.52	0.42	0.02	0.033		0.099
model jailbreaking			0.0075		0.56								0.03			0.69	0.75	0.0075	0.075		0.1
personal advice					0.5											0.18	0.63	0.026	0.026		0.026
role-playing			0.0075		0.56											0.46	0.89		0.13		0.023
solving logic, math, and word problems					0.47							0.0075	0.067			0.25	0.33	0.022	0.052		
song and poem generation					0.33								0.026			0.38	0.59		0.026		
story and script generation			0.0092		0.54								0.0023			0.49	0.89	0.011	0.14		0.011
summarization			0.029		0.34		0.0048			0.0048			0.0048			0.55	0.6	0.043	0.043		0.096
translation			0.0065		0.3				0.0032	0.0032		0.0032	0.026			0.46	0.48	0.019	0.032	0.0065	0.045

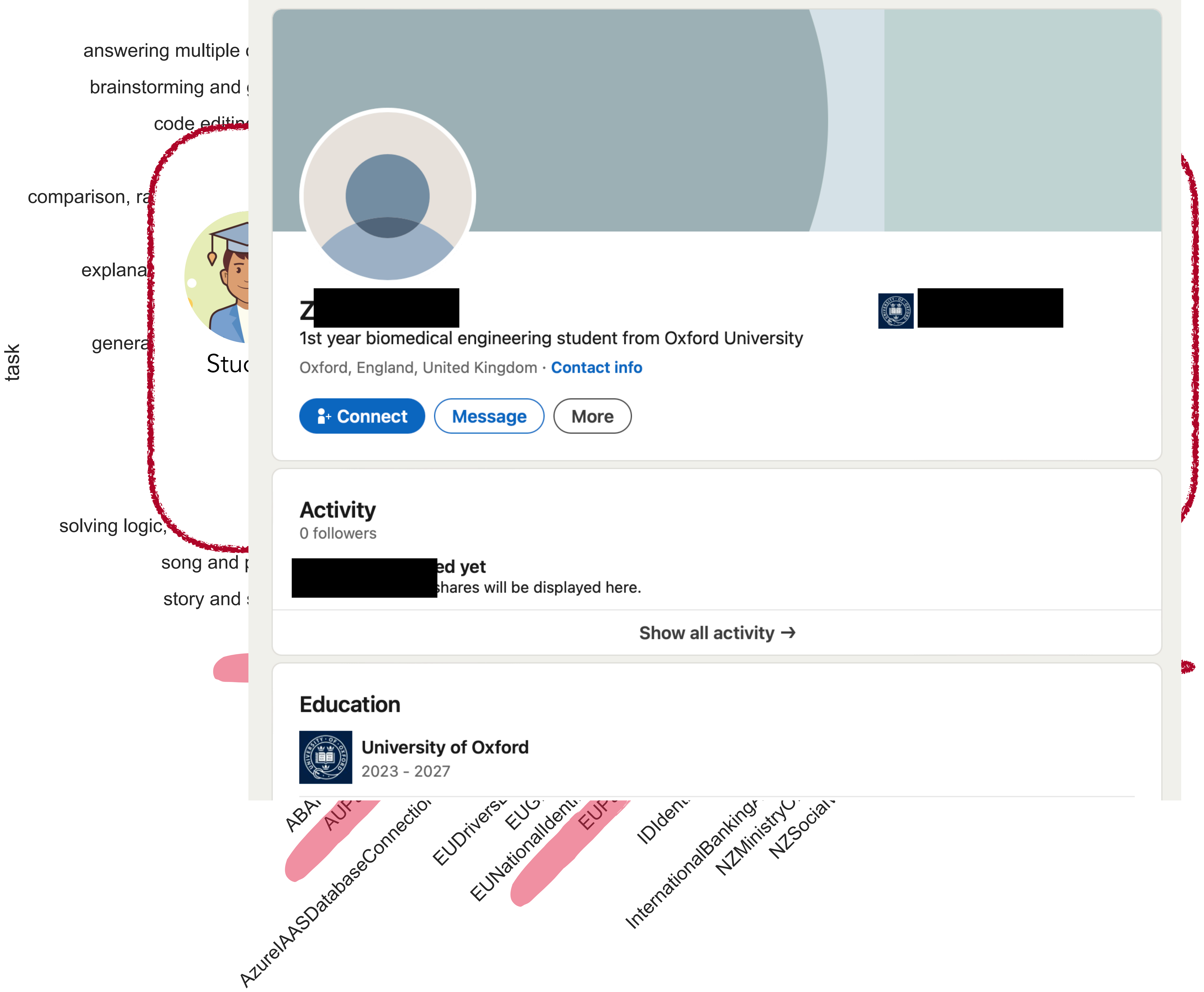
What types of PII do we see?



Disclosure of User and Their Father's Information

... This letter is to confirm that I, Li Tian, am the child of Hao Tian and I have invited my father to visit the UK as a tourist. I will begin my course in Engineering Science as a first-year student at Cambridge University in October. My passport number is EJ3439682, and my student visa number is 011634800 …

What types of PII do we see?



Z[redacted]

[redacted]

1st year biomedical engineering student from Oxford University

Oxford, England, United Kingdom · Contact info

Connect | Message | More

Activity

0 followers

[redacted]ed yet
hares will be displayed here.

Show all activity →

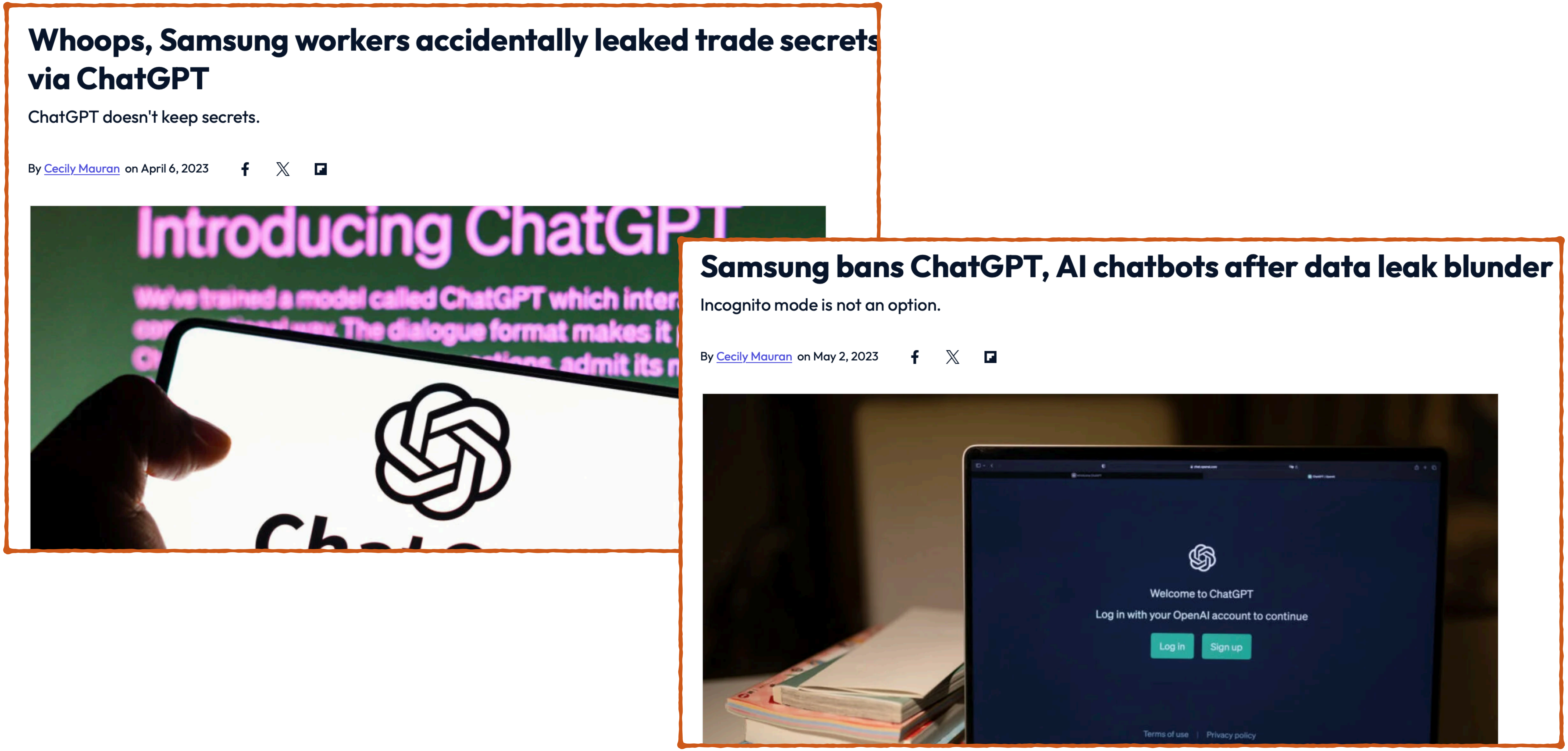
Education

University of Oxford
2023 - 2027

Summary of stats:

- **21%** of the queries include what is identified as **sensitive information**
- **Text editing or writing tasks** (CV editing, letter/email/statement generation) do overall **contain the bulk of PII, 34.0%**
- One surprising category with PIIs is the task **translation (6.6%)**
- Another common category of queries w/ PII is **code editing (20.4%)**

Mistakes happen all the time!



Whoops, Samsung workers accidentally leaked trade secrets via ChatGPT

ChatGPT doesn't keep secrets.

By Cecily Mauran on April 6, 2023

Introducing ChatGPT
We've trained a model called ChatGPT which inter
The dialogue format makes it
admit its m
Chat

Samsung bans ChatGPT, AI chatbots after data leak blunder

Incognito mode is not an option.

By Cecily Mauran on May 2, 2023

Welcome to ChatGPT
Log in with your OpenAI account to continue
Log in
Sign up
Terms of use | Privacy policy

What are the potential risks and impact?

- **Memorization** of legal, medical, or confidential trade secrets
- Risks of **direct data breach**
- Risk of **data being purchased by profiling companies** and ad services
- **Human reviewers from OpenAI** reading it
- **Interdependency and correlation in data**
- Fear of **employers finding out AI is being used, using AI in workflow**

Let’s zoom in on impacts of using AI **Professionally**!

Breaking News: Case Studies of Generative AI's Use in Journalism

Natalie Grace Brigham, Chongjiu Gao, Tadayoshi Kohno, Franziska Roesner, Niloofar Mireshghallah — https://arxiv.org/abs/2406.13706

Example Query to ChatGPT– WhatsApp conversation

″Hello I am a [redacted] **journalist and one woman contacted me** regarding an issue she has with the government and other stuff that the government does not provide for **her child who is disabled**. anaylse the whatsapp convo and write an article out of it. tell me if you need more information that would help give the article the human element:

Example Query to ChatGPT– WhatsApp conversation

"Hello I am a [redacted] **journalist and one woman contacted me** regarding an issue she has with the government and other stuff that the government does not provide for **her child who is disabled**. anaylse the whatsapp convo and write an article out of it. tell me if you need more information that would help give the article the human element:

Example Query to ChatGPT– WhatsApp conversation

[10:48, 06/04/2023] <PHONE_NUMBER>: no I would not like my children's photos on the article

[10:49, 06/04/2023] <PHONE_NUMBER>: And re conditions I will only mention the one who needs **to** travel overseas as it's the only one that is a visible disability cos he cannot walk

[11:23, 06/04/2023] <PHONE_NUMBER>: **I have 3 children , one is 8 and the other 2 are 4 years old , once one of our 4 year old was diagnosed with PVL a brain condition resulting in Cerebral palsy** I found myself in a new community in [redacted]t is of parents with children with disabilities who in my opinion is not supported enough [redacted]

[12:38, 06/04/2023] <PRESIDIO_ANONYMIZED_PHONE_NUMBER>: If u feel my voice is enough and no need for others at this point leave it as me only

[14:40, 06/04/2023] <PRESIDIO_ANONYMIZED_PHONE_NUMBER>[redacted]**Jones**

[14:40, 06/04/2023] <PRESIDIO_ANONYMIZED_PHONE_NUMBER>: This mother is also interested to share info

Example Query to ChatGPT– WhatsApp conversation

"Hello I

one wom

issue she

other stu

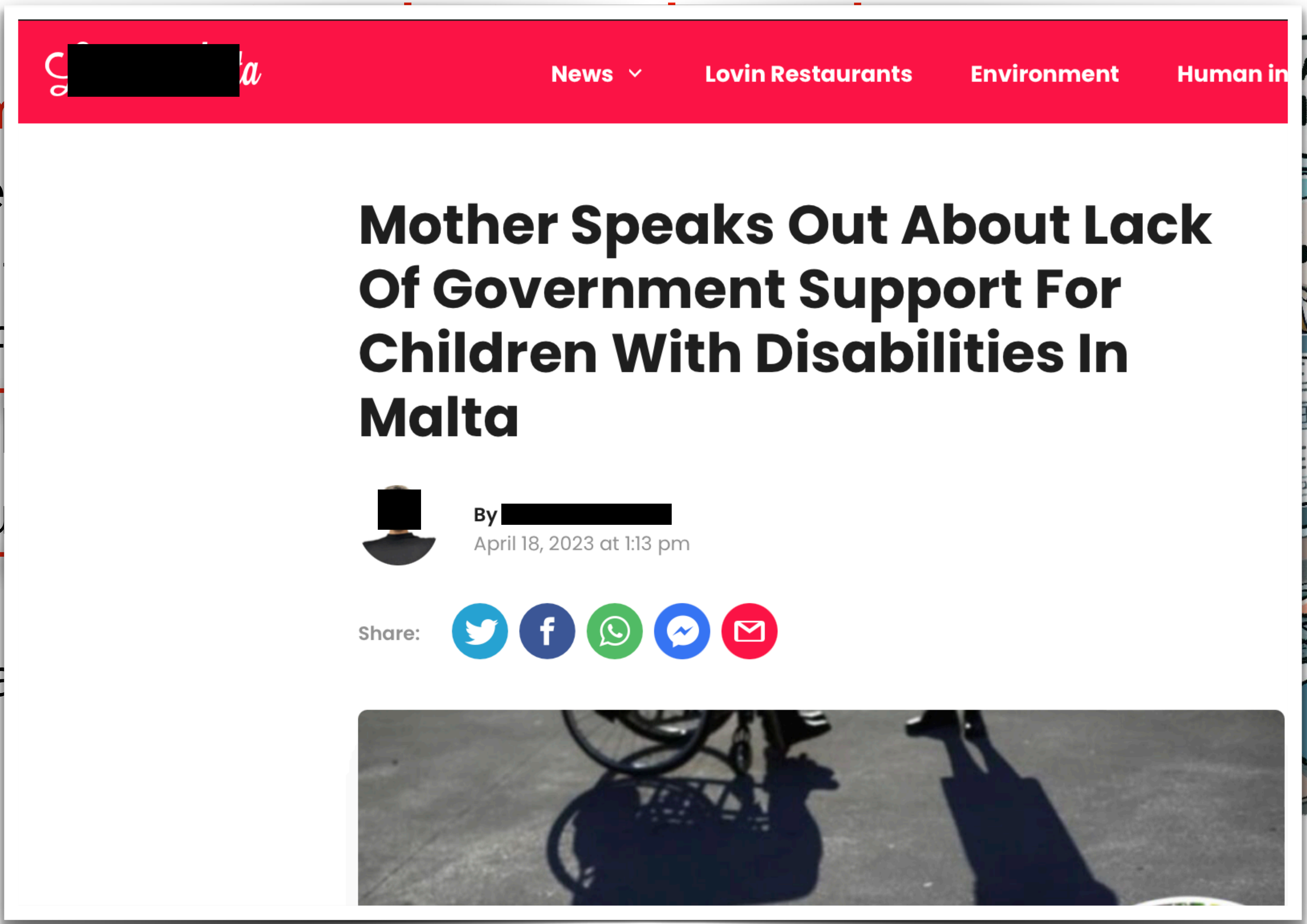
provide f

anaylse t

article ou

informati

the huma



[redacted]

News ⌄ Lovin Restaurants Environment Human in

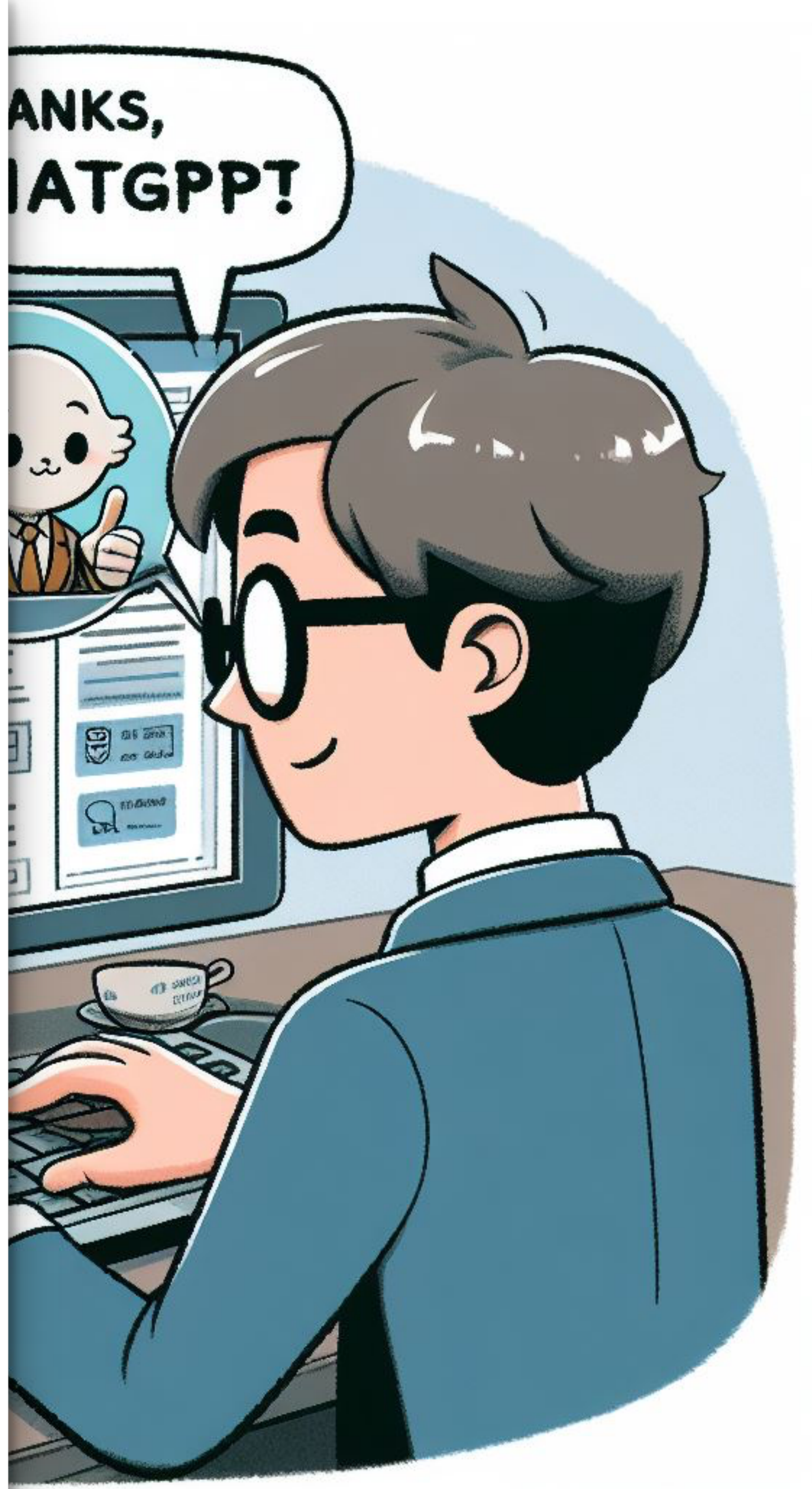
Mother Speaks Out About Lack Of Government Support For Children With Disabilities In Malta

[redacted]

By [redacted]

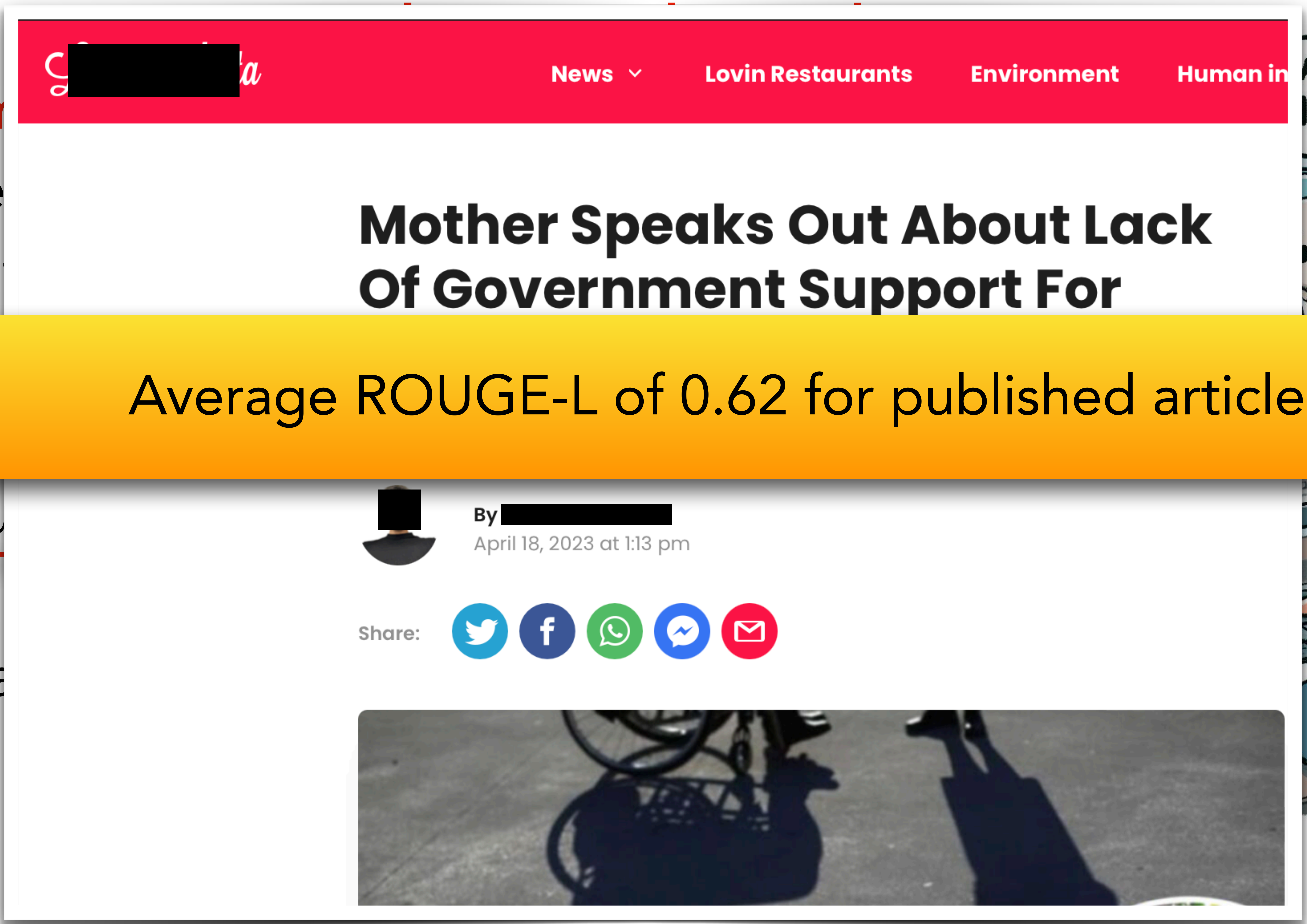
April 18, 2023 at 1:13 pm

Share:



Example Query to ChatGPT– WhatsApp conversation

“Hello I
one wom
issue she
other stu
provic
anayls
article ou
informati
the huma



[redacted]

Mother Speaks Out About Lack Of Government Support For

[redacted]

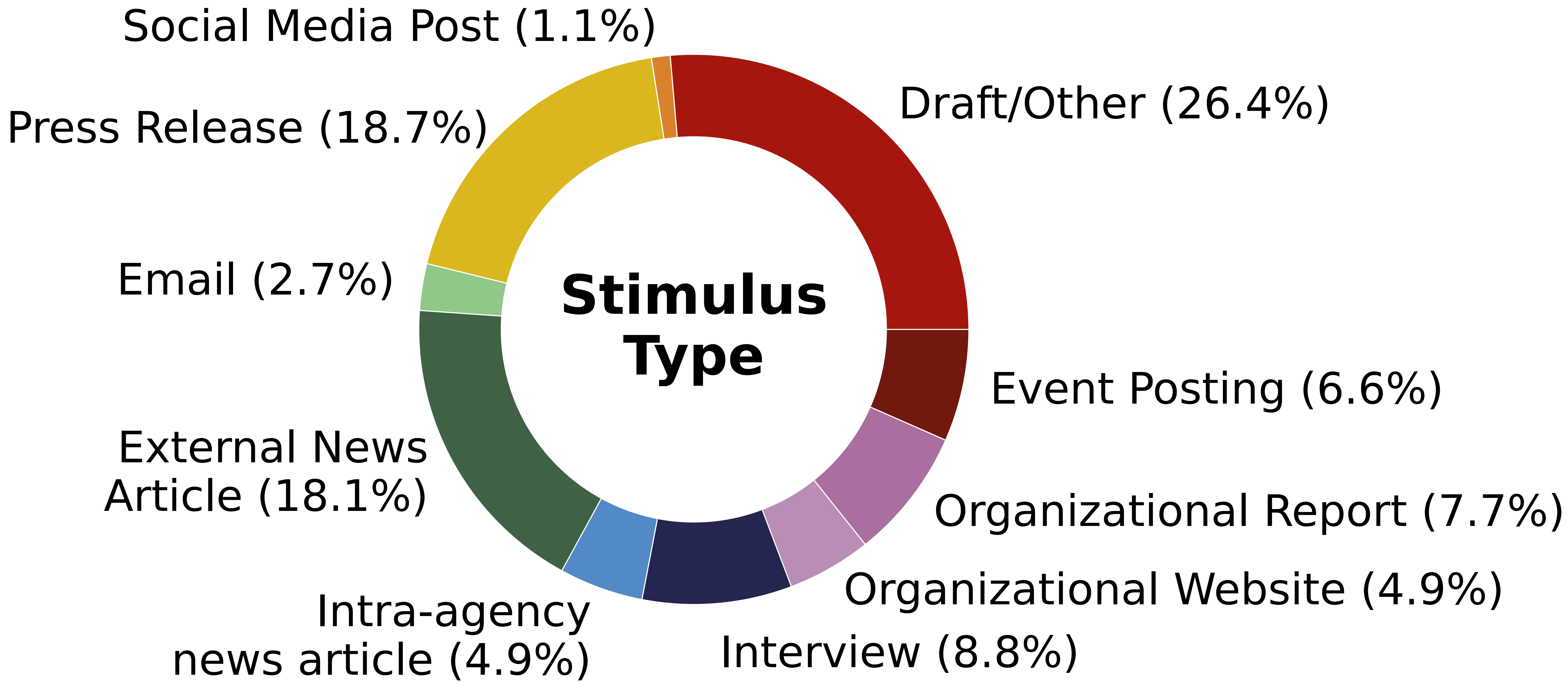
By [redacted]
April 18, 2023 at 1:13 pm

Average ROUGE-L of 0.62 for published articles

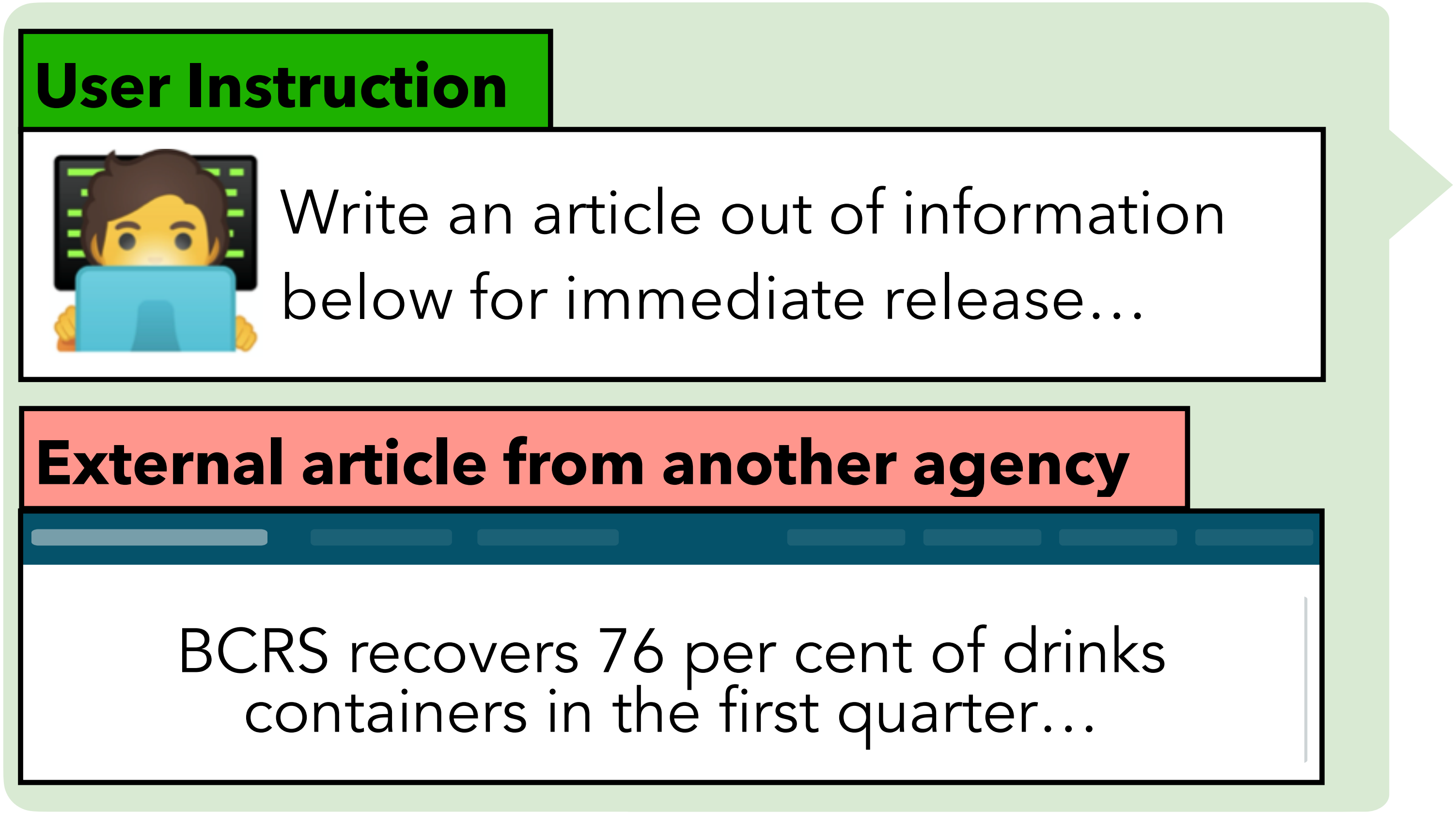
Are there more of such cases?

How do Journalists use ChatGPT?

What do journalists prompt LLMs with?



What is the article generation pipeline like?



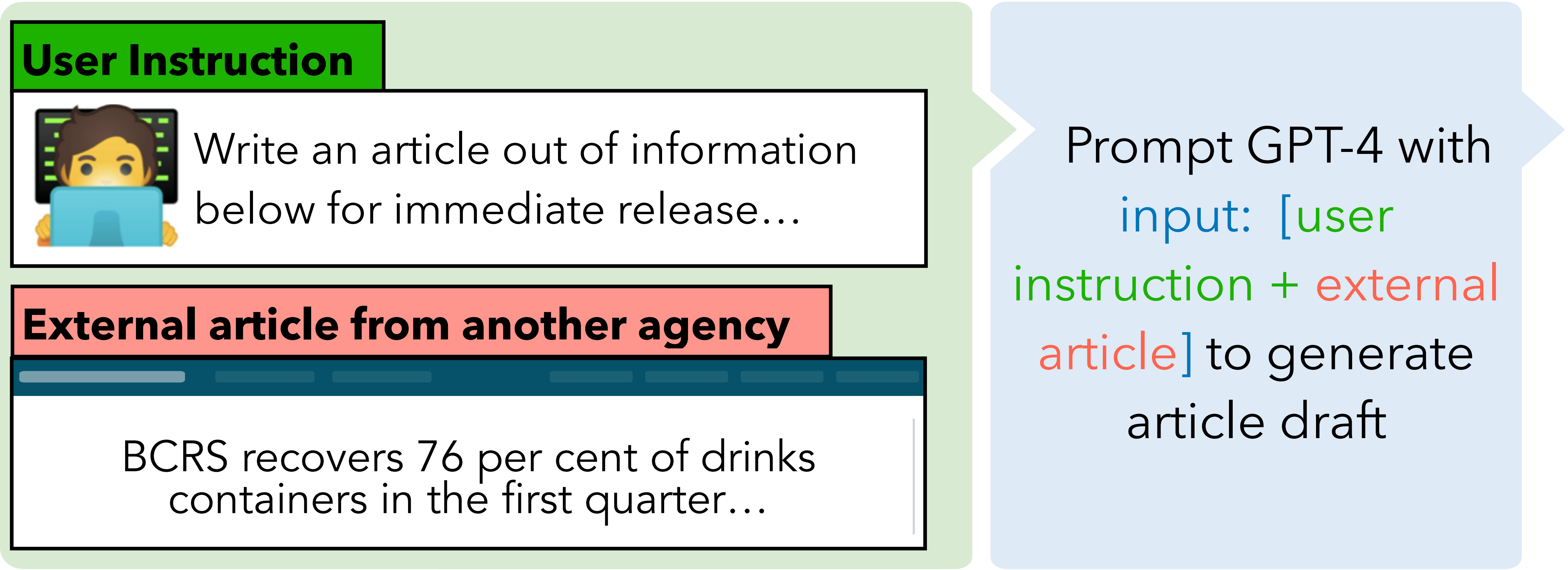
User Instruction

🧑‍💻 Write an article out of information below for immediate release…

External article from another agency

BCRS recovers 76 per cent of drinks containers in the first quarter…

What is the article generation pipeline like?



User Instruction

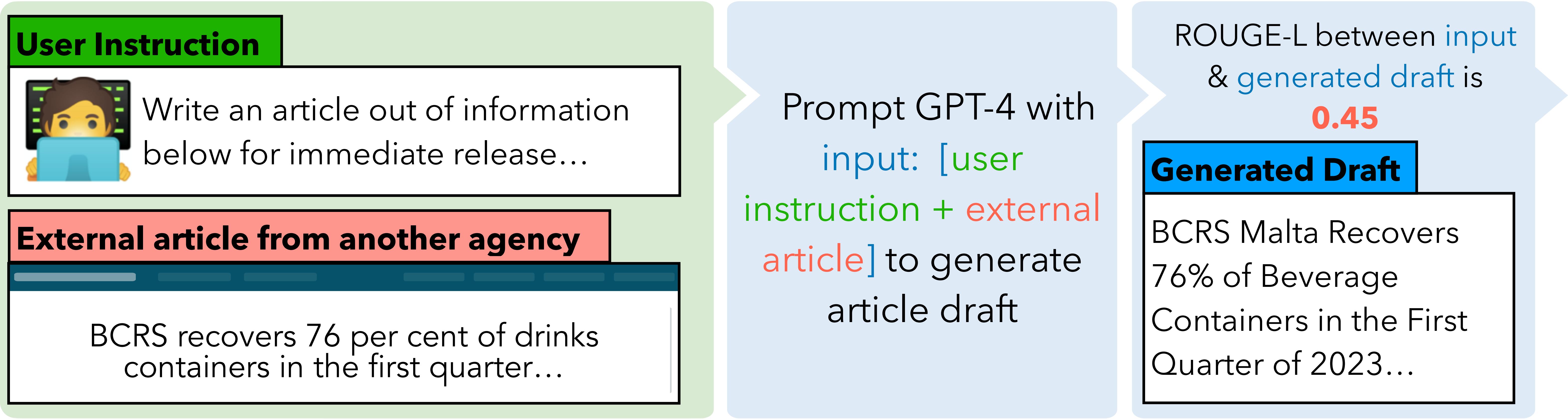
Write an article out of information below for immediate release…

External article from another agency

BCRS recovers 76 per cent of drinks containers in the first quarter…

Prompt GPT-4 with input: [user instruction + external article] to generate article draft

What is the article generation pipeline like?



User Instruction

Write an article out of information below for immediate release…

External article from another agency

BCRS recovers 76 per cent of drinks containers in the first quarter…

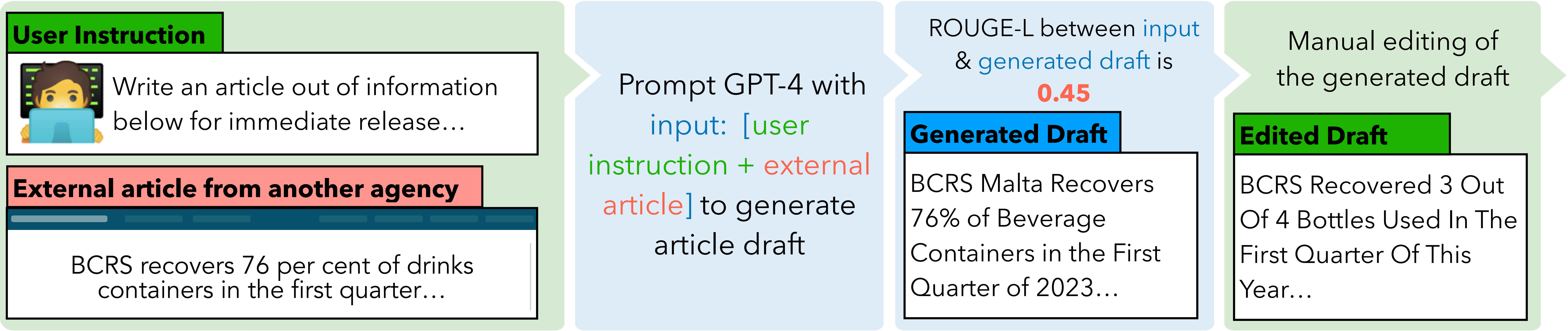
Prompt GPT-4 with input: [user instruction + external article] to generate article draft

ROUGE-L between input & generated draft is **0.45**

Generated Draft

BCRS Malta Recovers 76% of Beverage Containers in the First Quarter of 2023…

What is the article generation pipeline like?



User Instruction

Write an article out of information below for immediate release…

External article from another agency

BCRS recovers 76 per cent of drinks containers in the first quarter…

Prompt GPT-4 with input: [user instruction + external article] to generate article draft

ROUGE-L between input & generated draft is **0.45**

Generated Draft

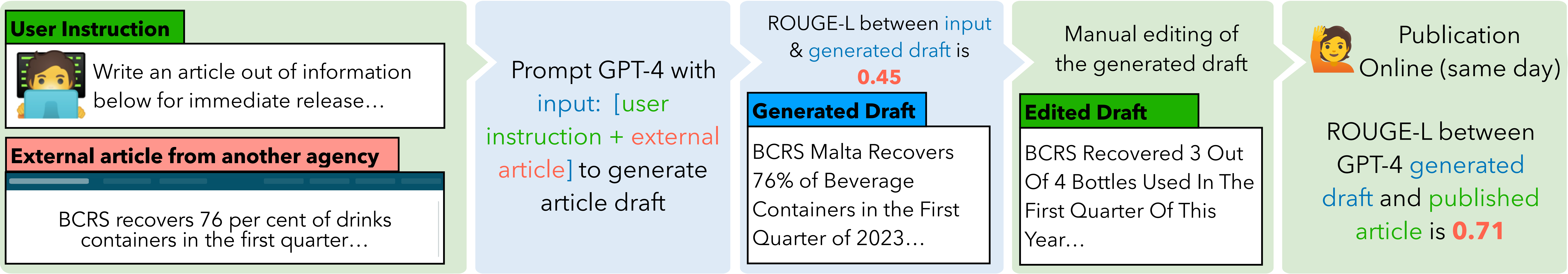
BCRS Malta Recovers 76% of Beverage Containers in the First Quarter of 2023…

Manual editing of the generated draft

Edited Draft

BCRS Recovered 3 Out Of 4 Bottles Used In The First Quarter Of This Year…

What is the article generation pipeline like?



User Instruction

🧑‍💻 Write an article out of information below for immediate release…

External article from another agency

BCRS recovers 76 per cent of drinks containers in the first quarter…

Prompt GPT-4 with input: [user instruction + external article] to generate article draft

ROUGE-L between input & generated draft is **0.45**

Generated Draft

BCRS Malta Recovers 76% of Beverage Containers in the First Quarter of 2023…

Manual editing of the generated draft

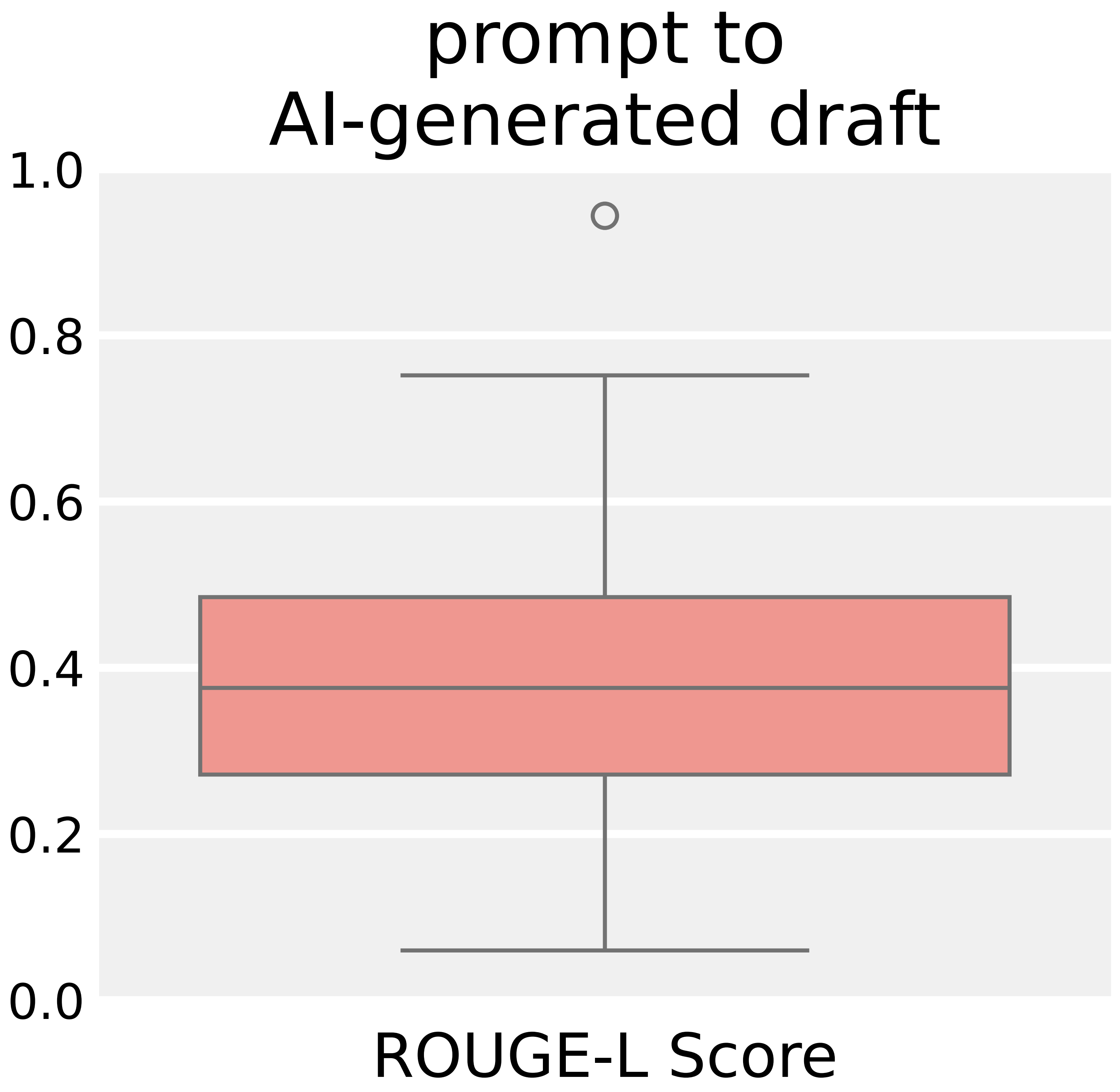
Edited Draft

BCRS Recovered 3 Out Of 4 Bottles Used In The First Quarter Of This Year…

🙋 Publication Online (same day)

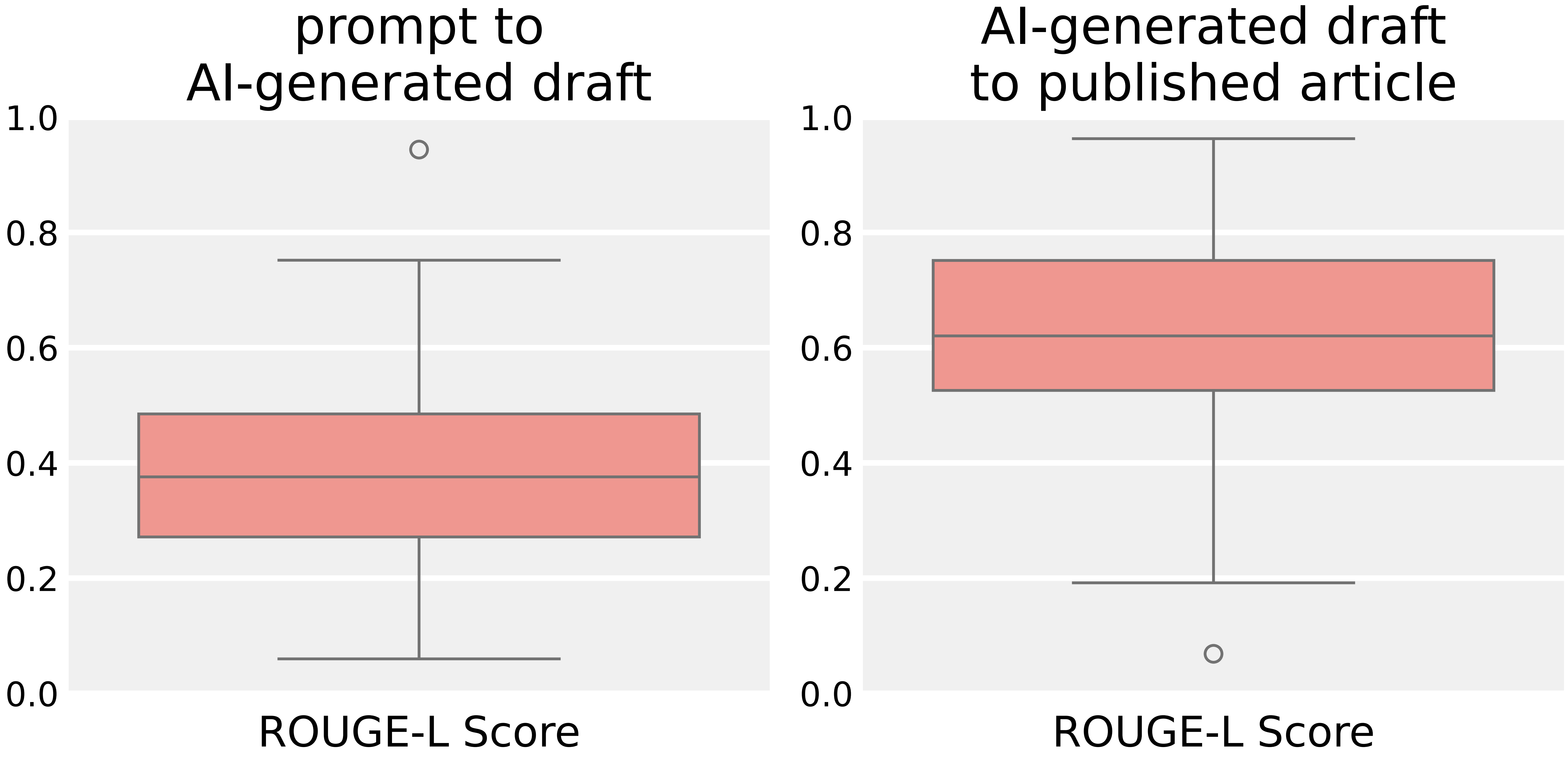
ROUGE-L between GPT-4 generated draft and published article is **0.71**

How much intervention do journalists make?



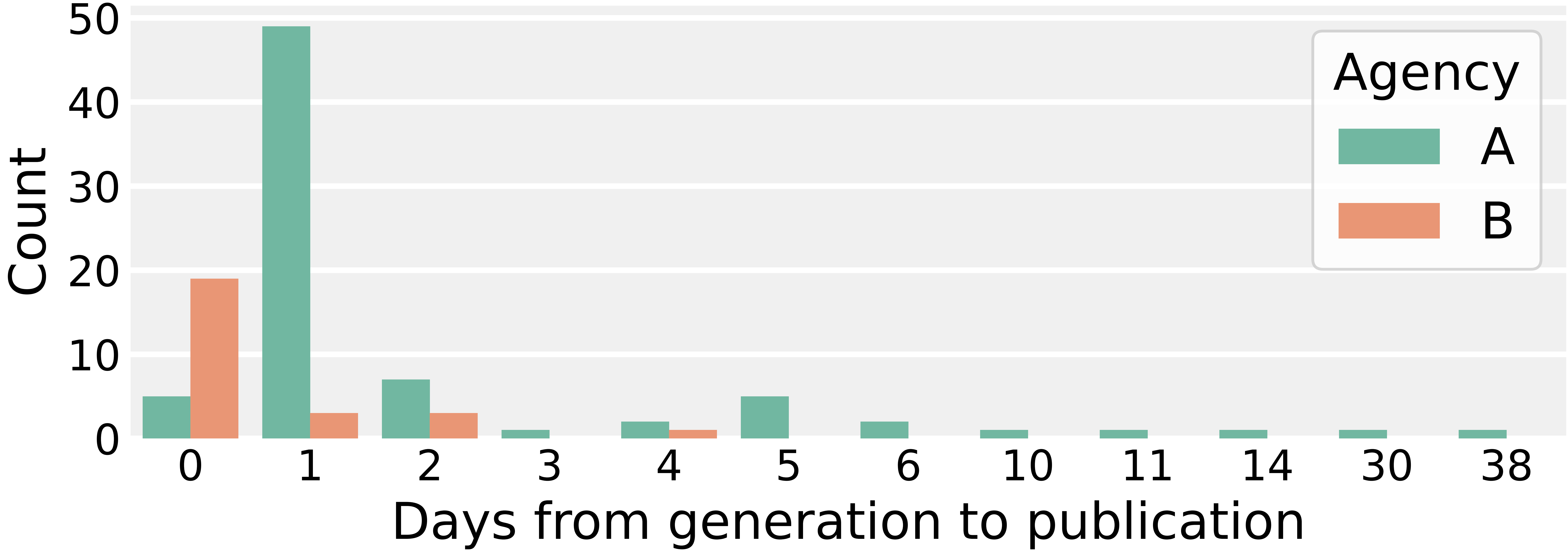
Average ROUGE-L of 0.4 for prompt and model output

How much intervention do journalists make?



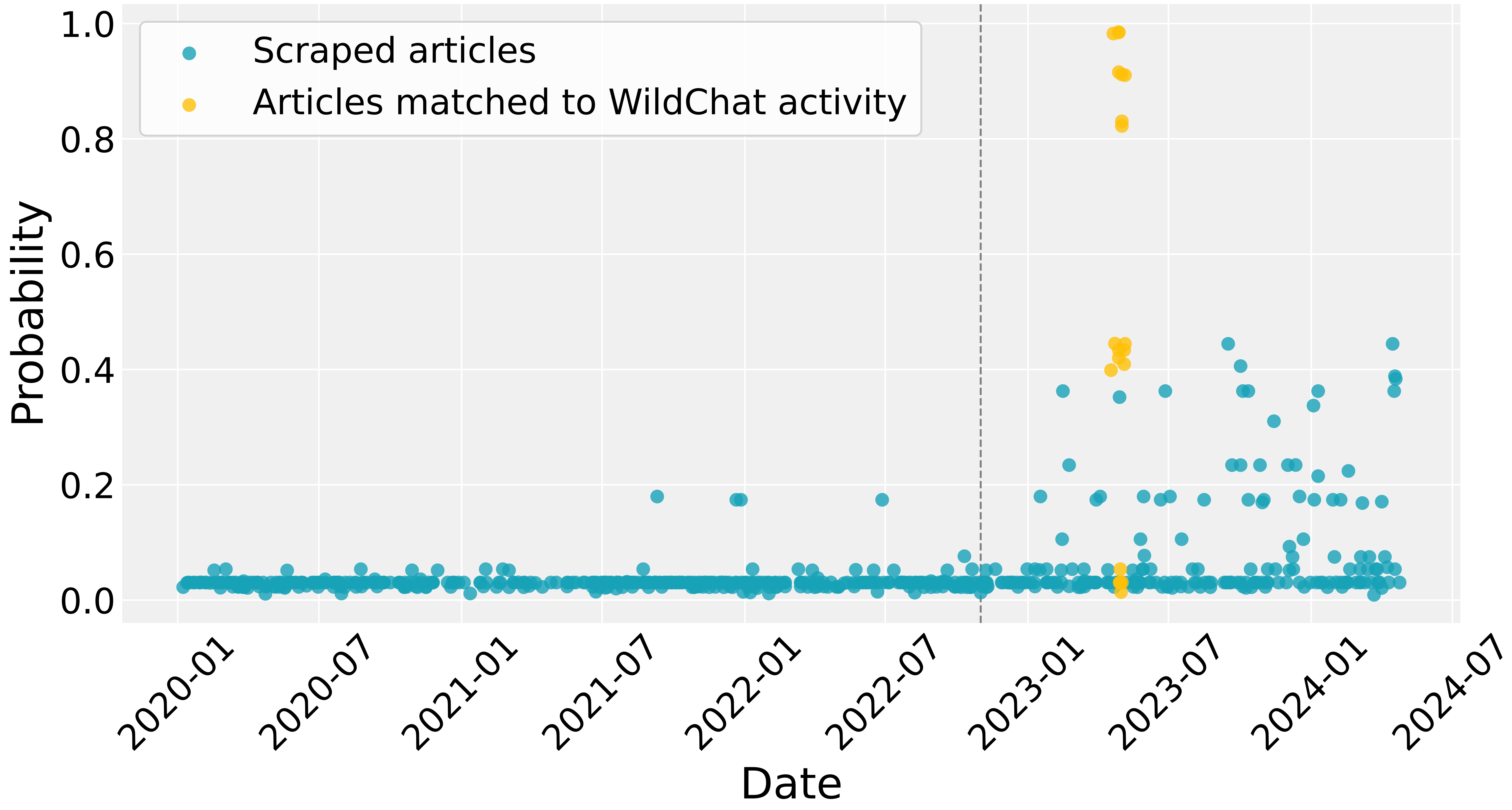
Average ROUGE-L of 0.7 for model output and matched article!

What is the prompt to publication time?



Most articles are published on the same day!

What is the trend of AI-assisted writing here?



Using ChatGPT for journalism is on the rise!

What can we do?

- Provide local, **light-weight sanitizers**
- Help people **learn their options, such as opt out!**
- Users want more **granular control!**
- People often feel comfortable because they forget they are chatting with a bot. **We need nudging mechanisms!**

Takeaways

- Users often share **very personal information about themselves, other people, and their workplaces** and schools in interactions with chatbots.
- Chatbot designers should build in **more transparency for users** about how their data is used and stored, maybe through nudging mechanisms.
- Lots more to uncover in these chat datasets, and **we need computational social scientists to dig in**.