OpenAI and Anthropic Disregard Rule Against Bots Scraping Web Content

Description

The top two AI startups in the world, OpenAI and Anthropic, are **disregarding** requests from media publishers to stop **scraping** their web content for free model training data, according to Business Insider. These companies have been found to be either ignoring or **bypassing** the established web rule robots.txt, which is meant to prevent **automated scraping** of websites.

A startup called TollBit, which aims to facilitate paid licensing deals between publishers and AI companies, discovered that several AI companies, including OpenAI and Anthropic, are not adhering to robots.txt rules. In a letter sent to large publishers on Friday, TollBit highlighted this issue, as reported earlier by Reuters. The letter did not name the AI companies accused of skirting the rule.

Although OpenAI and Anthropic have stated publicly that they respect robots.txt and blocks to their specific web crawlers, GPTBot and ClaudeBot, TollBit's findings suggest otherwise. These AI companies are **allegedly bypassing** robots.txt to **scrape** all content from websites or pages.

Robots.txt, a code used since the late 1990s, allows websites to indicate to bot crawlers that they do not want their data **scraped**. However, the demand for high-quality data for generative AI models has led to a **disregard** for this code.

OpenAI and Anthropic have popular chatbots, ChatGPT and Claude, respectively, that rely on **scraped** data from the web. OpenAI has secured deals with publishers for content access, including Axel Springer, the owner of BI.

As the US Copyright Office prepares to update its guidance on AI and copyright, the debate over AI training data and copyright continues. Tech companies have argued that web content should not be considered under copyright when used for AI training data.

Vocabulary List

1. **Disregarding** /d?s.r?????.d??/ (verb): Ignoring something or treating it as unimportant.
2. **Bypassing** /?ba??pæs.??/ (verb): Avoiding or going around.
3. **Automated** /???.t??me?.t?d/ (adjective): Operated by largely automatic equipment.
4. **Scraping** /?skre?.p??/ (verb): Extracting data from websites.
5. **Allegedly** /??le.d??d.li/ (adverb): Used to convey that something is claimed to be the case or have taken place, although there is no proof.

Vocabulary List:

1. **Disregarding** /ˌdɪs.rɪˈgɑːr.dɪŋ/ (verb): Ignoring something or treating it as unimportant.
2. **Bypassing** /ˈbaɪ.pæs.ɪŋ/ (verb): Avoiding or going around.
3. **Automated** /ˈɔː.tə.meɪ.tɪd/ (adjective): Operated by largely automatic equipment.
4. **Scraping** /ˈskreɪ.pɪŋ/ (verb): Extracting data from websites.
5. **Allegedly** /əˈlɛdʒ.əd.li/ (adverb): Used to convey that something is claimed to be the case or have taken place although there is no proof.
6. **Content** /ˈkɒn.tɛnt/ (noun): The information or material contained in a document or digital medium.

Comprehension Questions

Multiple Choice

1. What are the top two AI startups mentioned in the text?

   Option: OpenAI and Anthropic
   Option: Tesla and Amazon
   Option: Google and Microsoft
   Option: Facebook and Apple

2. What is the purpose of robots.txt according to the text?

   Option: Preventing automated scraping of websites
   Option: Enhancing web design
   Option: Improving website security
   Option: Increasing website traffic

3. Which specific web crawlers are mentioned in the text that OpenAI and Anthropic use?

   Option: GPTBot and ClaudeBot
   Option: Alexa and Siri
   Option: Cortana and Watson
   Option: Echo and Bixby

4. What roles do ChatGPT and Claude play?

   Option: Chatbots
   Option: Virtual assistants
   Option: Search engines
   Option: Web browsers

5. Who has secured deals with publishers for content access including Axel Springer?

Option: OpenAI
Option: Anthropic
Option: TollBit
Option: US Copyright Office

6. What is the purpose of TollBit according to the text?

Option: Facilitate paid licensing deals between publishers and AI companies
Option: Create AI chatbots
Option: Develop web crawler bots
Option: Analyze copyright laws

Answer

Multiple Choice: 1. OpenAI and Anthropic 2. Preventing automated scraping of websites 3. GPTBot and ClaudeBot 4. Chatbots 5. OpenAI 6. Facilitate paid licensing deals between publishers and AI companies

Vocabulary quizzes

Multiple Choice (Select the Correct answer for each question.)

1. Which industry is known for offering high-end premium products and services?

Option: Technology
Option: Fast Food
Option: Luxury
Option: Textiles

2. What is essential for maintaining mental health during challenging times?

Option: Isolation
Option: Support
Option: Neglect
Option: Denial

3. What process is initiated when a product is found to have safety issues?

Option: Launch
Option: Remodel
Option: Recall
Option: Rebrand

4. Which component is responsible for transferring power from the engine to the wheels in a vehicle?

Option: Brakes
Option: Suspension
Option: Transmission
Option: Steering Wheel

5. What do obstacles and hurdles represent in personal growth?

Option: Achievements
Option: Challenges
Option: Rewards
Option: Complacency

6. What is required to achieve success in any endeavor?

Option: Luck
Option: Efforts
Option: Connections
Option: Appearance

7. What is the term for giving special importance or focus to a specific point?

Option: Dismiss
Option: Avoid
Option: Emphasize
Option: Forget

8. Which aspect related to wellness encompasses emotional psychological and social well-being?

Option: Physical Fitness
Option: Mental Health
Option: Financial Wealth
Option: Career Growth

9. Which term refers to the production of goods using machinery in a factory?

Option: Create
Option: Design
Option: Build
Option: Manufacture

10. What action is taken when an event is called off and not rescheduled?

Option: Confirmation
Option: Revival
Option: Cancellation
Option: Postponement

Gap-Fill (Fill in the blanks with the correct word from the vocabulary list.)

11. Various _____ require specific skills and expertise.

12. Education plays a key role in raising _____ about important issues.

13. Some unethical practices involve _____ established protocols.

14. During economic downturns companies may resort to _____ to cut costs.

15. Workplaces should provide _____ for employees to express themselves.

16. Companies seek to improve _____ through efficient operations.

17. Ongoing _____ is essential for professional development.

18. Utilizing available _____ effectively can lead to success.

19. The artist is _____ for his innovative approach to painting.

20. Technology is constantly _____ to meet changing needs.

Matching Sentences (Match each definition to the correct word from the vocabulary list.)

21. surrounding mental health can prevent individuals from seeking help.
22. The essential function of a car's system is to deliver power to the wheels.
23. have a responsibility to provide a safe working environment for their staff.
24. Educational can help individuals acquire new skills and knowledge.
25. It is important to address underlying rather than just surface problems.
26. are essential for maintaining visibility in rainy conditions while driving.
27. The consists of the engine and transmission responsible for vehicle propulsion.
28. The design elements such as the of a car can enhance its overall look.

29. Regular exercise and a balanced diet contribute to overall .
30. Success is often a result of persistent and dedication.

Answer

Multiple Choice: 1. Luxury 2. Support 3. Recall 4. Transmission 5. Challenges 6. Efforts 7. Emphasize 8. Mental Health 9. Manufacture 10. Cancellation
Gap-Fill: 11. Professions 12. Awareness 13. Bypassing 14. Downshifting 15. Safe Spaces 16. Profitability 17. Training 18. Resources 19. Renowned 20. Evolving
Matching sentence: 1. Stigma 2. Transmission 3. Employers 4. Programs 5. Issues 6. Windshield wipers 7. Powertrain 8. Exterior trim 9. Well-being 10. Efforts

CATEGORY

1. Business - LEVEL5

Date Created
2024/06/22
Author
aimeeyoung99