

Colin Jones, Omshi Samal, James Daus, Calvin Lee, Hieu Do

Professor Gember-Jacobson

COSC465A

5 May 2022

Analyzing Internet Providers' Ethics: An NLP-Powered Study of ISPs' Social Responsibility

Introduction

In 2022, the Internet has achieved unprecedented importance in human society. Humans' daily lives are as defined by their encounters and experiences online as by those in the real world, and the Internet serves as a crucial arena for cultural interchange, political discourse, and global commerce. Few of us, however, turn our attention to the companies providing Internet services. What is the agenda of these ISPs? What are their positions on certain important issues, such as user data security and environmental stewardship? Our group set out to pull back the curtain on these oftentimes-mysterious corporations and discover exactly from whom America receives Internet access, basing our study upon Professor Gember-Jacobson and Emily Huff's prior research [1], which saw them utilize an NLP framework to analyze the practices and beliefs of several core ISPs. We saw an opportunity to expand upon this project by dramatically increasing the size of our ISP dataset and employing GPT-3, one of the world's most advanced language models, as our NLP framework. The pipelines we created analyze publicly-available documentation produced by these ISPs and produce answers to a battery of questions investigating ISPs' practices in five key areas: data security, net neutrality, environmental stewardship, censorship, and routing security practice. Our results revealed much about the ethics of America's ISPs and shone a light on these frequently-ignored but key power brokers of the Internet.

Methodology

Our methodology contains three main parts: data gathering, GPT-3 configuration, and accuracy measurement. Using publicly available API from Federal's Communication Commission (FCC), we retrieve a total of 68 network transparency statements from different Internet Service Provider's (ISP). We

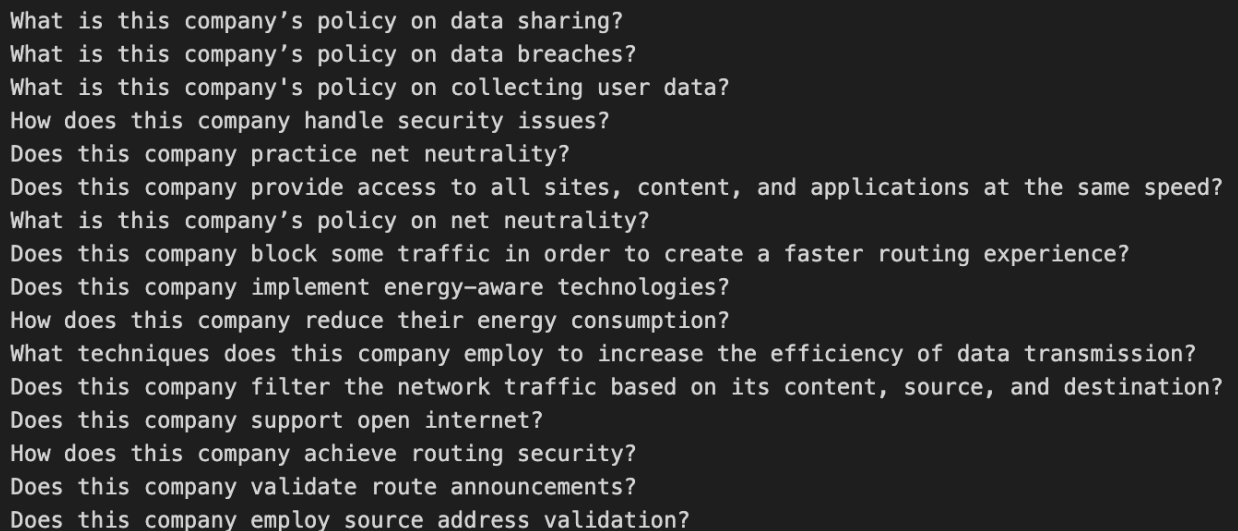
then convert these documents into readable texts for GPT-3 to parse and provide answers for various questions regarding their social responsibilities. The GPT-3 is configured to be able to consume large amount of texts and provide an answer to 16 predetermined questions. We then manually go through the answers to measure how accurate GPT-3's responses are.

Using FCC's Electronic Comment Filing System's (ECFS) public API [2] and the requests package, we were able to automatically download documents of interest. By filtering out submission types to only compliance filing, we fetched 68 documents published by various ISPs, 16 of which are in docx format and 52 of which are in pdf format. Among the 68 documents, we filtered out 16 filings which either are corrupted (unreadable by the text parser), or contain insufficient data to work with, or are mislabeled. The 52 remaining valid documents include network transparency statements, network management policy disclosures, and compliance filings. After having downloaded all the documents, we tried to parse them into text (.txt format) which is readable by GPT-3 using docx library to parse docx filing and pdfminer3 library for pdf filing.

Once we had acquired our data, we set about designing a method to analyze it. Fortunately, GPT-3, being a one-shot learning language model, is more than capable of parsing and comprehending text upon "first glance". We designed a set of sixteen questions that covered crucial subtopics within our five primary areas of inquiry, and by using OpenAI's online GPT-3 playground feature, were able to test the model's success at answering these questions upon input of data even before writing any code. Once we had refined our questions into a form that consistently allowed GPT-3 to answer accurately, we used OpenAI's provided API to write a Python script that would automate the process. After we collected all of our scraped data into a directory, our script prepended each ISP's documentation to the list of questions and fed it to GPT-3, which produced answers to our questions and wrote them to a log file. The concise, easily-human-readable answers given by GPT-3 comprise a substantial set of results in and of themselves, but we also thought it might be useful to produce even more concise output. Initially, we considered rating each company on a scale of 1 to 5 based on the ethics of their positions in each of our 5 categories, but after finding that GPT-3 struggled to accurately assign a numerical rating based on our data, we decided a

binary approach -- having GPT-3 answer a simple yes-or-no question about whether the ISP holds responsible positions -- would be more achievable. We wrote a separate pipeline, modeled after the first one, to achieve this task and produced another set of responses for each ISP.

As aforementioned, there are 16 questions that we want our GPT-3 model to answer regarding the practices of each ISP. These questions were curated empirically and covered topics such as data privacy, securities, and net neutrality.



```
What is this company's policy on data sharing?  
What is this company's policy on data breaches?  
What is this company's policy on collecting user data?  
How does this company handle security issues?  
Does this company practice net neutrality?  
Does this company provide access to all sites, content, and applications at the same speed?  
What is this company's policy on net neutrality?  
Does this company block some traffic in order to create a faster routing experience?  
Does this company implement energy-aware technologies?  
How does this company reduce their energy consumption?  
What techniques does this company employ to increase the efficiency of data transmission?  
Does this company filter the network traffic based on its content, source, and destination?  
Does this company support open internet?  
How does this company achieve routing security?  
Does this company validate route announcements?  
Does this company employ source address validation?
```

Figure 1. Questions to be answered by GPT-3

We answered these questions for each ISP by reading through their compliance filings. Each statement was read by at least two different people who then gave their own answers to each of the 16 questions above. For the binary yes-no questions, if the two answers were the same, we logged that as the “true” answer. If the two answers were different, we asked for third person’s opinion to break tie. Afterwards, we compared GPT-3’s answer to this “true” answer. For the non-binary answer, we checked if GPT-3 successfully summarized the ISP’s policy from the compliance filing or successfully returned that such information is not available. This check is also conducted by two people, with the third person breaking ties.

Results

Given our resources, we were able to fully evaluate the GPT-3 output for 19 of the companies' policies. We believe this is a large enough sample that these results can be extrapolated to the entire dataset we created using GPT-3. Each point is one survey question where GPT-3 properly used the FCC filing and answered correctly and clearly. If the information was correct but not contained in the filing, it was still marked as incorrect, as we cannot trust that any outside sources stored in GPT-3 are up-to-date or accurate. Shown below are the accuracies of GPT-3 on the companies we were able to analyze. In some cases, our pipeline or GPT-3 was unable to parse the files correctly and resulted in a failure, which we have included.

<u>Corporation Name:</u>	<u>Score:</u>
ADT Systems	13/16
Agile Network Builders	11/16
Antietam Broadband	Failure
BPM	Failure
Benjamin Wold (Walker Wifi)	16/16
Cape Ann	13/16
CastleBerry	Failure
Cedar Creek	12/16
Center Junction Telephone	14/16
Central Communications	13/16
City of Chanute Kansas	15/16
Clear Rate	6/16
Consolidated Broadband Systems	11/16
Corn Belt	15/16

Decatur	12/16
Edward_H._Winters (Yellowstone Media Design)	15/16
Ellsworth_Cooperative_Telephone Association	Failure
Extreme_Broadband_Inc	15/16
Gunby_Communications_Inc	14/16
Total	195/304 (64.1%)

Table 1. Accuracies of GPT-3 Survey Results (Unfiltered)

These results above take into account every document we checked from GPT-3, which included several failures and an anomaly in “Clear Rate” where GPT-3 answered the same prompt several times. We believe these results could be corrected with proper filtering, and displayed below are the results corrected for these anomalies.

<u>Corporation Name:</u>	<u>Score:</u>
ADT Systems	13/16
Agile Network Builders	11/16
Benjamin Wold (Walker Wifi)	16/16
Cape Ann	13/16
Cedar Creek	12/16
Center Junction Telephone	14/16
Central Communications	13/16
City of Chanute Kansas	15/16
Consolidated Broadband Systems	11/16
Corn Belt	15/16
Decatur	12/16
Edward_H._Winters (Yellowstone Media Design)	15/16

Extreme_Broadband_Inc	15/16
Gunby_Communications_Inc	14/16
Total	189/224 (84.4%)

Table 1. Accuracies of GPT-3 Survey Results (Correct for Anomalies)

The complete dataset of human-checked GPT-3 results is available [here](#), and includes the accuracy for each survey question. Full results of the GPT-3 generated artifacts and the ground truth policies can be found in the GitHub repository.

Discussion

The much larger dataset, as well as the more advanced power provided by GPT-3, have allowed us to draw much more general conclusions about ISPs' ethics than the previous study, which had a smaller sample size. Our accuracy rate of 84.4% also represented an improvement over the 71% accuracy of the original study, giving us more confidence in our results. As for the results themselves, based on the publicly-available data, it seems that most ISPs do employ ethical practices in the categories we investigated. However, the fact that these are public documents with information that was volunteered by the ISPs, we must still remain skeptical about whether these documents fully represent reality. Given the fact that they were submitted to the FCC, a regulatory body with high standards for accuracy and truthfulness, we can place more trust in these results than if the original data were sourced elsewhere, but the fact remains that any self-reported information on the part of these large companies must be taken with a grain of salt. Still, our results have real-world usage in that they represent a concise, easily human-readable summation of ISPs' stated principles, information that, if disseminated among the public, might allow citizens to make more informed decisions about their Internet usage and (if the choice is available to them) which ISPs they choose to patronize.

Conclusion

We set out with this research to explore potential ways to inform the public around company policies, and to provide valuable experimentation on scalable AI analysis. A large value our results provide is that this method can be scaled to handle far greater volumes of data and an increasingly expansive survey. GPT-3 is able to rapidly handle large amounts of data and a wide range of subjects, given the proper data and training, which far outclasses any potential manual efforts. Each question included in the survey adds a multiplicative amount of work as the dataset increases, and quickly becomes prohibitive for human checks. However, this automated process allows for hundreds more corporations to be added, and for the survey to be expanded to include an even larger range of ethical and economic considerations.

With our current pipeline, GPT-3 was often unable to handle failures gracefully, and instead outputted fake or unrelated answers in the face of corrupted data. This is a significant concern because a company whose files we cannot parse may be incorrectly identified to consumers or researchers. In future research, we believe adding large amounts of additional training data to GPT-3 that includes corrupted data and a failure response would be a suitable solution, and though the pipeline may still encounter difficulties, this will avoid any dissemination of misinformation.

We believe this research is a step in a promising direction of how AI can be used to assist civilians in becoming well-informed consumers. Because of the flexibility of this pipeline, future works could expand the dataset to include several more industries, increasingly large and politically relevant surveys, and more robust training data to enable GPT-3 to safely and correctly classify companies numerically. We look forward to seeing how awareness and accountability for ISPs and other corporations continues to develop into the future.

Bibliography

[1] Emily Huff and Aaron Gember-Jacobson. 2021. Divesting in Socially (Ir)responsible Internet Service Providers. In ACM SIGCOMM 2021 Workshop on Technologies, Applications, and Uses of a

Responsible Internet (TAURIN '21), August 23, 2021, Virtual Event, USA. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3472951.3473504>

[2] Anon. 2022.. Retrieved May 5, 2022 from <https://www.fcc.gov/>

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.