Terms of Deception: Exposing Obscured Financial Obligations in Online Agreements with Deep Learning (Extended Abstract)

Elisa Tsai
University of Michigan
eltsai@umich.edu

Anoop Singhal
National Institute of Standards and Technology
anoop.singhal@nist.gov

Atul Prakash
University of Michigan
aprakash@umich.edu

***Abstract*—This paper investigates one type of social engineering scam, where unsuspecting users inadvertently consent to hidden financial obligations by performing routine online actions, such as making a purchase. Terms and conditions, often dense and overlooked, can be a vehicle for these scams, embedding deceptive or confusing terms to manipulate users. This paper highlights the suitability of a deep learning approach to address the wordplay and nuanced language used in these terms. We propose the design of TermLens, a browser plugin that leverages Large Language Models (LLMs) to detect obscured financial terms hidden within the fine print, a task that traditional security checks often miss. We show the feasibility of Termlens detecting obscured financial terms through a case study. We also discuss challenges and future plans.**

1. Introduction and Background

In the United States alone, the Federal Trade Commission (FTC) reported a $5.8 billion financial loss due to fraud in 2021 [15]. Online scams and fraud can take many forms. One new form of social engineering scam that is increasingly pervasive relies on the typical behavior of users on the Internet: agreeing to a website's terms and conditions as a result of some action on a website, such as clicking a button to purchase an item.

Figure 1 shows a concrete real-world example—a website that claims to sell smartwatches for fitness (product name redacted to maintain neutrality and avoid potential bias). However, the catch comes when the user actually makes a purchase. Making a purchase causes the unwitting user to also agree to a subscription to a fitness app that charges the user a recurring monthly fee of $70, which per the website's terms and conditions, will be charged to the user 6 days later as shown in Figure 1. The screenshot in Figure 2 demonstrates a complaint by a victim of such a website.

Such scams exist both as fly-by-night websites that advertise themselves through social media and then disappear after scamming users and as more persistent websites on the Internet. Typically, the user is *not even required to read the terms and conditions* prior to the purchase of an item but is implicitly agreeing to additional unexpected financial obligations by simply completing the purchase. A user may have difficulty recovering the money lost, even upon disputing the additional charges with a credit card company [1], [2], because the website operator can simply state that the user agreed to the website's terms and conditions by making the purchase—not reading the terms and conditions was user's fault. These websites present not only financial dangers but also risks to privacy, as users expose sensitive personal information, including credit card details and contact information.

In addition to what we term "hidden charges scams," which involve undisclosed subscription terms leading to financial detriment for users, there are other concerning cases of obscured financial terms. For example, in the terms of use of Celsius [10], a now-bankrupt cryptocurrency lending company, it is stated:

> *"In the event that you, Celsius or any Third Party Custodian becomes subject to an insolvency proceeding, it is unclear how your Digital Assets would be treated and what rights you would have to such Digital Assets [...] You explicitly understand and acknowledge that the treatment of Digital Assets in the event of such an insolvency proceeding is unsettled, not guaranteed, and may result in [...] you being treated as* ***an unsecured creditor*** *and/or the total loss of any and all Digital Assets reflected in your Celsius Account, including those in a Custody Wallet."*

These terms imply that, in the event of bankruptcy, users could be treated as unsecured creditors of Celsius, meaning that they might not recover their digital investments. Later in court, a judge ruled that Celsius owned their users' money [29] based on these terms, highlighting the significant financial risks of not fully understanding terms and conditions.

As in any good scam, these agreements with obscured financial terms leverage *dark patterns* [8], [13], [24], [25], manipulating users for profit. Existing approaches to detect scams and dark patterns leverage word occurrences or use website features that potentially indicate a scam such as indicative images, link length, certificates, website structure, and redirection mechanism [5], [6], [14], [18], [31], [42].

Nevertheless, these approaches fall short as they typically overlook the terms and conditions, which are not covered by the standard features selected. For the examples we found for hidden charges scams, the websites often resemble legitimate small business websites, except that their terms and conditions obligate users to financial terms that are not disclosed explicitly otherwise. Moreover, the nuanced word-play nature of obscured terms necessitates a comprehensive understanding of the implications behind each term, further complicating detection efforts.

The prevalence of such deceptive practices and the potential harm they can inflict underlines the importance of our research. In this extended abstract, we introduce an obscured financial term detection system leveraging Large Language Models (LLMs).

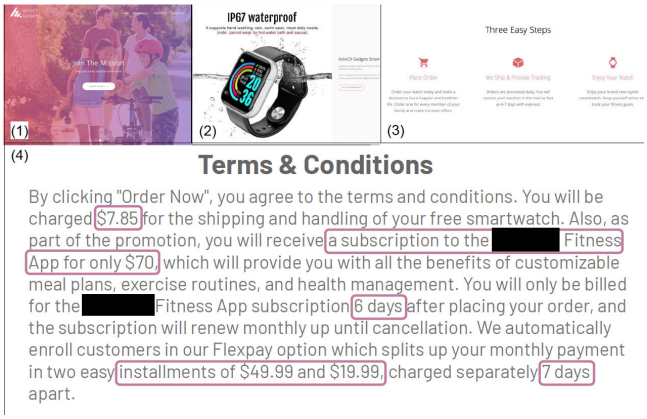

Figure 1. **Hidden Charges Websites Example**: (1)-(3): Example of hidden charges from a real-world website selling smartwatches; (4): Its terms and conditions contain a description of charges of an automatic subscription to the fitness app, but the website's purchase screens do not require the user to have reviewed those future subscription charges prior to the purchase of the smartwatch.

2. Method

We envisage TermLens to be deployed as a browser plugin that evaluates terms and conditions when a user is on a *payment page* for a purchase or a service. Designing

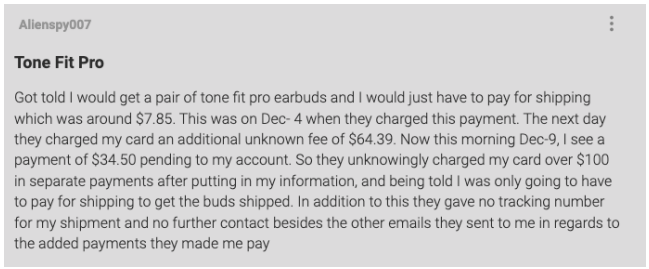

Figure 2. A victim complaint of a potential hidden charges website [4]. The hidden charges occur days after the purchase.

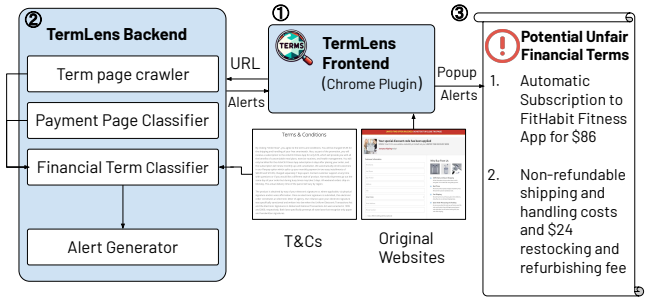

Figure 3. TermLens pipeline—① The user interface (chrome plugin) that sends back the URL to the backend; ② the server-side components that process data, including the crawling of terms pages, classification of financial terms, and the generation of alerts based on the analysis; ③ The warning displayed on the frontend, warning the user of potentially unfair financial terms detected on a website.

TermLens presents challenges since the terms and conditions need to be analyzed in the *context* of the *payment page* that the user is visiting to reduce false positives, e.g., terms already disclosed during the purchase process. We describe our prototype design and early results.

2.1. Prototype Plugin Design

When a user turns on TermLens on a webpage, the frontend sends the current URL to its backend to find potential terms and conditions associated with transactions on that page. As shown in Figure 3, the backend does a depth-first search to collect the website's term page(s). A financial term classifier then identifies terms with financial implications.

Concurrently, a payment page classifier assesses if the user is currently on a payment page. Analyzing the *context* presented by the payment page is important to determine whether the financial terms in the terms and conditions are disclosed or obscured. Therefore, when a user is on a payment page, TermLens also extracts product and financial information directly from that page. The alert generator compares the extracted information with that from the terms and conditions pages and identifies any financial obligations that are not clearly stated on the main website. The alert generator reports any obscured obligations to the user (see the last box on Figure 3).

The advent of Language Models (LMs) such as GPT-4 [3], BERT [12], RoBERTa [22], and LLaMa [35] has transformed text analysis in tools like TermLens. TermLens employs the GPT-4 API for tasks such as payment page classification, financial term identification, and alert generation. We plan to expand the range of LMs available to users, allowing customization based on accuracy, speed, and privacy considerations. As GPT-4 excels in most natural language processing tasks [3], providing a variety of LMs will require creating specific datasets for fine-tuning, particularly for analyzing terms and conditions (see §4).

2.2. Case Studies

Using the website depicted in Figure 1 as a case study, we extract the following product information from the payment page (*context*) and financial terms from T&Cs:

Extracted Product and Financial Terms

```
"Product": [{
    "Name": "[Redacted] Smartwatch",
    "Type": "Wearable Technology",
    "Category": "Electronics",
    "Price": "$0",
    "Shipping Charge":"$7.85",
    ...
}]
"Financial Terms":
[{
  "Shipping Charge:": "You will be charged $7.85
  ↪ for the shipping and handling of your free
  ↪ smartwatch.",
  "Subscription Fee": "Also, as part of the
  ↪ promotion, you will receive a subscription
  ↪ to the [Redacted] Fitness App for only
  ↪ $70.",
  "Flexpay Option Enrollment": "Automatic
  ↪ enrollment in Flexpay, splitting the
  ↪ monthly payment into two installments of
  ↪ $49.99 and $19.99, charged 7 days apart.",
  "Refund Policy": "Customers may request a
  ↪ refund within 60 days of purchase if not
  ↪ content with the product.",
  "Restocking and Refurbishing Fee": "There is a
  ↪ $24 restocking and refurbishing fee per
  ↪ unit for returned merchandise."
  ...
}]
```

The alert generator detects obscured terms, such as "Subscription Fee" and "Restocking and Refurbishing Fee", as shown in Figure 3. These terms are displayed via a popup from the TermLens plugin, effectively alerting them to these obscured financial obligations including the subscription to a fitness app that is not shown on the original website.

In another case study involving Celsius' [29] terms and conditions, TermLens identified 15 terms, including the dubious term discussed in §1, all characterized by complex legal language, with 13 presented in all caps.

3. Related Work

Dark patterns. Dark patterns, intentionally deceptive user interface designs, are widespread in websites, apps, and products, as detailed by Mathur et al. who found 1,818 instances across 11K shopping sites, encompassing 15 types of such patterns [24]. Recent research delves into the psychological, ethical, and cognitive impacts of dark patterns on user decision-making [25], [27], [28], [37], along with their legal implications and potential regulation and prevention strategies [16], [23]. Our research contributes to this field by identifying a novel category of dark pattern: the concealment of obscured terms with significant financial consequences within terms and conditions.

(Social engineering) scam detection. Scam website detection methodology largely relies on two categories of features: external features (URLs, certificates, and website logos) [7], [14], [26], [30], [31], [36], [43] and website content features (page content e.g. visual structures, HTML structures, scripts, and hyperlinks) [17], [18], [40], [41]. These features, chosen based on domain knowledge and used to build rule-based or machine learning classifiers, have limitations—attackers can evade detection by altering these features. As the Figure 1 case shows, traditional detection methods can miss scams, so we suggest using NLP to better understand website services to detect social engineering scam that resides in terms and conditions.

Legal analysis of terms and conditions. Research on using NLP for analyzing legal documents such as online contracts is limited [9], [19], [20], [21]. Braun et al. [9] applied the HuggingFace transformers model [39] to study terms of service of German online stores under EU law, creating a dataset of 50 annotated documents. This task is challenging due to the legal expertise required, making the development of such datasets time-intensive and complex.

4. Future Plans and Conclusion

4.1. Open Issues

Tackling the detection of obscured financial terms involves the following key challenges: **(1) Generalizability**: While our case study shows promising results, scaling the detection system to a variety of obscured terms remains a significant challenge. **(2) Robustness**: Enhancing the effective against adversarial attacks is important as attackers can modify their terms to evade detection systems. **(3) Readability**: In the Celsius case study, some obscured financial terms are long paragraphs with legal jargon written in all caps. Enhancing the readability of terms and conditions to make them understandable by non-specialists is also an important challenge. Furthermore, a metric to rank the importance of obscured financial terms is needed since alerting 100 such terms defeats the purpose of TermLens. Future efforts could explore methods to simplify and summarize legal language without losing its original meaning and address the potential hallucination and explainability issue that arises with it.

In the future, additional security measures on online agreements can be included to alert users about potential scams. For example, employing Named Entity Recognition (NER) techniques to extract company contact details and cross-referencing them with auxiliary datasets could verify the legitimacy of the business, providing a more comprehensive defense against scams by analyzing online agreements with deep learning.

5. Future Plans

To the best of our knowledge, there is not an open-source dataset available that specifically covers terms and conditions in English. To address this gap, we have so far gathered a dataset of 96 websites from varied sources. We sourced legitimate e-commerce and shopping sites from

SimilarWeb [34] and cryptocurrency sites from Coinmarket [11]. Malicious sites were manually collected from scam-reporting platforms like ScammerInfo [33], ScamAdvisor [32], and ScamWatcher [38] over three months. For future work, we will expand this dataset and create a comprehensive annotation scheme for identifying obscured obligation terms. We aim to fine-tune language models such as BERT to provide alternatives for users who want to opt out of commercial APIs for privacy or other concerns. We will evaluate the system on a separate evaluation dataset, as well as evaluate the effectiveness of TermLens through a user study.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments on the paper. This work is supported by a gift from the OpenAI Cybersecurity Grant program. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of OpenAI.

Disclaimer. Commercial products are identified in order to adequately specify certain procedures. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the identified products are necessarily the best available for the purpose.

References

[1] AARP. Americans lost record-breaking $8.8 billion to scams in 2022. https://www.aarp.org/money/scams-fraud/info-2023/ftc-consumer-losses.html, 2023.

[2] AARP. Undisclosed fees under consumer protection laws. https://www.justia.com/consumer/deceptive-practices-and-fraud/undisclosed-fees/, 2023.

[3] ACHIAM, J., ADLER, S., AGARWAL, S., AHMAD, L., AKKAYA, I., ALEMAN, F. L., ALMEIDA, D., ALTENSCHMIDT, J., ALTMAN, S., ANADKAT, S., ET AL. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[4] ALIENSPY007. Tone Fit Pro Reports & Reviews, 2022. https://www.scampulse.com/tone-fit-pro-reviews.

[5] BILGE, L., KIRDA, E., KRUEGEL, C., AND BALDUZZI, M. Exposure: Finding malicious domains using passive dns analysis. In *Ndss* (2011), pp. 1–17.

[6] BITAAB, M., CHO, H., OEST, A., ZHANG, P., SUN, Z., POURMOHAMAD, R., KIM, D., BAO, T., WANG, R., SHOSHITAISHVILI, Y., ET AL. Scam pandemic: How attackers exploit public fear through phishing. In *2020 APWG Symposium on Electronic Crime Research (eCrime)* (2020), IEEE, pp. 1–10.

[7] BLUM, A., WARDMAN, B., SOLORIO, T., AND WARNER, G. Lexical feature based phishing url detection using online learning. In *Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security* (2010), pp. 54–60.

[8] BONGARD-BLANCHY, K., ROSSI, A., RIVAS, S., DOUBLET, S., KOENIG, V., AND LENZINI, G. " I am Definitely Manipulated, Even When I am Aware of it. It's Ridiculous!"-Dark Patterns from the End-User Perspective. In *Designing Interactive Systems Conference 2021* (2021), pp. 763–776.

[9] BRAUN, D., AND MATTHES, F. Nlp for consumer protection: Battling illegal clauses in german terms and conditions in online shopping. In *Proceedings of the 1st Workshop on NLP for Positive Impact* (2021), pp. 93–99.

[10] CELSIUS. 1.7 million people call Celsius their home for crypto, 2023. https://celsius.network/.

[11] COINMARKET. CoinMarketCap: Cryptocurrency Prices, Charts And Market Capitalizations, 2024. https://coinmarketcap.com/.

[12] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] DI GERONIMO, L., BRAZ, L., FREGNAN, E., PALOMBA, F., AND BACCHELLI, A. UI dark patterns and where to find them: a study on mobile applications and user perception. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (2020), pp. 1–14.

[14] DRURY, V., AND MEYER, U. Certified phishing: taking a look at public key certificates of phishing websites. In *15th Symposium on Usable Privacy and Security (SOUPS'19). USENIX Association, Berkeley, CA, USA* (2019), pp. 211–223.

[15] FTC. New Data Shows FTC Received 2.8 Million Fraud Reports from Consumers in 2021, 2022. https://tinyurl.com/yetzdvj8.

[16] GRAY, C. M., SANTOS, C., BIELOVA, N., TOTH, M., AND CLIFFORD, D. Dark patterns and the legal requirements of consent banners: An interaction criticism perspective. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–18.

[17] JAIN, A. K., AND GUPTA, B. B. Phishing detection: analysis of visual similarity based approaches. *Security and Communication Networks 2017* (2017).

[18] KHARRAZ, A., ROBERTSON, W., AND KIRDA, E. Surveylance: Automatically detecting online survey scams. In *2018 IEEE Symposium on Security and Privacy (SP)* (2018), IEEE, pp. 70–86.

[19] LAGIOIA, F., MICKLITZ, H.-W., PANAGIS, Y., SARTOR, G., AND TORRONI, P. Automated detection of unfair clauses in online consumer contracts. In *Legal Knowledge and Information Systems: JURIX 2017: The Thirtieth Annual Conference* (2017), vol. 302, IOS Press, p. 145.

[20] LIMSOPATHAM, N. Effectively leveraging BERT for legal document classification. In *Proceedings of the Natural Legal Language Processing Workshop 2021* (2021), pp. 210–216.

[21] LIPPI, M., PAŁKA, P., CONTISSA, G., LAGIOIA, F., MICKLITZ, H.-W., SARTOR, G., AND TORRONI, P. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law 27* (2019), 117–139.

[22] LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., ZETTLEMOYER, L., AND STOYANOV, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[23] LUGURI, J., AND STRAHILEVITZ, L. J. Shining a light on dark patterns. *Journal of Legal Analysis 13*, 1 (2021), 43–109.

[24] MATHUR, A., ACAR, G., FRIEDMAN, M. J., LUCHERINI, E., MAYER, J., CHETTY, M., AND NARAYANAN, A. Dark patterns at scale: Findings from a crawl of 11K shopping websites. *Proceedings of the ACM on Human-Computer Interaction 3*, CSCW (2019), 1–32.

[25] MATHUR, A., KSHIRSAGAR, M., AND MAYER, J. What makes a dark pattern... dark? Design attributes, normative considerations, and measurement methods. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (2021), pp. 1–18.

[26] MOGHIMI, M., AND VARJANI, A. Y. New rule-based phishing detection method. *Expert systems with applications 53* (2016), 231–242.

[27] NARAYANAN, A., MATHUR, A., CHETTY, M., AND KSHIRSAGAR, M. Dark patterns: Past, present, and future: The evolution of tricky user interfaces. *Queue 18*, 2 (2020), 67–92.

[28] NOUWENS, M., LICCARDI, I., VEALE, M., KARGER, D., AND KAGAL, L. Dark patterns after the gdpr: Scraping consent pop-ups and demonstrating their influence. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (2020), pp. 1–13.

[29] POST, T. W. Bad news for thousands of crypto investors: They don't own their accounts, 2023. https://www.washingtonpost.com/technology/2023/01/05/celsius-crypto-bankruptcy-ruling/.

[30] SAHINGOZ, O. K., BUBER, E., DEMIR, O., AND DIRI, B. Machine learning based phishing detection from urls. *Expert Systems with Applications 117* (2019), 345–357.

[31] SAKURAI, Y., WATANABE, T., OKUDA, T., AKIYAMA, M., AND MORI, T. Discovering httpsified phishing websites using the tls certificates footprints. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (2020), IEEE, pp. 522–531.

[32] SCAMADVISER. ScamAdviser — Check a website for risk, 2023. https://www.scamadviser.com/.

[33] SCAMMER.INFO. Scammer Info - Scambait Forum and Scam Number Database, 2023. https://scammer.info/.

[34] SIMILARWEB. Website Traffic - Check and Analyze Any Website, 2023. https://www.similarweb.com/.

[35] TOUVRON, H., LAVRIL, T., IZACARD, G., MARTINET, X., LACHAUX, M.-A., LACROIX, T., ROZIÈRE, B., GOYAL, N., HAMBRO, E., AZHAR, F., ET AL. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[36] VAN DEN HOUT, T., WABEKE, T., MOURA, G. C., AND HESSELMAN, C. Logomotive: detecting logos on websites to identify online scams-a tld case study. In *Passive and Active Measurement: 23rd International Conference, PAM 2022, Virtual Event, March 28–30, 2022, Proceedings* (2022), Springer, pp. 3–29.

[37] WALDMAN, A. E. Cognitive biases, dark patterns, and the 'privacy paradox'. *Current opinion in psychology 31* (2020), 105–109.

[38] WATCHER, S. Online Scams: Reports, Informations and Victim Assistance, 2023. https://www.scamwatcher.com/.

[39] WOLF, T., DEBUT, L., SANH, V., CHAUMOND, J., DELANGUE, C., MOI, A., CISTAC, P., RAULT, T., LOUF, R., FUNTOWICZ, M., ET AL. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[40] XIANG, G., HONG, J., ROSE, C. P., AND CRANOR, L. Cantina+ a feature-rich machine learning framework for detecting phishing web sites. *ACM Transactions on Information and System Security (TISSEC) 14*, 2 (2011), 1–28.

[41] YANG, P., ZHAO, G., AND ZENG, P. Phishing website detection based on multidimensional features driven by deep learning. *IEEE access 7* (2019), 15196–15209.

[42] ZHANG, H., LIU, G., CHOW, T. W., AND LIU, W. Textual and visual content-based anti-phishing: a bayesian approach. *IEEE transactions on neural networks 22*, 10 (2011), 1532–1546.

[43] ZOUINA, M., AND OUTTAJ, B. A novel lightweight url phishing detection system using svm and similarity index. *Human-centric Computing and Information Sciences 7*, 1 (2017), 1–13.