CRADLE Law and Economics Papers

About the series

An Empirical Test of Pretrial Signaling: Text Analysis of GitHub Copyright Notices

By Pengfei Zhang and Ji Li

CRADLE Cornell Research Academy of Development, Law, and Economics

Einaudi Center for International Studies

Part of Global Cornell

An Empirical Test of Pretrial Signaling: Text Analysis of GitHub Copyright Notices

Pengfei Zhang* Ji Li†

March 13, 2024

Abstract

This paper presents an empirical test of disputants' settlement behavior using online copyright notices. The Section 512(c) notice-and-takedown regime provides a natural setting to study the signaling aspect of pretrial bargaining. We apply text analysis to quantify the attributes of notice as a pretrial signal, and we use the text data to evaluate how different factors help to close the information gap and improve the settlement rate. The three primary determinants that help settlement are found to be text features of the complaints, legal representation, and platform mediation. A strong signal is short, easy to read, and more specific. Legal representation improves the credibility of the signal. Platform mediation, on the other hand, adds commitment to the signal. Interestingly, how the lawyers draft a notice compromises the positive effect of legal representation. Lawyers prefer long sentences, big words, and more terminology, whereas an effective notice is much more concise. Most of our empirical findings support theoretical predictions, but we also discuss some discrepancies between the two. [1]

*School of Economic, Political and Policy Sciences, The University of Texas at Dallas, pengfei.zhang@utdallas.edu

†School of Economic, Political and Policy Sciences, The University of Texas at Dallas, jxl190007@utdallas.edu

[1]We are grateful to Benjamin Artz (discussant) at EEA 2024. All errors are our own.

Keywords: *Pretrial Settlement, Signaling, Notice and Takedown, Text Analysis*
JEL Codes: C72, D82, K24, K40

1 Introduction

A large majority of civil cases are settled prior to trial. [2] Understanding why some lawsuits go to trial while many others are resolved out of court is a primary question in law and economics. Theoretical models have analyzed how strategic information transmission would affect the outcomes of pretrial bargaining. The signaling model aka Reinganum and Wilde (1986), in particular, has been a workhorse model in the theory of litigation. There are few studies, however, that quantify this information transmission behavior in an empirical context.

This paper provides an empirical test of the disputants' pretrial signaling behavior. The Section 512(c) notice-and-takedown regime, enacted as part of the Digital Millennium Copyright Act (DMCA), is a natural setting to study such behavior. Under this regime, copyright owners can send takedown notices to online service providers to remove the content they believe to be infringing. A notice must include (a) claiming under penalty of perjury that the plaintiff owns the copyright to the original work, (b) identifying unauthorized and infringing works on the platform, and (c) the plaintiff's legal demand (remove, modify, or go private). Online platforms have a strong incentive to comply otherwise they risk losing their safe harbor protection (which shields them from lawsuits). [3] The notice and takedown procedure creates a signaling game between the plaintiff and the defendant. For each copyright dispute, the relevant victim issues a notice that can be viewed as a settlement offer. The defendant decides either to accept the notice (in which case, she complies with the request) or to reject the notice (in which case, she has one opportunity to send a counter-notice). The majority of DMCA notices are not subject to the scrutiny of a court. As Urban and Quilter (2005) put it, "this was precisely the point behind Section 512: the efficient removal of infringing materials from the Internet in a fair process, with (in most cases) no need for court review".

We begin by reviewing the signaling model and extending it to two related institutional settings-one where the plaintiff can send verifiable signals (which we call the

[2] Less than 4% of civil cases that are filed in the US state courts go to trial. In the US Federal Courts, only about 2% of civil cases go to trial.

[3] See Grimmelmann and Zhang (2023) for a discussion on the safe harbor liability rule.

disclosure model), and another where a mediator facilitates the settlement (which we call the mediated settlement model). The three models correspond to the owner-sent notices, the attorney-sent notices, and the GitHub-reviewed notices in our sample respectively. By comparing the three models, we derive three sets of hypotheses that set the stage for our empirical tests. Our focus is on how different institutional factors help to close the information gap and improve the settlement rate. The models show that both legal representation and platform mediation increase the settlement rate. Legal representation uniformly increases the credibility of the signals, whereas platform mediation adds commitment and polarizes the signals (i.e., make the strong signals stronger). We also derive hypotheses on the heterogeneous effects of the factors when the winning probability changes.

We collect 4,684 takedown notices received by GitHub and extract different text features from them. The text information falls into four categories: main features, the platform's action, textual characteristics, and precautionary technology. The main features include whether the notice is prepared by an attorney hired by the owner, the number of infringing URLs, and the demand of the plaintiff. We also extract GitHub's annotations for the takedown notices. These annotations document instances where GitHub allows plaintiffs to make revisions to their notices and permits defendants to make necessary adjustments in order to prevent the removal of content. We quantify the signal conveyed by the notices by the following textual characteristics. For each notice, we summarize the *length, readability, specificity, similarity, and redundancy* of the notices' descriptions on ownership and copyrighted work. We first test whether the text attributes are different across legal representation and platform mediation. We then conduct regression analysis to see what factors predict the settlement rate. Finally, the notices also report whether the plaintiff has adopted precautionary measures such as open-source licenses and anti-circumvention technologies. We use them as proxies for the winning rate in testing the heterogeneous effect of signaling, legal representation, and platform mediation.

Text matters. We find that our measure of text features, especially readability and specificity, is an important determinant of the settlement rate. When the respondent describes the ownership and the copyrighted work, higher readability improves the settlement rate. Mentioning more specific named entities, such as the names of companies,

organizations, and locations, also enhances the odds of settlement. For both ownership and infringement descriptions, overly lengthy answers make the settlement harder. The presence of repeated words in the descriptions has minimal impact on the likelihood of settlement. In summary, an effective notice involves concise yet informative descriptions, emphasizing named entities, and avoiding excessive wordiness.

Lawyers and owners differ significantly in their writing style, demands, and investigation efforts. In terms of writing style, a lawyer-written notice is significantly longer and harder to read. Lawyers also include more specific named entities and use more repeated words, perhaps providing additional context for the dispute. Next, lawyers are more demanding. They more often demand a complete takedown of the repository instead of allowing the content to go private or giving the defendant a chance to modify it. Last but not least, lawyers demonstrate a higher commitment to investigation and legal research. They identify more infringing links and the associated URLs than the owners. They meticulously review the URLs before GitHub's inspection. All these indicate a more thorough evidence-gathering process.

Hiring a lawyer helps with the settlement. Consistent with our theoretical prediction, legal representation has a significant positive effect on the settlement rate. The virtue of legal representation most likely lies in sufficient investigation rather than a distinct writing style or demand. This is consistent with our theory that the presence of a lawyer helps to increase the credibility of the signal. Lawyers' writing style, interestingly, might hurt settlement. While their notices do provide more useful information, their writing tends to be more lengthy and convoluted in readability, both of which actually reduce the settlement rate.

An active role of the platform helps settlement too. In our case, GitHub is, to some extent, mediating the copyright disputes between the two parties. The GitHub Trust & Safety team reviews the takedown notice to see if it complies with the statutory requirements. It will let the plaintiff amend their copyright claims whenever appropriate. The team also grants a chance to the defendant to modify the content and avoid removal. Consistent with theoretical prediction, platform mediation significantly increases the settlement rate. Both of the mediation practices - review and revision - increase the odds of

settlement substantially.

Most of our empirical findings support theoretical predictions. As predicted by the models, legal representation, platform mediation, and lower plaintiff's demand all increase the settlement rate. The results are robust for different specifications. The regression results are consistent with the main findings of the three pretrial signaling models on the settlement rate and on the signaling mechanism. When using open-source licenses and anti-circumvention technology as winning rate proxies, our regression analysis provides mixed evidence for the heterogeneous effects with regard to winning rates predicted by the models.

This paper contributes to the burgeoning literature on pretrial settlement. A large fraction of this literature uses game-theoretic models to investigate the litigation and resolution of civil disputes. The literature highlights two explanations for why settlement fails, resulting in a costly trial. One argues that asymmetric information is a barrier to pretrial bargaining (Bebchuk (1984), Reinganum and Wilde (1986), and much of the subsequent literature that builds upon them). The other argues that mutual optimism - that is, both parties think they are more likely to win - leads to a bargaining impasse (Priest and Klein (1984), Yildiz (2011)). [4] In this paper, we extend the signaling model to predict the effects of legal representation and platform mediation.

The empirical evidence that is needed to provide the foundation for the models is, however, in more limited supply. A common approach is to test the predictions of different models. Waldfogel (1998) use data describing contracts, torts, and intellectual property cases in the Southern District of New York from 1984-1987 to evaluate the information asymmetry hypothesis versus the mutual optimism hypothesis. Pecorino and Van Boening (2018) compares the screening model and the signaling model in a lab experiment setting. Silveira (2017) estimates a structural model of plea bargaining based on the screening model. The first challenge to this approach is that the empirically testable implications of the models depend strongly on (often-times) unobservable details of the bargaining process, e.g., who has private information, who makes the offer, the number of rounds (Spier (2007)). Assumptions on the bargaining process have to be made on

[4] See Spier (2007) for an excellent review of the theoretical literature.

the sample to fit the model. [5] In our context of notice-and-takedown, the role of the sender, the role of the receiver, and the two-step procedure are all statutorily defined. The second challenge is that information transmission in the pre-trial stage is not directly measured but rather inferred from the settlement offers. This can be problematic because the wealth effect of an offer can contaminate the identification of the signaling effect. In this paper, we quantify the information transmitted in a notice using text analysis techniques. The text-as-data techniques allow us to quantify the attributes of the pretrial signals and thereby study their credibility and informativeness using the open-ended questions found in GitHub takedown notices. [6] We apply these methods to extract textual features from the takedown notice that reflect the plaintiff's signal and evaluate their influence on the settlement outcome. Our study cannot solve the two empirical challenges entirely but we believe we are making insightful progress.

Previous literature has investigated different determinants of the settlement rate. They found that legal representation (Poppe and Rachlinski (2016)), the type of plaintiff (Eisenberg and Farber (1997)), the stakes of the case (Chang and Hubbard (2021)), mediation (Klerman and Klerman (2015)), and risk aversion (Viscusi (1988)) all play a role in the success of the settlement. [7] The existing literature has also delved into different industries, including antitrust (Perloff et al. (1996)), pharmaceutical patent (Ahn et al. (2023)), computer patent (Somaya (2003)), product liability (Viscusi (1988)), medical malpractice (Danzon and Lillard (1983)), and foreign investment dispute (Vu (2021)). This paper makes two primary contributions to this literature. This paper is, to our knowl-

[5] For example, in Waldfogel (1998), the information asymmetry hypothesis is based on Bebchuk (1984) which posits that the case is more likely to settle if the better-informed defendant expects a lower winning rate. The prediction will be reversed if the plaintiff is informed or if the defendant makes the offer.

[6] Text analysis in legal research just begins to emerge (see Choi (2023)). Our textual measures have been used in the field of accounting to study various attributes of financial statements. For instance, Dyer et al. (2017) employs measures such as length, readability, and informativeness to capture the characteristics of the 10-K filings. See Grimmer et al. (2022) for a general reference for applying text analysis to social science.

[7] Literature on the effect of Legal representation is substantial. Poppe and Rachlinski (2016) reviews this line of research and suggests that lawyers generally assist litigants in achieving more favorable outcomes. Similarly, Earnhart and Rousseau (2019) finds that defendants tend to get better outcomes when represented by a lawyer, although the benefits may not always cover the costs. Huang (2008), on the other hand, finds that clients, rather than lawyers, play a more dominant role in deciding whether to settle or litigate a case. The low settlement rate in cases where both parties are represented could be attributed to a representation selection effect, where parties inclined to litigate are more likely to seek legal representation. Literature on the effect of mediation is limited. Klerman and Klerman (2015) finds that mediation facilitates settlement and the mediator's proposal technique contributes to the facilitation. Parties tend to make larger concessions in the early stages of negotiation, especially when a mediator's proposal is accepted.

edge, the first paper to show that text feature is an important determinant of settlement rate. We also show how the text feature interacts with other factors such as legal representation and mediation. Second, we investigate the settlement question in the digital copyright domain, which is novel to the literature. We show how legal representation and mediation are also important in the online dispute context.

This paper also contributes to the growing literature on content moderation. [8] Urban and Quilter (2005) presents the first set of descriptive statistics on the notice and takedown process under DMCA Section 512. They found that corporations and business entities were the primary senders of the notices, a majority of the notices were sent for competition purposes, one-third of the notices were questionable regarding the validity of the copyright infringement claim, and few individual users responded with a counter-notice. [9] The literature begins to have rigorous econometric studies on the impact of the notice-and-takedown. Empirical evidence of over-removal is accumulating, indicating the chilling effects of the policy (see Keller (2015) for a survey). For instance, Penney (2019) runs a survey experiment with hypothetical scenarios of receiving a takedown notice. He finds that respondents broadly reported being less likely to share any content in the future and only one-third said they would send a counter-notice or challenge the takedown they believed was wrong or mistaken. Zhang (2021) leverages the timing of notices for an event study and finds a persistent drop in original contributions following the takedown of a repository. This paper complements the above literature by conducting a text analysis of the notices and investigating the different factors that affect a counter-notice.

The remainder of the paper proceeds as follows. Section 2 reviews the necessary institutional background of the Digital Millenium Copyright Act and the notice-and-takedown process. Section 3 reviews the signaling model, compares it with two other related models, and derives hypotheses for empirical exercise. Section 4 describes the GitHub notice data and empirical strategies. Section 5 presents the empirical results. Section 6 concludes.

[8] See Keller and Leerssen (2020) for surveying information released by platforms and independent research.

[9] In a follow-up study, Urban et al. (2017) emphasizes the role of automation in sending complaints, and compares how the automated notices differ from the manual notices by small rightsholders. See also **?** on the questionable validity of many takedown notices, especially those generated by automated systems.

2 Institutional Background

2.1 Digital Millenium Copyright Act

Copyright law protects literary, scientific, and artistic works, giving their creators the ability to control certain uses of their works. Copyright, in its current shape, is a bundle of two major rights: (i) the exclusive right to make copies and distribute them, (ii) the exclusive right for further derivative works.[10] If a work is under copyright, the copyright owner has the exclusive right to make derivative works based on the copyrighted work. If someone else makes an unauthorized derivative work, it may infringe the copyright. The primary defense for unauthorized use is the doctrine of fair use. A particular use may be fair if it only uses a small amount of copyrighted content, uses that content in a transformative way, uses it for educational purposes, or some combination of the above. Because code naturally lends itself to such uses, each use case is different and must be considered separately.

In 1998, Congress passed the Digital Millennium Copyright Act (DMCA), codified in Section 512, to address copyright issues in the digital age. The Act is designed to protect the rights of copyright holders, encourage technological innovation, and balance the interests of content creators and internet service providers. The DMCA has three innovative pieces of legislation. First, it offers provisions for notice and takedown procedures, making it possible for copyright owners to request the removal of infringing content from online platforms. Second, it creates a safe harbor provision for internet service providers hosting allegedly infringing user-generated content. As long as the online platform follows the notice-and-takedown rules, it will not be liable for copyright infringement of third-party content, providing incentives for platforms to maintain the safe harbor status. Third, it also includes anti-circumvention measures, which prohibit the bypassing of digital rights management (DRM) and other protective technologies.

[10] A "derivative work" is defined as a work based upon one more preexisting works, such as translation, musical arrangement, dramatization, fictionalization, motion picture conversion, sound recording, art reproduction, abridgment, condensation, or any other form in which a work may be recast, transformed, or adapted.

2.2 GitHub and its Copyright Issues

Founded in 2008, GitHub is an online platform hosting open-source software. GitHub provides a bundle of collaborative tools built on top of the Git version control system, making it very popular among software developers. Git allows the developers to collaborate on the same project at the same time, access previous versions of the project, and merge individual contributions into the code base. To launch a project on GitHub, developers create a repository (referred to as a repo). This code warehouse comprises all files related to the project, including source code, project files, resource files, and configuration information. The leading developer initiates and tracks the development of the repository, allowing other fellow developers to contribute improvements to the code within the repository. As of January 2023, GitHub had over 100 million developers and hosted more than 372 million repositories, with at least 28 million of them being public. It is the largest source code hosting platform.

Copyright issues arise on GitHub when the codes or other contents in a user's repository infringe on someone else's intellectual property rights. The GitHub team handles copyright infringement in accordance with the DMCA. If copyright owners believe that their intellectual property rights have been infringed on GitHub, they can submit a takedown notice to GitHub. This notice includes information about the infringing content, the rights being infringed, and contact information for the copyright owner. The copyright owners are not necessary users of GitHub. For instance, if a company finds its unauthorized codes appear in files under some repositories of GitHub users, its employees engaged in legal affairs may submit a takedown notice to require GitHub to remove those codes.

2.3 Notice and Takedown

The DMCA provides a notice-and-takedown process for complaints about copyright infringement and the process consists of two parts: (i) a takedown-notice procedure for copyright holders to request that content be removed; and (ii) a counter-notice procedure for users to get content reenabled when content is taken down by mistake or misidentifi-

cation.

Section 512(c) creates the notice-and-takedown regime. To receive safe harbor, hosting services and search engines are required by 512(c) to respond "expeditiously" to notices of copyright infringement. For instance, Section 512(c)(1)(C) removes the platform's safe harbor protection if it receives a "notification of claimed infringement" and fails to remove it. Section 512(c) applies to hosted content and requires the online platforms to establish and maintain a structured process as discussed below.

In GitHub, DMCA takedown notices are frequently used by copyright owners to ask GitHub to take down content they believe to be infringing. Upon an initial investigation, the plaintiff prepares and sends a takedown notice to GitHub. Assuming the takedown notice meets the minimum requirements of the DMCA, GitHub will post the notice to its public repository and pass the complaint along to the affected user. [11] GitHub will disable access to the defendant user's content if: (i) the copyright owner has alleged copyright over the user's entire repository or package; (ii) the user has not made any changes after being given an opportunity to do so; or (iii) the copyright owner has renewed their takedown notice after the user had a chance to make changes.

Section 512(g) outlines the counter-notification procedure. If a user believes their content was wrongly removed due to a DMCA takedown notice, they can file a counter-notification to have it restored. Some statutory requirements for a counter-notice include the user's contact information, a statement of good faith belief that the material was removed or disabled as a result of mistake or misidentification, and consent to jurisdiction in a U.S. district court. Once the service provider receives a valid counter-notice, it must promptly inform the copyright complainant that it will put back the material unless the complainant files a court action. Unless a copyright owner chooses not to seek judicial remedy following receipt of the counter-notification, a takedown target must be prepared to defend the challenged use in a Federal District Court.

In GitHub, counter notices can be used to dispute a takedown notice. If a user

[11] The minimum requirements include (a) identifying copyrighted works that are allegedly being infringed, (b) claiming under penalty of perjury that the plaintiff owns the copyright to the original work, (c) that the content on GitHub is unauthorized and infringing. GitHub exercises little discretion in the process, and it is up to the parties (and their lawyers) to evaluate the merit of their claims.

believes that her content was disabled as a result of a mistake or misidentification, she may respond with a counter notice. GitHub will post it to its public repository and pass the counter notice back to the copyright owner. If a copyright owner wishes to keep the content disabled after receiving a counter notice, he will need to initiate a legal action seeking a court order to restrain the user from engaging in infringing activity relating to the content on GitHub. If the copyright owner does not give GitHub notice within 10-14 days, by sending a copy of a valid legal complaint filed in a court of competent jurisdiction, GitHub will re-enable the disabled content.

Protections for the target of the notice (the alleged infringer) are relatively few, and judicial protection is not available unless three things occur: the target elects to submit a counter notice; the complainant then files suit; and a court reviews the issue. However, the statute provides encouragement to online platforms to replace wrongfully or mistakenly targeted material. An online service provider might be subject to some tort or contractual liability for a wrongful takedown of content. GitHub has been active in doing simple screening of fraudulent notices. GitHub frequently contacts the defendants to give them a chance to make their software compliant. GitHub sometimes also checks the validity of the notices. [12]

The vast majority of DMCA notices likely are never subject to the scrutiny of a court. As Urban and Quilter (2005) put it, "this was precisely the point behind Section 512: the efficient removal of infringing materials from the Internet in a fair process, with (in most cases) no need for court review".

3 Model and Theoretical Predictions

In this section, we consider and compare three related models of pretrial settlement: the signaling model, the disclosure model, and the mediated settlement model. Legal representation and platform mediation are two primary institutional factors in our setting

[12]Not every online service provider does that. In practice, platforms limit their liability with their terms of service. Although the statute seeks to encourage putback by providing a safe harbor against liability for wrongful takedown, the platforms' service contracts limit most legal or financial incentives for doing so.

that help to close the information gap. We start with a brief review of the signaling model aka Reinganum and Wilde (1986). Informed readers may skip the subsection at no cost. The signaling model resembles the owner-sent notices in our sample. We then incorporate costly disclosure to the signaling model. The disclosure model resembles the attorney-sent notices in our sample. We show that legal representation increases the settlement rate by increasing the credibility of the signals. Lastly, we incorporate a mediator to the signaling model. The mediator model resembles the GitHub-reviewed notices in our sample. We show that platform mediation increases the settlement rate by polarizing the signals.

We derive three sets of hypotheses from analyzing the models, which sets the stage for our empirical tests. Collectively, these theoretical predictions guide our empirical exercise in Section 5 where we use different characteristics of the plaintiff's notice to predict the likelihood of receiving a rejection (counter-notice) from the defendant.

3.1 Benchmark Model

There is a plaintiff and a defendant. The plaintiff and the defendant fight for the property right of a good valued at v. In the event of a trial, there is a probability $\pi \in [\underline{\pi}, 1]$ that the plaintiff will prevail and obtain the property right. With probability $1 - \pi$, the defendant prevails and obtains the property right instead. Regardless of the outcome, the litigation would impose legal costs c_p for the plaintiff and c_d for the defendant.

Settlement negotiations are conducted against the outside option of the litigation. The plaintiff and the defendant have different beliefs regarding π. Suppose the plaintiff believes the winning probability at trial to be π_p while the defendant believes it to be π_d. These divergent beliefs may arise because the two litigants receive different signals of the true state, and may be influenced by their different backgrounds and experiences. Let S be the plaintiff's demanded settlement. The two parties will settle the dispute if $S \geqslant \pi_p v - c_p$ and $v - S \geqslant (1 - \pi_d)v - c_d$. Rearranging terms gives us the settlement zone, $S \in [\pi_p v - c_p, \pi_d v + c_d]$. The settlement condition will be met if $\pi_p - \pi_d \leqslant \frac{c_p + c_d}{v}$ when the gap in beliefs is capped by the cost-value ratio of the litigation. In what follows, we

present theoretical predictions from three institutional arrangements: unverifiable signaling, costly disclosure, and mediation. All three have different implications for the belief gap $\pi_p - \pi_d$ and change the settlement rate.

3.2 Signaling Model: A Review of Reinganum and Wilde (1986)

Nature randomly selects a state $\pi \in \pi$ according to some non-degenerate commonly known distribution $F(\pi)$, and reveals it to the plaintiff only. The informed plaintiff makes a take-it-or-leave-it offer S, which will signal his private information π. The uninformed defendant will then form a Bayesian belief regarding the plaintiff's type and decide how to respond to the offer. Reinganum and Wilde (1986) characterize a fully-separating equilibrium of this game where the plaintiff's demand perfectly reveals his type and the defendant mixes between accepting and rejecting the offer.

Let $S(\pi)$ be the plaintiff's demanded settlement given his type π. For such a mixed strategy to work, the defendant must be indifferent between accepting and rejecting the request such that $(1-\pi)v - c_d = v - S(\pi)$. This implies that the plaintiff's equilibrium demand would be

$$S(\pi) = \pi v + c_d, \tag{1}$$

which gives the defendant exactly the same payoff that she would get from litigation.

Let $\sigma(S)$ be the probability that the defendant accepts the offer S. Given the acceptance probability $\sigma(S)$, the plaintiff will make an offer that maximizes his expected payoffs $\sigma(S)S + (1-\sigma(S))(\pi v - c_p)$. Combining the first-order condition of S with (1), we get $\frac{\sigma'(S)}{\sigma(S)} = -\frac{1}{c_p + c_d - v}$, which equalizes the hazard rate with the inverse of the rent dissipation. This is a first-order linear ordinary differential equation. Solving this differential equation with the boundary condition $\sigma(\underline{S}) = 0$, we obtain $\sigma(S) = e^{1 - \frac{s}{c_p + c_d - v}}$. Notice that the probability of settlement $\sigma(S)$ is decreasing in the demand S. In equilibrium, higher-type plaintiffs must have their demanded offers rejected more frequently in order to discourage lower-types from bluffing. There is a one-to-one mapping between the type

and the equilibrium acceptance

$$\sigma(\pi) = e^{-\pi/(c_p+c_d-v)} \tag{2}$$

such that the offer is fully revealing. Also note that the settlement rate is always less than 1 for any type. This happens because of asymmetric information and the non-verifiability of the signals.

Identifying the signaling effect, however, is an empirical challenge. There are two competing effects of a higher offer: on the one hand, higher demand signals a higher probability for the plaintiff (defendant) to prevail (lose) in court, and thus shall increase the settlement rate. We call this the "signaling effect". On the other hand, higher demand decreases the defendant's payoff in settlement, and thus shall decrease the settlement rate. We call this the "wealth effect". The wealth effect dominates the signaling effect in Reinganum and Wilde (1986), and that is why rejection probability $\sigma(S)$ is increasing in the demand S. An empirical exercise aiming to identify the signaling effect of pretrial offers must shut down the wealth effect.

The signaling model described above closely resembles the notice and counter-notice procedure in GitHub. The property right in this case is the copyright of a software. The copyright owner - the plaintiff - first files a complaint against a disputed software, after which the developer of the software - the defendant - has the opportunity to make a counter-notice to rebut the complaint. The notice is a settlement offer that signals the plaintiff's estimate of the legal case. A counter-notice is a rejection by the defendant. The law requires the notice and counter-notice procedure before the case can proceed to court. The goal of the law is to foster efficient bargaining between the two parties outside the court. The process can thus be viewed as a simple form of settlement negotiation.

Comparative statics exercises of the equilibrium acceptance rate with respect to S and π give us the following hypotheses.

Hypothesis 1 (Signaling)**.**

Hypothesis 1a. *The more demanding the plaintiff is, the lower the settlement rate is.*

Hypothesis 1b. *The settlement rate is higher if the plaintiff's winning probability is lower.*

3.3 Incorporating Costly Disclosure in the Signaling Model

Not all notices are sent by owners themselves. In fact, more than half of the notices in our sample are sent by legal representatives. Attorneys are costly but they also possess legal expertise that can help the case. In particular, they can produce tangible evidence that proves the strength of the case and show it in court. In this section, we add costly disclosure to the signaling model to predict the effects of legal representation. [13]

The informed plaintiff decides whether to show the evidence $e \in E$ to the defendant before making an offer. The evidence fully reveals π and is only obtainable after paying a cost $c_e > 0$ (we assume $c_e < c_p + c_d$). If the plaintiff is silent, the game is the same as above in which the defendant decides whether to accept the offer. If the plaintiff reveals the evidence, the defendant decides whether to accept or reject after observing the evidence.

At the final stage, the defendant will accept the plaintiff's demand S if $v - S \geqslant (1 - \pi_d)v - c_d$, or $S \leqslant \pi_d v + c_d$. If a plaintiff reveals π, the defendant's belief becomes $\pi_d = \pi$, the maximum demand the defendant will accept is $S(\pi) = \pi v + c_d$, and the plaintiff's payoff from disclosure is $\pi v + c_d - c_e$. If instead the plaintiff is silent, the plaintiff can only ask for a single settlement $\hat{S}$. Silence is optimal if $\hat{S} \geqslant \pi v + c_d - c_e$, or $\pi \leqslant \frac{\hat{S} - c_d + c_e}{v}$. The set of silent plaintiffs is thus $[\underline{\pi}, \frac{\hat{S} - c_d + c_e}{v}]$.

The plaintiff will choose $\hat{S}$ to maximize his expected payoffs:

$$\max_{\hat{S}} \int_{\underline{\pi}}^{\frac{\hat{S} - c_d + c_e}{v}} \hat{S} dF(\pi) + \int_{\frac{\hat{S} - c_d + c_e}{v}}^{1} [S(\pi) - c_e] dF(\pi)$$

where the first term is the expected payoff from settlement and the second is that of the trial. The optimality condition of $\hat{S}$ gives us

$$F(\frac{\hat{S} - c_d + c_e}{v}) = (c_d - 2c_e) f(\frac{\hat{S} - c_d + c_e}{v}).$$

The left hand is the marginal cost of raising the demand due to asking less to those de-

[13] Shavell (1989) considers a screening game where the informed plaintiff can disclose verifiable information at no cost before the uninformed defendant makes the offer. Farmer and Pecorino (2005) study how costly discovery affects the bargaining outcome in the signaling and the screening models.

fendants already willing to settle. The right hand side is the marginal benefit due to inducing, at the margin, $f(\frac{\hat{S}-c_d+c_e}{v})$ more defendants to settle, thereby earning $c_d - 2c_e$ per defendant.

In this model, all plaintiffs settle. In the extreme case when $c_e = 0$, complete unraveling will happen. The plaintiff with the strongest case would reveal it and demand $S = \pi v + c_d$ which would be accepted with probability one. The plaintiff with the second strongest case would do the same and his demand would be accepted as well. In the end, all plaintiffs but the lowest type are led to reveal their type and all would settle with the defendants for $S(\pi)$. In the general case where $c_e > 0$, the plaintiffs still never go to trial even though some of the plaintiffs may choose to remain silent. If the plaintiff chooses to disclose π, he would settle for $S(\pi)$; otherwise, he would settle for $\hat{S}$.

Comparing the disclosure model with the signaling model, we have the following three hypotheses.

Hypothesis 2 (Legal Representation)**.**

Hypothesis 2a. *Legal representation increases the credibility of the signals.*

Hypothesis 2b. *Legal representation leads to a higher settlement rate.*

Hypothesis 2c. *The positive effect of legal representation on the settlement rate is larger if the plaintiff's winning probability is higher.*

3.4 Incorporating Mediator in the Signaling Model

GitHub is taking an increasingly active role in the notice-and-takedown process. GitHub frequently contacts the defendants to give them a chance to make their software compliant. GitHub sometimes also checks the validity of the notices. In the dispute resolution framework, GitHub is mediating the disputes between the two parties on the software it hosts. In this section, we add a mediator to the signaling model to predict the effects of platform mediation.

There is a plaintiff, a defendant, and a mediator. The mediator commits to a mediation plan. The plaintiff first privately reports a type $\hat{\pi} \in [\underline{\pi}, 1]$ to the mediator. The

mediator then either recommends a proposal S or terminates the process. The mediator's action is publicly observed. Mediation is successful if both parties agree to the proposal. Otherwise, they proceed to the trial. Formally, a mediation plan is a tuple $\{\sigma(\hat{\pi}), S(\hat{\pi})\}$ where $\sigma(\hat{\pi})$ is the probability of reaching an agreement, and $S(\hat{\pi})$ is the recommended settlement.[14]

A mediation plan is efficient if it maximizes the *ex-ante* total payoffs, i.e., $\mathbb{E}[u_p + u_d]$ subject to the plaintiff's and defendant's incentive compatibility and participation constraints. In this simple model, efficiency is equivalent to maximizing the settlement rate.

Define $U_p(\pi, \hat{\pi})$ and $U_d(\pi, \hat{\pi})$ as the expected payoff of the plaintiff and the defendant respectively. They are

$$\begin{aligned} U_p(\pi, \hat{\pi}) &= \sigma(\hat{\pi})S(\hat{\pi}) + (1 - \sigma(\hat{\pi}))(\pi v - c_p) \\ U_d(\pi, \hat{\pi}) &= -\sigma(\hat{\pi})S(\hat{\pi}) - (1 - \sigma(\hat{\pi}))(\pi v + c_d). \end{aligned}$$

If the two parties settle with probability $\sigma(\hat{\pi})$, the settlement is $S(\hat{\pi})$. Since mediation is self-enforcing, the harshest punishment is to ask the two parties to proceed to trial if they fail to settle.

The efficient mediation plan determines $\{\sigma(\pi), S(\pi)\}$ to maximize the total payoffs:

$$\begin{aligned} \max_{\sigma(\pi), S(\pi)} \quad & \int_{\underline{\pi}}^{1} \sum_i U_i(\pi) \mathrm{d}F(\pi) && (3) \\ \text{s.t.} \quad & U_p(\pi, \pi) \geqslant \pi - c_p, \ \ \forall \pi && \text{(IR-P)} \\ & U_d(\pi, \pi) \geqslant -\mathbb{E}_\mu [\pi | S] - c_d, \ \ \forall \pi && \text{(IR-D)} \\ & U_p(\pi, \pi) \geqslant U_p(\pi, \hat{\pi}), \ \ \forall \pi, \hat{\pi} && \text{(IC)} \end{aligned}$$

where the first set of constraints is individual rationality for the plaintiff, the second set of constraints is individual rationality for the defendant, and the last set of constraints is incentive compatibility for truth-telling.

[14] By the revelation principle, we can restrict our attention to direct mechanism w.l.o.g. where the report is a type and the message is a settlement.

Myerson (1981) establishes the necessary and sufficient conditions for a direct mechanism to be incentive compatible. Applying Myerson's lemma to the (IC) constraint, we know that a mediation plan is incentive compatible, if and only if the following two conditions hold: (i) $\sigma(\pi)$ is non-increasing in π, and (ii) for any $\pi \in [\underline{\pi}, 1]$, $U_p(\pi) = \int_{\underline{\pi}}^{\pi}[1-\sigma(\tilde{\pi})]\mathrm{d}\tilde{\pi} + U_p(\underline{\pi})$. The first condition ensures the weak monotonicity of the function $\sigma(\pi)$ such that a higher type is always rewarded with a weakly higher probability of reaching an agreement. The second condition shows that the expected payoff of different types of the plaintiff is pinned down by the settlement probability $\sigma(\pi)$ and by the expected payoff of the lowest type of the plaintiff $U_p(\underline{\pi})$. Any two indirect mechanisms, which give rise to the same function $\sigma(\pi)$ and $U_p(\underline{\pi})$ once the plaintiff optimizes, therefore imply the same expected payoff for all types of the plaintiff.

Rewrite the objective of the mediation problem as follows:

$$\begin{aligned} \min_{\sigma(\pi)} \quad & (c_p + c_d)\int_{\underline{\pi}}^{1}[1-\sigma(\pi)]\mathrm{d}F(\pi) \\ s.t. \quad & \text{(i), (ii), (IR-P), (IR-D)} \end{aligned}$$

Observe that this is a linear programming problem in the $\sigma(\pi)$ functional space. We can then apply the extreme point theorem and simplify the mediation problem further. [15] Instead of considering all weakly monotone functions, it is sufficient to restrict our attention to the set of extreme points which requires $\sigma(\pi) \in \{0, 1\}$ for all π. And an extreme point is monotone if and only if it is a step function. There exists a threshold type π^* such that

$$\sigma(\pi) = \begin{cases} 1 & \text{if } \pi \leqslant \pi^*, \\ 0 & \text{if } \pi > \pi^*. \end{cases} \tag{4}$$

The mediator always recommends an agreement for reports below π^*, and always rec-

[15] The extreme point theorem states that a function σ that is an extreme point and that maximizes the total payoffs among all extreme points also maximizes the total payoffs among all functions.

ommends the outside option for reports above π^*. It follows that

$$U_p(\pi) = \begin{cases} U_p(\underline{\pi}) & \text{if } \pi \leqslant \pi^*, \\ \pi - \pi^* + U_p(\underline{\pi}) & \text{if } \pi > \pi^*. \end{cases}$$

By definition, this implies $S(\pi) = S(\underline{\pi}) = \pi^* v - c_p$ for $\pi \in [0, \pi^*]$. Therefore,

$$S(\pi) = \begin{cases} \pi^* v - c_p & \text{if } \pi \leqslant \pi^*, \\ \pi v - c_p & \text{if } \pi > \pi^*. \end{cases} \tag{5}$$

It is straightforward to check that this satisfies (IR-P), i.e., $S(\underline{\pi}) \geqslant \pi v - c_p$ for $\pi \in [0, \pi^*]$ and $\pi v - c_p \geqslant \pi v - c_p$ for $\pi \in (\pi^*, 1]$. (IR-D) requires whenever $\pi \in [0, \pi^*]$, $S(\underline{\pi}) \leqslant \mathbb{E}[\pi | \pi \leqslant \pi^*] + c_d$. That is, for a π^* to be feasible, it has to satisfy

$$\pi^* - \mathbb{E}[\pi | \pi \leqslant \pi^*] \leqslant \frac{c_p + c_d}{v}. \tag{6}$$

Notice that so long as $c_p + c_d > 0$, such a π^* always exist. As either c_p or c_d becomes larger, this set possibly grows larger. [16]

Comparing the mediated settlement model with the signaling model, we have the following three hypotheses.

Hypothesis 3 (Platform Mediation)**.**

Hypothesis 3a. *Platform mediation adds commitment to the signals.*

Hypothesis 3b. *Platform mediation leads to a higher settlement rate.*

Hypothesis 3c. *The positive effect of platform mediation is unchanged if the plaintiff's winning probability is higher.*

[16] In general, there could be multiple π^* satisfying this condition. To have a unique π^*, we need $\pi' - \mathbb{E}[\pi | \pi \leqslant \pi']$ to be a monotone increasing function of π'. As shown by Burdett (1996), the necessary and sufficient condition would be a restriction on $\mu^0(\pi)$ such that $\int_{-\infty}^{\pi} F(x)\mathrm{d}x$ is log-concave. This can be satisfied if the distribution $F(\pi)$ is log-concave (but not necessarily). Several well-known distributions are log-concave, e.g., uniform, normal, and exponential distributions.

4 Data Description and Empirical Strategy

4.1 Data Description

We apply text analysis to examine the takedown notices and counter-notices received by GitHub. The primary objective is to gain a deeper understanding of the textual content and context surrounding these notices, uncovering any latent attributes. Copyright holders should submit GitHub's takedown notice by filling in a copyright claim form. This form functions as a questionnaire with some mandatory questions. GitHub publicly posts all takedown notices received, from 2011 to the present, after removing private information and unavailable URLs in some takedown notices.

To send a takedown notice, the plaintiffs should fill out a copyright claim contact form to answer the questions listed by GitHub. We collect 4,684 takedown notices received between March 2021 and August 2023, with 67 of them receiving counter notices. We put the monthly counts of takedown notices and counter notices in Figure 1. Within GitHub's copyright claims form, nine questions are relevant to our research objectives (see Table 1). We exclude questions related to contact information and focus on the remaining nine questions and their corresponding answers. We extract features from both the questions and anatomies, categorizing them into four categories: main features, textual characteristics, the plaintiff's precautions, and the intermediary's actions. Variables and their values under each category are in Tables 2. Table 3 contains the summary statistics of owner respondents, attorney respondents, and all respondents. Table 4 records the information on the dummy variables.

The first question on the copyright claims form inquires whether the submitter is the copyright holder or authorized to act on their behalf, such as an attorney. Among our observations, 55% of the plaintiffs are legal professionals or third-party representatives to represent the copyright holder and 45% of them are the copyright holder themselves. The dummy variable "Legal_representative" equals 1 if the notice submitter is an attorney. This is consistent with Urban et al. (2017) which finds a growing trend of professionalization when it comes to infringement searches. Copyright holders are increasingly

turning to third-party rights enforcement organizations (REOs) to locate infringing repositories and issue takedown notices. Our dataset reveals that a significant 55% of copyright holders opt to engage attorneys to assert their copyright claims. There are noticeable differences between takedown notices drafted by individual claimants and those made by REOs.

GitHub allows the authors to revise their takedown notices if their answers do not meet the platform's requirements. If plaintiffs modified and re-submitted their notices, they should report that in Question 2 and the variable "GitHub_revision" equals 1. Only 9.48% of the takedown notices were revised ones. Question 2 refers to a mediation method facilitated by GitHub.

Questions 3 and 4 are open-ended. Question 3 requires descriptions of the ownership of the copyrighted work or the authorization to act on the copyright owner's behalf. Owners should reveal their identities and describe their ownership, while lawyers should provide the lawyer-client relationship. Question 4 asks the respondents to describe the original copyrighted work. We extract five textual features from the answers to these questions: length, readability, specificity, similarity, and redundancy.

Firstly, we count the clean words in an answer to represent the length. Clean words refer to expressions that do not include stop words, punctuation, or URLs. On average, the length of responses to Question 3 is 20.48 words and that of Question 4 is 25.58 words.

Secondly, we employ the fog index to estimate the textual answers' readability of Question 3. The fog index is the weighted average of sentence length and complex word usage. Responses to question 3 exhibit an average fog index of approximately 13.8, suggesting a person with a 13th or 14th-grade education level can understand the answers.

Thirdly, We calculate Named Entity Recognition occurrences within the text to represent the specificity. NER identifies the specific mentions of named entities such as company names, organization names, locations, and more. On average, each answer contains around 2.2 entities. We use the same way to analyze the answers to Question 4. Question 4 requires a comprehensive description of the original copyrighted work allegedly subject to infringement. Responses have an average fog index of 13.13 and an average word

count of 25.58, along with a NER occurrence of 2.66.

Fourthly, we apply Term Frequency-Inverse Document Frequency (TF-IDF) and cosine similarity to measure the resemblance between answers to the open-ended questions. TF-IDF measures the importance of a term (word or n-gram) within a document relative to a collection of documents. In our context, each answer is regarded as a separate document. An n-gram is a contiguous sequence of n words. For instance, "submit a takedown notice" is a 4-gram. TF-IDF measures two components, Term Frequency (TF) and Inverse Document Frequency (IDF). TF measures how frequently a term appears within a specific document. A higher term frequency indicates that the term is important within the document. IDF measures the rarity of a term across the entire corpus. A term that appears in many documents will have a lower IDF score, while a term that appears in only a few documents will have a higher IDF score. TF-IDF helps identify keywords and important terms that might be overshadowed by common words when using TF alone. We compute TF-IDF of 4-grams in the answers. Cosine similarity is then used to calculate the distance between these TF-IDF-weighted vectors, providing a measure of similarity between documents. A higher value of cosine similarity indicates a greater similarity between two documents.

Lastly, we calculate the word redundancy, the number of the same words that are used in answers to both Question 3 and Question 4, within identical takedown notices. In total, we summarize seven variables about these two questions.

Question 5 asks claimants to provide URLs of the allegedly infringing files or repositories. Plaintiffs submit 5.72 URLs on average. Notably, the alleged infringing repositories may have forks and plaintiffs can report URLs of forks in Question 7. The average number of the reported forks is 3.48. The distributions of URL and fork numbers are uneven. For instance, the 95th percentile of the URL count is 17, but the largest value is $1,418$.

Questions 6 and 8 explore the technological precautions taken by the copyright holders. Question 6 asks the plaintiffs if the repository has Anti-Circumvention Technology to safeguard copyrighted content and approximately 15.56% of them assert its usage. Besides, according to the answers to Question 8, only 4.42% of infringed works hold an open-source license that governs usage, modification, and sharing. fall under such li-

censes. If the alleged infringing code violates the terms of the open-source license, the copyright holder has legal grounds to request its removal. Additionally, if the work is protected by effective anti-circumvention measures, it may be more challenging for infringers to bypass these measures, making it easier for the plaintiff to demonstrate a clear case of infringement. Therefore, these two variables serve as proxies for the winning rate, and having these measures indicates a higher probability of winning.

The final one, Question 9, asks for the preferred solution for alleged infringement. In our sample, 83.45% of plaintiffs have sent a strong signal by demanding GitHub delete the allegedly infringing contents rather than just making them private. If the plaintiff is resolute to delete the infringing content, the value of variable "Demand" is one, otherwise zero.

GitHub also provides annotations to help readers better understand how they processed the notice, referred to as anatomies. We consider two anatomies in this paper and represent them by dummy variables "Chance_to_change" and "GitHub_verification". Anatomy 1 records the scenario that the takedown notice either did not claim that the entire reported repository is infringing the copyrighted work, or the copyright holder suggested modifications to resolve the alleged infringement rather than directly taking it down. GitHub will contact the users of that reported repository and give them approximately one business day to delete or modify the content. In 52.33% of the cases, the defendants obtained this opportunity. When a takedown notice alleges more than one repository or file is infringing the copyrighted work and GitHub determines that some of the reported repositories or files are innocent, the platform uses Anatomy 2 to annotate this case. The percentage of this case is only 8.16%. These two variables are methods of platform mediation.

4.2 Empirical Strategy

Our main specification is to test the effects of different characteristics of a takedown notice on the settlement rate. Our outcome variable is "settled", which indicates whether the case is privately settled or might proceed to trial. In our case, "settled" is equal to 1 if no

counter-notice is received, and 0 otherwise. We use three alternative regression models: logistic regression, probit regression, and complementary log-log (cloglog) regression. A positive coefficient associated with an independent variable in the logistic and probit models indicates a propensity for that feature to increase the settlement rate. In logistic regression, a one-unit change in a predictor changes the dependent variable's log odds by its coefficient value when keeping other variables constant. The coefficients in a probit model represent the change in the z-scores, or standard deviations, of the dependent variable for a one-unit change in the predictor, holding all other variables constant.

In our dataset, a trial is a rare event: there are 67 counter-notices among the 4,684 takedown cases, and therefore, only 1.43% of the takedown notices prompted a counter-notice. This leads to an imbalance issue in the sample, which occurs if the distribution of the dependent variable is heavily skewed towards one of the outcomes. The imbalance of the outcome variable can be challenging for estimation because the logistic and probit regressions might not be robust to rare events.

To address the robustness of our results, we employ two statistical techniques: the cloglog regression and resampling. The asymmetric transformation of the probability in the cloglog model (Cox, 1972; McCullagh, 1980) assigns more weight to the tails of the probability distribution, so it focuses on modeling the probabilities associated with rare events. Cloglog helps to address the issue of unbalanced data. When conducting the cloglog regression, we interchange the values of 0 and 1 for the dependent variable since $Y = 1$ represents the minority in this model. The coefficients indicate the extent to which a one-unit change in the predictor influences the change in the cloglog transformation of the probability of a takedown notice being countered. [17] A negative coefficient implies that the associated variable reduces the likelihood of a counter-notice.

To fix the imbalance issue and a potential bias towards the majority class, we also re-sample the observations by applying the Synthetic Minority Over-sampling Technique (SMOTE) proposed by Chawla et al. (2002). The SMOTE algorithm finds the neighbors in the feature space of the minority class and creates new instances based on the properties of those minority samples, the takedown notices that received a counter-notice in our

[17] The complementary log-log transformation function is defined as $cloglog(p) = log(-log(1-p))$.

case. We begin by employing SMOTE to extend the number of samples in the minority class to reach half the number of the majority class.

5 Empirical Results

In this section, we present our empirical results on the determinants of pretrial settlement. Three factors show prominence in facilitating settlement: text features of the notices, legal representation, and platform mediation. We dig into the mechanisms and possible heterogeneous effects.

Our statistical tests and regression analysis support the main predictions of the three signaling models. Legal representation, platform mediation, and lower demand are all found to increase the settlement rate. This is consistent with the three hypotheses on the settlement rate - Hypotheses 1a, 2b, and 3b. We also find suggestive evidence for the two hypotheses on how institutional factors change the signals - Hypotheses 2a and 3a. The results for the heterogeneous effect with regard to winning rates are mixed. Hypothesis 3c in the mediated settlement is supported, while the evidence for Hypotheses 1b and 2c is mixed. Part of the reason is that we do not have an indisputable measure of winning rates because of data limitations.

5.1 Baseline and Text Features

5.1.1 Baseline Features

In the baseline model, we regress the outcome variable on three fundamental features of a notice - who writes the notice (the owner or the attorney), the number of infringing URLs reported, and the demand of the plaintiff.

$$Settled_i = \beta_1 Legal\_representative_i + \beta_2 Infringing\_URL\_count_i + \beta_3 Demand_i + \varepsilon_i \quad (7)$$

The regression results are reported in Table 5. In Column (1), legal representation leads to an approximately 2.99 times increase in the settlement rate. Likewise, the z-score associated with the settlement rate rises by 0.43 standard deviation as shown in Columns (2). Column (3) shows that the log-transformed hazard of being countered decreases by 1.09. The results support Hypothesis 2b, which suggests that hiring an experienced copyright lawyer contributes significantly to the plaintiff's ability to settle the dispute. The effect of locating more infringing URLs, though, is not statistically significant. This suggests that identifying more files or repositories as infringing does not contribute significantly to dispute resolution. In terms of the respondent's demand, requesting a removal of the allegedly infringing work, instead of making it private, significantly reduces the odds of settlement by 54.51%. Columns (2) and (3) indicate a 0.31 standard deviation decrease in the z-score of the settlement probability and a 0.78 increase in the log-transformed hazard of being countered are associated with a demand of removal. Hence, the empirical results provide support for Hypothesis 1a, which postulates that the more demanding the plaintiff is, the lower the settlement rate becomes. Table 6 displays the regression results with a re-sampling of minority observations. The signs and significance of the three variables remain consistent before and after re-sampling. This confirms the robustness of our findings.

5.1.2 Precautionary Technology and Licensing as Winning Rate Proxy

In our sample, a fraction of copyright holders adopt anti-circumvention technologies and open-source licenses as protective measures for their works. DMCA strictly prohibits the circumvention of technical measures that control access to copyrighted works. Hence, having anti-circumvention in place will greatly increase the plaintiff's chance to prevail in court. In contrast, having an open-source license will put the plaintiff at a disadvantage because many defense strategies are available such as fair use and public domain. Therefore, an open-source license lowers the winning rate. In what follows, we use these two measures as proxies for the winning rate. To test Hypothesis 1b, we run regression 8

to check whether the settlement rate changes with the winning rate.

$$Settled_i = \beta_1 Anti\_circumvention_i + \beta_2 License_i + \beta_3 X_i + \epsilon_i \tag{8}$$

The coefficients on both variables are significantly positive after re-sampling in Table 8. The presence of an open-source license effectively doubles the odds of settlement, elevating the z-score by 0.72 standard deviations, and reducing the log-transformed hazard of encountering a counter notice by 0.96. Figure 6 illustrates the robustness of coefficient values as the re-sampling size increases. This supports Hypothesis 1b which predicts that the settlement rate is higher when the plaintiff has a lower winning probability. An open-source license not only contributes more content to the public domain, in this case also becomes a focal point enabling the two parties to settle.

However, when the copyright holder has adopted anti-circumvention measures, settlement odds increase by a remarkable factor of 6.8, accompanied by a rise of 1.18 standard deviations in the z-score. The log-transformed hazard of being countered declines by 1.76. This contradicts Hypothesis 1b as a higher winning rate increases the settlement rate. Overall, the regression analysis provides mixed evidence on Hypothesis 1b of the signaling model.

5.1.3 Textual Features as a Determinant of Settlement Rate

We use equation 9 to check the effects of the readability, specificity, length, and redundancy of the answers to Questions 3 and 4 on the settlement rate and record the results in Table 9. X_i represents the three baseline variables in the previous section.

$$\begin{aligned} Settled_i = &\beta_1 Ownership\_fog_i + \beta_2 Ownership\_NER_i + \beta_3 Ownership\_word_i \\ &+ \beta_4 Infringement\_fog_i + \beta_5 Infringement\_NER_i + \beta_6 Infringement\_word_i \\ &+ \beta_7 Redundancy_i + \beta_8 X_i + \varepsilon_i \end{aligned} \tag{9}$$

The coefficients associated with textual features of the answer to Question 3, the ownership or authorization description, are presented in Columns (1) to (3) of Table 9. Although

in all three regressions, we observe a close-to-zero and insignificant coefficient for the fog index, those coefficients become significant and have larger magnitudes in Table 10 after re-sampling. Balanced samples indicate that when the response requires an additional year to understand, the log odds of settlement decrease by 0.03. Lower readability weakens the signal and adversely affects the settlement rate.

Both unbalanced and re-sampled datasets confirm the significance of specificity and length. When the author mentions one additional entity, we see an 11.82% increase in the odds of settlement, a corresponding z-score increase of 0.05 standard deviations, and a 0.11 reduction in the log-transformed hazard of being countered. Specificity aids in highlighting key elements such as names, dates, organizations, and especially legal terminology, thereby potentially making the claims more assertive and informative. For instance, the following response explicitly identifies the company names of the owner and legal representative:

> "The copyrighted work that is reproduced and made available for copying by the repository identified below is **Apple's** internal unreleased software source code. The copyright therein is owned by **Apple**. I am an attorney with **Reed Smith LLP**, which advises **Apple Inc.** in certain intellectual property matters. I am authorized by **Apple** to act on their behalf in this matter."

The preceding response contains more detailed information compared to the one below:

> "I am an employee of the copyright owner."

An overly lengthy description is not advisable, as adding just one more word reduces the probability of settlement by 1.33%, lowers the z-score by 0.0057 standard deviations, and increases the log-transformed hazard of being countered by 0.01. These findings suggest that an inclination to provide an excessive description of the issue, especially by the attorneys may be counterproductive.

Columns (4) to (6) in Table 9 show how the answer to Question 4, the description of the copyrighted work, influences the settlement rate. In all three models, the coefficients

for both the fog index and the number of NER are nearly zero and statistically insignificant. However, these coefficients become significant in logistic and probit regressions after re-sampling. Columns (4) to (5) in Table 10 reveal that an additional year of comprehension requirement for the response leads to a minimal increase in the log odds of settlement and a 0.01 standard deviation rise in the settlement's z-score. The coefficient on the NER count becomes significantly positive after re-sampling. According to the re-sampled data, each additional named entity contributes to a 0.03 increase in the log odds of settlement, a 0.02 standard deviation improvement in the settlement's z-score, and a 0.02 reduction in the log-transformed hazard of receiving a counter notice. Hence, the specificity in the description of the copyrighted work strengthens the plaintiff's signal. Furthermore, the length has a negative effect on the settlement rate as it does in answers to Question 3. In Table 9, an additional word decreases the odds of settlement rate by 0.93% and reduces the rate by 0.0046 standard deviations in the z-score and 0.0034 in the log-transformed hazard. The results are robust after re-sampling. A too-lengthy answer negatively impacts the settlement outcome.

About the redundancy, in Columns (7) to (9) of Table 9, only the cloglog regression shows a significant coefficient at the 10% confidence level, whereas the logistic and probit models yield significantly positive results after re-sampling in Columns (7) and (8) of Table 10. The inclusion of one repeated word in the two responses leads to a 0.07 increase in the settlement's log odds and a 0.04 standard deviation improvement in the z-score. The redundancy helps bolster the plaintiff's signal. Additionally, the six aforementioned textual features maintain their signs and significance when included together in the regressions in Columns (7) to (9) of Table 9 and Columns (7) and (8) of Table 10, confirming the robustness of these coefficients.

In summary, our measure of text features is an important determinant of the settlement rate. To effectively convey a strong signal and avoid being countered, respondents should include more pivotal information (specificity), write concisely (length), and make the descriptions more reader-friendly (readability).

5.2 The Role of Legal Representation

5.2.1 Comparing Attorney-Written and Owner-Written Notices

We see previously that legal representation has a significant positive effect on settlement rate. In this section, we investigate the mechanism behind it. It turns out that the way an attorney drafts the notice varies greatly from that of the owners.

We begin by examining whether the text attributes between attorney submitters and owner submitters are different by Welch's t-test. This tests Hypothesis 2a, which investigates whether an attorney helps to increase the credibility of the signal. The Welch's t-test assumes that the variances of numeric features between the two types of submitters are not equal. Figure 2 illustrates the differences between owners and attorneys and the corresponding confidence intervals. The difference is statistically significant if the confidence interval does not include zero.

In panel A of Figure 2, all variables are binary. Owners who hire an attorney for the takedown notice are more inclined to use anti-circumvention technology but less likely to obtain an open-source license. Attorneys are more likely to ask for the removal of the reported work than the owners. Regarding the intermediary's actions, owner respondents' takedown notices are more likely to lead to opportunities to modify their repositories to prevent content removal. There is no significant difference in the likelihood of revising and re-submitting the takedown notice between the two types of respondents. The number of takedown notices submitted by attorneys, where not all repositories or files are found to be infringing, is lower compared to those submitted by owners, as attorneys conduct more thorough investigations.

There is a noteworthy distinction regarding the textual attributes of the responses to the two open-ended questions on ownership and infringement. In panel B of Figure 2, attorney respondents, on average, exhibit a higher fog index and use a greater number of named entities in their answers to both Question 3 and Question 4 when compared to owner respondents. Specifically, the fog indices for the two descriptions, as reported by attorneys, exceed those provided by owners by 1.41 and 2.15 units, respectively. This im-

plies that the comprehension of texts generated by attorneys requires an additional one to two years of education in comparison to owner-authored texts, making them harder to read. Furthermore, attorneys use 0.86 and 1.56 more named entities in their responses to Questions 3 and 4, respectively. The attorneys mention more names of individuals, locations, programs, and companies to help readers understand the parties and the alleged infringement involved. For example,

> "... It's an **OPL-1** license under the **Odoo Proprietary License v1.0**. This software and associated files (the "Software") may only be used (executed, modified, executed after modifications) if you have purchased a valid license from the authors, typically via **Odoo Apps**, or if you have received a written agreement from the authors of the Software. You may develop **Odoo** modules that use the Software as a library (typically by depending on it, importing it and using its resources), but without copying any source code or material from the Software. You may distribute those modules under the license of your choice, provided that this license is compatible with the terms of the **Odoo Proprietary License** (For example: **LGPL**, **MIT**, or proprietary licenses similar to this one)..."

NER assists in clarifying which software license should be acquired for usage and which licenses allow the distribution of modules. It is helpful when assessing whether a defendant's repository infringes on copyright or not. Additionally, attorney respondents use 0.86 more repeated words in their answers to Questions 3 and 4 compared to owner respondents.

In panel C of Figure 2, the length of responses excluding stop words, numbers, and punctuation is also significantly different. Attorneys, on average, write 4.08 more words than owners in their responses to Question 3, and this disparity expands to 8.58 words for Question 4. Besides, attorneys incorporate 0.86 more repeated words, referred to as redundancy, in their responses to those two questions. While this difference is statistically significant, the magnitude of the 0.86-word discrepancy may not be substantial enough to capture readers' attention or affect the settlement rate as we show in the following regression results. Furthermore, attorneys demonstrate a heightened commitment to identify-

ing infringing repositories and their associated forks. On average, attorneys report 2.68 more repositories and 2.59 more forks in comparison to owners.

Panel D demonstrates the similarity between answers. The attorney-written responses have greater similarity compared to those written by the owners in the descriptions of both ownership and the copyrighted work. Attorneys often introduce their authorization and the information of their law firms and employ a standardized process of investigation. This leads to a higher similarity of attorney-written answers.

In conclusion, the results of the t-tests support Hypothesis 2a. The legal representative helps amplify the credibility of the signals in two ways. First, attorney-authored notices include more pivotal information in the descriptions regarding ownership and copyrighted work. Second, the legal representative conducts a more thorough investigation and detects more infringing URLs and forks than the owners themselves.

5.2.2 Heterogeneous Effects

What explains the effect of legal representation? We know that owners and attorneys differ significantly in writing styles. But the attorney also investigates more and has more aggressive demands. Which of these differences explain the benefit of hiring an attorney?

We explore this question by testing whether the effect of legal representation is a function of other related variables. We start with textual features. We examine whether the attorney effect is a function of textual features by including the interaction terms between the legal representative and the textual features in equation 9. In Table 13, none of the coefficients on the interaction terms between legal representatives and textual features are significant. In fact, an overly complex and lengthy notice weakens the attorney's signal. In Table 14, some coefficients are significant after re-sampling. If the answer to Question 3 requires an additional year of education for comprehension, it leads to a decrease of 0.07 in the log odds of the settlement rate, a reduction of 0.04 in the z-score, and an increase of 0.06 in the log-transformed hazard of being countered. The inclusion of one additional word by the attorney results in a 0.02 decrease in the log odds of settlement, a 0.01 standard deviation drop in the z-score, and a 0.01 increase in the log-transformed

hazard of being countered. If we keep the textual features fixed, legal representative has no additional effects on facilitating the settlement. Therefore, writing styles do not contribute significantly to the positive effect of legal representation.

If anything, the attorney respondents' writing styles, actually hurt the effectiveness of the signaling. Lawyers prefer long sentences, big words, and more terminology, whereas an effective notice is much more concise. Here are two excerpts describing the copyrighted work, one from an attorney and the other from an owner. They have similar textual features but the latter was more effective.

> **Attorney**: "It has come to our attention that a project containing copyrighted content of our project which was published on GitHub 21 days ago on [private] account with the title of Online-Airline-Platform is infringing our copyright right. We have tried to contact the account owner but due to lack of contact details, we were unable to do so. Since, the project account is the identical version of a copyrighted project, which exposes our organization's proprietary data, we request GitHub to get this project removed from [private] account permanently from GitHub without leaving any trail or backup."

> **Owner**: "A developer of mine has accidentally published a GitHub repository that is part of my business. The developed code is part of a copyright and is subject of an upcoming investment. It falls under the German Business Secret Protection Act. Any sort of making this code available for the public is harmful for the future of our upcoming company. Unfortunately, this data breach has led to a user creating a fork of an older version."

Next, we examine whether legal representation has an additional effect on settlement when the attorneys have different demands by incorporating an interaction term between the legal representative and the respondent's demand. In Column (1) of Table 11, the log odds of settlement show a decrease of 17.06 units, but that implies almost a zero change in odds. The coefficient from the probit model is not statistically significant either in Columns (2). The results in Table 12 after re-sampling align with those obtained

before re-sampling. Hence, we do not observe any additional demand effect of an attorney on the settlement rate. An unyielding attorney does not enhance the strength of the plaintiff's signal. Legal demand does not explain the attorney effect. In other words, if the owner is determined to remove the infringing content anyway, the attorney cannot mitigate the negative impact of higher demand on settlement.

The merit of hiring a lawyer appears to rest on the legal expertise and the efforts put into legal research. We test whether the attorney effect is a function of the investigation efforts by introducing the interaction terms between legal representation and the investigation effort - the number of reported infringing URLs and whether they reported forks. Although reporting more infringing URLs or forks does not help the settlement, when the investigation of URLs and forks is conducted by attorneys, it has a positive impact on the settlement rate. The first two columns in Tables 21 and 22 show that an attorney's investigation of infringing URLs significantly increases the settlement rate. After re-sampling, reporting one additional infringing URL results in an increase of 0.15 in the log odds of settlement and a 0.08 increase in the z-score of settlement. The effect of reporting a fork by the attorney becomes significant after re-sampling. In Columns (3) and (4) of Table 22, reporting a fork by the attorney increases the log odds of settlement by 0.02 and the z-score of settlement by 0.01 standard deviations. The magnitude of this effect is smaller than that of reporting an infringing URL. Since attorneys put in more investigative efforts than owners, their investigation of infringing content, especially infringing URLs, enhances the signal and increases the settlement rate.

Last but not least, we test Hypothesis 2c by including the interaction between the legal representative and the precautionary measures. Before re-sampling, the interaction terms show no significance, as indicated in Table 23. Following re-sampling, the interaction term between the legal representative and anti-circumvention becomes significant in Columns (1) and (3) of Table 24. Nevertheless, the coefficient from the probit regression remains insignificant in Column (2). Additionally, the interaction between the legal representative and open-source license consistently lacks significance. Consequently, we find mixed evidence on whether the positive effect of legal representation is amplified when the plaintiff's winning probability is higher.

5.3 The Role of Intermediary

5.3.1 Comparing Revised and Original Notices

GitHub requires the plaintiffs to revise and re-submit the takedown notice to ensure compliance with the platform's regulations. As depicted in Panels A and B of Figure 3, the revision results in a significant increase in readability, specificity, length, and redundancy within the answers to the two open-ended questions. Additionally, plaintiffs tend to report a greater number of infringing URLs and forks in their revised notices. Panel C illustrates that plaintiffs who submit revised takedown notices tend to be more demanding in requesting the removal of reported content compared to those who were not required to revise their notices. Furthermore, fewer opportunities are allowed for the defendants to avoid removal by modifying their repositories in response to these revisions. All these changes collectively indicate that the commitments for the revised notices are considerably stronger than those in the original ones.

Panel D illustrates the similarity between the answers to the two open-ended questions. After revision, the similarities among both the ownership descriptions and infringement descriptions are higher. Platform mediation pools the strong signals. Strong notices are revised to be stronger and more detailed, and weak notices are more diverse. This suggests that platform mediation adds commitment to the signals, which supports 3a.

5.3.2 The Effects of Platform Mediation

We test Hypothesis 3b by exploring whether the proactive involvement of the intermediary platform in the pretrial phase has an impact on the settlement rate. The regression model 10 is as follows.

$$\begin{aligned} Settled_i =& \beta_1 GitHub\_revision_i + \beta_2 Chance\_to\_change_i + \beta_3 GitHub\_verification_i \\ &+ \beta_4 X_i + \epsilon_i \end{aligned} \tag{10}$$

In Table 18, Column (1) shows a significant fourfold increase in the odds of settlement when the takedown notice is a revised one. To add more validity, we also check the robustness of the coefficient for various re-sampling sizes. In our earlier regressions, we re-sampled minority samples (countered takedown notices) by adjusting their number to a 50% proportion relative to those that did not receive counter-notices. Figure 7 demonstrates that as the re-sampling size increases, the coefficient on GitHub revision in the logistic regression transitions from slightly negative to positive and eventually stabilizes. Therefore, the revision policy by the platform contributes to dispute resolution.

When the defendant is granted a 24-hour window to make modifications before GitHub takes down the reported repository, the odds of settlement surge by a substantial 53.75%, the z-score rises by 0.31 standard deviations, and the log-transformed hazard of being countered decreases by 0.19 as shown in Table 17. The re-sampled data also yield significantly positive coefficients. This grace period, provided by the intermediary platform, can substantially reduce the probability of the takedown notice being countered during the pretrial process. The platform serves as a mediator by assessing allegedly infringing content before taking any actions. GitHub provides defendants with the chance to amend their repositories to address copyright concerns, thereby averting immediate removal and enhancing the settlement rate.

Furthermore, GitHub's examination of URLs before removal plays a role in dispute resolution. The coefficient on GitHub verification becomes significant after re-sampling, and Figure 7 validates its sign and significance. In Columns (1) to (3) of Table 18, when non-infringing URLs are present in a takedown notice, the odds of settlement increase fourfold, the z-score of settlement rises by 0.97 standard deviations, and the log-transformed hazard of being countered decreases by 1.25. Once more, the mediator's investigation contributes to an increased settlement rate. The results for all three variables support Hypothesis 3b, indicating that platform mediation leads to a higher settlement rate.

5.3.3 Heterogeneous Effects

We examine three heterogeneous effects associated with GitHub's revision policy. First, we introduce the interaction term between the legal representative and the GitHub revision into equation 10. The coefficients related to these interaction terms become significant after re-sampling. Columns (4) to (6) in Table 18 show that if GitHub requires an attorney to modify the takedown notice, the settlement odds decline by 50% and the z-score drops down by 0.54. This implies that the attorneys should finish the takedown notice as effectively as possible when they first submit it.

Second, we delve into how textual features affect the effectiveness of the revision. We consider interaction terms between GitHub's revision and the seven textual variables related to the responses to Questions 3 and 4. In Tables 19 and 20, only the interaction term between the GitHub revision and the fog index of the response to Question 4 is significant. In Columns (4) to (6) of Table 19, if the response to Question 4 in a revised takedown notice demands an additional year of education for comprehension, we observe a 10.67% decrease in the odds of settlement, a corresponding reduction of 0.05 standard deviations in the z-score, and a rise of 0.11 in the log-transformed hazard of being countered. The results imply that altering the writing style in the revised takedown notice does not contribute to enhancing the plaintiff's signal while maintaining a simplified readability in the description of copyrighted work helps.

Lastly, we test whether the effect of platform mediation varies with winning rate proxies. None of the interaction terms between the precautionary measures and GitHub revision demonstrate significance. When there is a higher winning probability, the positive effect of platform mediation remains unchanged. This supports Hypothesis 3c.

5.4 Robustness Test

We apply two methods to perform the robustness test. Firstly, we assess the stability of all individual coefficients by testing various proportions of the minority to the majority. Figures 4 to 7 depict the coefficient trends and their 95% confidence intervals. As the mi-

nority cohort size expands, the coefficient values converge to a stable level. Additionally, the sign and significance remain consistent with the results mentioned earlier.

Secondly, we employ the complementary log-log regression as a means to validate the robustness of the results obtained from the logistic and probit regressions. The asymmetric nature of the inverse cumulative distribution function in cloglog helps mitigate the dataset's imbalance issue. It's important to note that due to the different coding strategies of the dependent variable between cloglog and the other two regressions, a negative sign in cloglog corresponds to the same interpretation as a positive sign in logistic and probit. In the empirical results, we observe that all signs and the majority of significance in cloglog are consistent with those in logistic and probit.

6 Conclusion

In this paper, we test pretrial signaling in two ways: (i) we apply text analysis to quantify the information conveyed in the signals, and (ii) we extend the signaling model to derive testable hypotheses and check whether they are consistent with empirical findings. The availability of online copyright notices and counter-notices posted by GitHub has provided us with a rich textual dataset for analysis. In our text analysis, we measure every copyright notice by five attributes including length, readability, specificity, similarity, and redundancy. Studying these text features helps us to understand what a strong pretrial signal is. A strong signal is concise, easy to read, and more specific.

Our second step is to show how different institutional factors change the pretrial signals and the settlement rate. We analyze and compare three related models of pretrial settlement: the baseline signaling model, the disclosure model, and the mediated settlement model. Both legal representation and platform mediation strengthen the signal and help to close the information gap between the disputants. The "attorney effect" is best explained by investigation efforts rather than writing techniques or a tougher attitude. In fact, lawyers can make their writing more effective by avoiding long sentences, big words, and excessive terminologies.

We conclude our paper by discussing a few limitations and promising extensions. One data limitation is that we do not observe the actual litigation if the two parties proceed to court. Because of that, we do not have good estimates of the stake of the case, the winning rate, etc., which prevents us from testing more implications of the models. Our empirical test is most likely applicable in disputes between strangers. Most of the disputants in our setting have no prior contractual relationship. Reputation and relationship-specific investment play little role in our setting, which might be critical to settlement in contract cases.

It would be interesting to see whether our findings generalize to copyright notices in other platforms or other kinds of online takedown notices. Google, for example, publishes annual transparency reports on content delisting, government requests, security and privacy. There are also takedown requests based on trademark, defamation, domain name ownership, etc. This data has the potential to offer many more insights into how disputes are resolved online and whether the pretrial bargaining models are relevant. Our study is just a glimpse of the agenda.

References

Ahn, Kiefer, Antonio Trujillo, Jason Gibbons, Charles L. Bennett, and Gerard Anderson, "Settled: Patent characteristics and litigation outcomes in the pharmaceutical industry," *International Review of Law and Economics*, 2023, *76*, 106169.

Bebchuk, Lucian Arye, "Litigation and settlement under imperfect information," *The RAND Journal of Economics*, 1984, pp. 404–415.

Burdett, Kenneth, "Truncated means and variances," *Economics Letters*, 1996, *52* (3), 263–267.

Chang, Yunchien and William Hubbard, "New Empirical Tests for Classic Litigation Selection Models: Evidence from a Low Settlement Environment," *American Law and Economics Review*, 10 2021, *23* (2), 348–394.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer, "SMOTE: Synthetic Minority over-Sampling Technique," *Journal of Artificial Intelligence Research*, 2002, *16* (1), 321–357.

Choi, Jonathan H, "How to use large language models for empirical legal research," *Journal of Institutional and Theoretical Economics (Forthcoming)*, 2023.

Cox, D. R., "Regression Models and Life-Tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1972, *34* (2), 187–202.

Danzon, Patricia Munch and Lee A. Lillard, "Settlement out of Court: The Disposition of Medical Malpractice Claims," *The Journal of Legal Studies*, 1983, *12* (2), 345–377.

Dyer, Travis, Mark Lang, and Lorien Stice-Lawrence, "The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation," *Journal of Accounting and Economics*, 2017, *64* (2), 221–245.

Earnhart, Dietrich and Sandra Rousseau, "Are lawyers worth the cost? Legal counsel in environmental criminal court cases," *International Review of Law and Economics*, 2019, *60*, 105857.

Eisenberg, Theodore and Henry S. Farber, "The Litigious Plaintiff Hypothesis: Case Selection and Resolution," *The RAND Journal of Economics*, 1997, *28*, S92–S112.

Farmer, Amy and Paul Pecorino, "Civil litigation with mandatory discovery and voluntary transmission of private information," *The Journal of Legal Studies*, 2005, *34* (1), 137–159.

Fournier, Gary M. and Thomas W. Zuehlke, "Litigation and Settlement: An Empirical Approach," *The Review of Economics and Statistics*, 1989, *71* (2), 189–195.

Grimmelmann, James and Pengfei Zhang, "An Economic Model of Intermediary Liability," *Berkeley Technology Law Journal, Forthcoming*, 2023.

Grimmer, Justin, Margaret E Roberts, and Brandon M Stewart, *Text as data: A new framework for machine learning and the social sciences*, Princeton University Press, 2022.

Grossman, Sanford J, "The informational role of warranties and private disclosure about product quality," *The Journal of Law and Economics*, 1981, *24* (3), 461–483.

Gunning, R, *The Technique of Clear Writing*, McGraw-Hill, New York, 1952.

Hörner, Johannes, Massimo Morelli, and Francesco Squintani, "Mediation and peace," *The Review of Economic Studies*, 2015, *82* (4), 1483–1501.

Huang, Kuo-Chang, "How Legal Representation Affects Case Outcomes: An Empirical Perspective from Taiwan," *Journal of Empirical Legal Studies*, 2008, *5* (2), 197–238.

Jovanovic, Boyan, "Truthful disclosure of information," *The Bell Journal of Economics*, 1982, pp. 36–44.

Keller, Daphne, "Empirical Evidence of 'Over-Removal'by Internet Companies under Intermediary Liability Laws," *The Center for Internet and Society at Stanford Law School*, 2015.

_ **and Paddy Leerssen**, "Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation," *Social media and democracy: The state of the field and prospects for reform*, 2020, *220*, 224.

Klein, Benjamin, Andres V Lerner, and Kevin M Murphy, "The economics of copyright" fair use" in a networked world," *American Economic Review*, 2002, *92* (2), 205–208.

Klerman, Daniel and Lisa Klerman, "Inside the Caucus: An Empirical Analysis of Mediation from Within," *Journal of Empirical Legal Studies*, 2015, *12* (4), 686–715.

Lee, Yoon-Ho Alex and Daniel Klerman, "The Priest-Klein hypotheses: Proofs and generality," *International Review of Law and Economics*, 2016, *48*, 59–76.

Lerner, Josh and Jean Tirole, "Some simple economics of open source," *The journal of industrial economics*, 2002, *50* (2), 197–234.

_ and _, "The economics of technology sharing: Open source and beyond," *Journal of Economic Perspectives*, 2005, *19* (2), 99–120.

Lessig, Lawrence, *Code: And other laws of cyberspace*, ReadHowYouWant. com, 2009.

McCullagh, Peter, "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1980, *42* (2), 109–127.

Meurer, Michael J, "The settlement of patent litigation," *The RAND Journal of Economics*, 1989, pp. 77–91.

Myerson, Roger B, "Optimal auction design," *Mathematics of operations research*, 1981, *6* (1), 58–73.

Pecorino, Paul and Mark Van Boening, "An Empirical Analysis of the Signaling and Screening Models of Litigation," *American Law and Economics Review*, 04 2018, *20* (1), 214–244.

_ and Mark Van Boening, "An empirical analysis of litigation with discovery: The role of fairness," *Journal of Behavioral and Experimental Economics*, 2019, *81*, 172–184.

Penney, Jonathon W, "Privacy and Legal Automation: The DMCA as a Case Study," *Stan. Tech. L. Rev.*, 2019, *22*, 412.

Perloff, Jeffrey M., Daniel L. Rubinfeld, and Paul Ruud, "Antitrust Settlements and Trial Outcomes," *The Review of Economics and Statistics*, 1996, *78* (3), 401–409.

Poppe, Emily S. Taylor and Jeffrey J. Rachlinski, "Do Lawyers Matter? The Effect of Legal Representation in Civil Disputes," *Pepp. L. Rev.*, 2016, *43*, 881–944.

Priest, George L and Benjamin Klein, "The selection of disputes for litigation," *The journal of legal studies*, 1984, *13* (1), 1–55.

Reinganum, Jennifer F and Louis L Wilde, "Settlement, litigation, and the allocation of litigation costs," *The RAND Journal of Economics*, 1986, pp. 557–566.

Seng, Daniel, "Copyrighting Copywrongs: An Empirical Analysis of Errors with Automated DMCA Takedown Notices," *Santa Clara High Tech. LJ*, 2020, *37*, 119.

Shavell, Steven, "Sharing of information prior to settlement or litigation," *The RAND Journal of Economics*, 1989, pp. 183–195.

Silveira, Bernardo S., "Bargaining With Asymmetric Information: An Empirical Study of Plea Negotiations," *Econometrica*, 2017, *85* (2), 419–452.

Somaya, Deepak, "Strategic determinants of decisions not to settle patent litigation," *Strategic Management Journal*, 2003, *24* (1), 17–38.

Spier, Kathryn E, "Litigation," *Handbook of law and economics*, 2007, *1*, 259–342.

Stanley, Linda R. and Don L. Coursey, "Empirical Evidence on the Selection Hypothesis and the Decision to Litigate or Settle," *The Journal of Legal Studies*, 1990, *19* (1), 145–172.

Urban, Jennifer M and Laura Quilter, "Efficient process or chilling effects-takedown notices under Section 512 of the Digital Millennium Copyright Act," *Santa Clara Computer & High Tech. LJ*, 2005, *22*, 621.

_ , Joe Karaganis, and Brianna Schofield, "Notice and takedown in everyday practice," *UC Berkeley Public Law Research Paper*, 2017, (2755628).

Viscusi, W. Kip, "Product Liability Litigation with Risk Aversion," *The Journal of Legal Studies*, 1988, *17* (1), 101–121.

Vu, Duy, "To Settle or to Fight to the End? Case-level Determinants of Early Settlement of Investor-State Disputes," *Review of Law Economics*, 2021, *17* (1), 133–166.

Waldfogel, Joel, "The Selection Hypothesis and the Relationship between Trial and Plaintiff Victory," *Journal of Political Economy*, 1995, *103* (2), 229–260.

_ , "Reconciling Asymmetric Information and Divergent Expectations Theories of Litigation," *The Journal of Law Economics*, 1998, *41* (2), 451–476.

_ , "Copyright research in the digital age: Moving from piracy to the supply of new products," *American Economic Review*, 2012, *102* (3), 337–42.

Ye, Bin, Shi Jinchuan, and Ma Yiran, "Narrower Door to Approach the 50Cost-Benefit-Driven Litigation and Plaintiff Win Rate," 2023.
Available at SSRN: https://ssrn.com/abstract=4501574 or http://dx.doi.org/10.2139/ssrn.4501574.

Yildiz, Muhamet, "Bargaining with optimism," *Annu. Rev. Econ.*, 2011, *3* (1), 451–478.

Zhang, Pengfei, "Piracy or Fair Use? Evaluating the Welfare of Software Copyright Takedowns: Theory and Evidence from GitHub," *Available at SSRN 4274527*, 2021.

Appendix A: Figures and Tables

Table 1: Selected Questions and Anatomies of the GitHub Takedown Notice

	Questions on the Takedown Notice
Q1	Are you the copyright holder or authorized to act on the copyright owner's behalf?
Q2	Are you submitting a revised DMCA notice after GitHub Trust & Safety requested you make changes to your original notice?
Q3	Please describe the nature of your copyright ownership or authorization to act on the owner's behalf.
Q4	Please provide a detailed description of the original copyrighted work that has allegedly been infringed. If possible, include a URL to where it is posted online.
Q5	What files should be taken down? Please provide URLs for each file, or if the entire repository, the repository's URL.
Q6	Do you claim to have any technological measures in place to control access to your copyrighted content? Please see our Complaints about Anti-Circumvention Technology if you are unsure.
Q7	Have you searched for any forks of the allegedly infringing files or repositories? Each fork is a distinct repository and must be identified separately if you believe it is infringing and wish to have it taken down.
Q8	Is the work licensed under an open source license?
Q9	What would be the best solution for the alleged infringement?
A1	GitHub gave repository owners a chance to make changes before we processed the notice.
A2	GitHub only processed the takedown notice with respect to some of the reported URLs.

We select a set of questions from the DMCA takedown notice on GitHub. We have omitted questions that request contact information or other private details and have retained those that are related to the plaintiff's properties, the copyrighted work, and the alleged infringement. These nine questions are mandatory to complete by the respondents. A1 and A2 are annotations provided by GitHub to help readers understand how the GitHub team handled the notice and the outcomes of their investigation into the reported infringing content.

Table 2: Variables and Their Values

Question	Variable	Attributes
Main Features		
Q1	Legal_representative	1 if attorney; 0 if owner
Q9	Demand	1 if takedown is mandatory; 0 other solutions are acceptable
Q5	Infringing_URL_count	Number of infringing repositories that were reported
Q7	Fork	1 if the infringing repositories were forked; 0 otherwise
Textual Characteristics		
Q3	Ownership_fog	Fog index
	Ownership_NER	Percentage of NER
	Ownership_word	Count of clean words
Q4	Infringement_fog	Fog index
	Infringement_NER	Percentage of NER
	Infringement_word	Count of clean words
	Redundancy	Number of words used in answers to both Q3 and Q4
Precaution		
Q6	Anti_circumvention	1 if yes; 0 if no
Q8	License	1 if yes; 0 if no
Intermediary		
Q2	GitHub_revision	1 if revised; 0 otherwise
Anatomy 1	Chance_to_change	1 if the infringing repositories were allowed to change; 0 otherwise
Anatomy 2	GitHub_verification	1 if not all reported repositories were infringing; 0 otherwise
Dependent variable	Settled	Logit/Probit: 1 if the takedown notice was not countered; 0 otherwise Cloglog: 1 if the takedown notice was countered; 0 otherwise

This table presents an overview of all variables extracted from the takedown notices and counter notices. As Questions 3 and 4 require open-ended responses, more than one variable is extracted from them. In the second column, we list the variable names. The third column provides detailed descriptions of the variables and the values of dummy variables. We have categorized these variables into four groups: baseline features, textual features, intermediary actions, and precautionary measures.

Table 3: Summary Statistics

Variable	Respondent	Mean	Std. Dev.	Min.	5%	25%	50%	75%	95%	Max.
Ownership_fog	Attorney	14.44	5.58	0	3.2	11.47	14.43	17.93	23.29	30.03
	Owner	13.02	5.86	0	2.8	8.90	12.91	16.73	22.84	30.03
	Overall	13.80	5.75	0	2.8	10.33	13.98	17.67	22.88	30.03
Ownership_NER	Attorney	2.59	4.08	0	0	1	2	3	7	62
	Owner	1.73	2.55	0	0	0	1	2	6	48
	Overall	2.20	3.50	0	0	0	1	3	7	62
Ownership_word	Attorney	22.31	34.07	0	2	7	14	27	55	561
	Owner	18.24	24.78	0	1	4	10	21	63	303
	Overall	20.48	30.30	0	1	6	12	25	60	561
Infringement_fog	Attorney	14.10	5.81	0	0	11.56	13.87	17.20	23.62	37.77
	Owner	11.95	6.48	0	0	8.20	13.06	16.16	21.75	34.04
	Overall	13.13	6.21	0	0	10.27	13.86	16.82	22.90	37.77
Infringement_NER	Attorney	3.37	5.40	0	0	1	2	4	13	84
	Owner	1.81	3.50	0	0	0	1	2	6	58
	Overall	2.66	4.71	0	0	0	1	3	12	84
Infringement_word	Attorney	29.45	37.70	0	0	8.75	19	34	89	682
	Owner	20.87	32.44	0	0	5	11	23	78	467
	Overall	25.58	35.68	0	0	7	16	30	84	682
Redundancy	Attorney	3.56	6.80	0	0	1	2	4	10	163
	Owner	2.71	5.87	0	0	0	1	3	8	120
	Overall	3.18	6.41	0	0	0	2	4	10	163
Infringing_URL_count	Attorney	6.93	39.40	0	1	1	1	2	20	1418
	Owner	4.25	12.59	0	1	1	1	3	15	261
	Overall	5.72	30.42	0	1	1	1	3	17	1418
Fork	Attorney	4.65	34.89	0	0	0	0	0	10	999
	Owner	2.06	20.61	0	0	0	0	0	4	685
	Overall	3.48	29.35	0	0	0	0	0	7	999

This table provides summary statistics for all non-binary variables. For detailed definitions of these variables, please refer to Table 2. For each variable, we have compiled statistics for three cohorts: notices authored by owners, notices authored by attorneys, and the overall dataset.

Table 4: Summary Statistics of Dummy Variables

Feature	Attribute	Percentage	Attribute	Percentage
Legal_representative	Attorney	54.91%	Owner	45.09%
GitHub_revision	Revised	9.48%	Original	90.52%
Anti_circumvention	Yes	15.56%	No	84.44%
License	Yes	4.44%	No	95.56%
Demand	Removal	83.45%	Others	16.55%
Chance_to_change	Yes	52.33%	No	47.67%
GitHub_verification	Yes	8.16%	No	91.84%

This table provides summary statistics for all dummy variables. For detailed definitions of these variables, please refer to Table 2.

Table 5: Baseline Features

	Logit	Probit	CLogLog
	(1)	(2)	(3)
Legal_representative	1.0968***	0.4262***	-1.0887***
	(0.271)	(0.101)	(0.269)
Infringing_URL_count	-0.002	-0.0008	0.002
	(0.002)	(0.001)	(0.002)
Demand	-0.7876*	-0.3138**	0.7813*
	(0.432)	(0.157)	(0.43)
Intercept	4.4718***	2.2808***	-4.4779***
	(0.431)	(0.156)	(0.429)
N	4684	4684	4684
Pseudo R square	0.03098	0.03130	0.004632

This table presents the baseline features of the takedown notices that influence the settlement rate, analyzed through logistic regression, probit regression, and complementary log-log regression (cloglog). Each column represents a distinct regression model. In logistic and probit regressions, the dependent variable is assigned a value of 1 when the dispute is settled (i.e., the takedown notice did not receive a counter notice). Conversely, in the cloglog regression, the dependent variable is assigned a value of 1 when the takedown notice did receive a counter notice. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 6: Baseline Features After Re-sampling

	Logit	Probit	CLogLog
	(1)	(2)	(3)
Legal_representative	1.9063***	1.1301***	-1.6144***
	(0.064)	(0.035)	(0.057)
Infringing_URL_count	0.0019	0.0014	-0.0013
	(0.001)	(0.001)	(0.001)
Demand	-0.2314***	-0.1666***	0.142**
	(0.07)	(0.042)	(0.056)
Intercept	0.239***	0.1663***	-0.5211***
	(0.066)	(0.04)	(0.053)
N	6925	6925	6925
Pseudo R square	0.1229	0.1234	0.1443

This table presents regressions on the takedown notice's baseline features using the re-sampled dataset. We apply SMOTE to increase the size of the samples that represent the countered takedown notices. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 7: Precautionary Technology

	Logit	Probit	CLogLog
	(1)	(2)	(3)
Legal_representative	1.053***	0.4086***	-1.0449***
	(0.27)	(0.101)	(0.268)
Infringing_URL_count	-0.0021	-0.0008	0.0021
	(0.002)	(0.001)	(0.002)
Demand	-0.8304*	-0.3375**	0.8216*
	(0.429)	(0.158)	(0.428)
Anti_circumvention	0.0145	-0.0045	-0.0165
	(0.346)	(0.137)	(0.342)
License	-0.9644**	-0.4084**	0.9484**
	(0.383)	(0.169)	(0.375)
Intercept	4.5991***	2.3396***	-4.6019***
	(0.426)	(0.157)	(0.425)
N	4684	4684	4684
Pseudo R square	0.03808	0.03846	0.005688

This table shows how the precautionary measures applied by the plaintiff impact the settlement rate. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 8: Precautionary Technology After Re-sampling

	Logit	Probit	CLogLog
	(1)	(2)	(3)
Legal_representative	1.9083***	1.1392***	-1.5814***
	(0.065)	(0.036)	(0.057)
Infringing_URL_count	0.0021	0.0015*	-0.0013
	(0.002)	(0.001)	(0.001)
Demand	-0.3274***	-0.2242***	0.205***
	(0.07)	(0.043)	(0.056)
Anti_circumvention	2.0561***	1.1797***	-1.7562***
	(0.148)	(0.077)	(0.14)
License	1.1957***	0.7248***	-0.9593***
	(0.173)	(0.099)	(0.151)
Intercept	0.1364**	0.1021**	-0.4453***
	(0.066)	(0.041)	(0.053)
N	6925	6925	6925
Pseudo R square	0.1617	0.1629	0.1846

This table shows how the precautionary measures applied by the plaintiff impact the settlement rate using the re-sampled dataset. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 9: Textual Features

	Logit (1)	Probit (2)	CLogLog (3)	Logit (4)	Probit (5)	CLogLog (6)	Logit (7)	Probit (8)	CLogLog (9)
Legal_representative	1.0604***	0.4208***	-1.0511***	1.1574***	0.4399***	-1.164***	1.1248***	0.4343***	-1.1269***
	(0.272)	(0.101)	(0.271)	(0.284)	(0.105)	(0.286)	(0.287)	(0.105)	(0.291)
Infringing_URL_count	-0.002	-0.0008	0.002	-0.0021	-0.0008	0.0021	-0.002	-0.0008	0.002
	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)
Demand	-0.7839*	-0.3042*	0.7786*	-0.8416*	-0.3356**	0.8311*	-0.8667*	-0.3284**	0.8737*
	(0.435)	(0.158)	(0.432)	(0.439)	(0.16)	(0.432)	(0.463)	(0.165)	(0.459)
Ownership_fog	-0.0097	-0.0048	0.0096				-0.0086	-0.0038	0.0088
	(0.023)	(0.009)	(0.023)				(0.023)	(0.009)	(0.023)
Ownership_NER	0.1117*	0.0475*	-0.1087*				0.1132*	0.0523*	-0.1082
	(0.065)	(0.028)	(0.064)				(0.069)	(0.028)	(0.069)
Ownership_word	-0.0134**	-0.0057*	0.0129**				-0.0136*	-0.0064**	0.0128**
	(0.006)	(0.003)	(0.006)				(0.007)	(0.003)	(0.006)
Infringement_fog				0.015	0.0084	-0.0126	0.0106	0.0063	-0.0086
				(0.019)	(0.008)	(0.018)	(0.018)	(0.007)	(0.018)
Infringement_NER				0.0141	0.0073	-0.0119	0.0066	0.0048	-0.0032
				(0.045)	(0.019)	(0.04)	(0.045)	(0.018)	(0.039)
Infringement_word				-0.0093*	-0.0046**	0.0083**	-0.0087*	-0.0046**	0.0075*
				(0.005)	(0.002)	(0.004)	(0.005)	(0.002)	(0.004)
Redundancy							0.0198	0.0095	-0.0184
							(0.015)	(0.006)	(0.015)
Intercept	4.6746***	2.3647***	-4.6782***	4.5473***	2.3054***	-4.5543***	4.7587***	2.3844***	-4.7753***
	(0.548)	(0.202)	(0.544)	(0.483)	(0.18)	(0.477)	(0.594)	(0.216)	(0.588)
N	4684	4684	4684	4684	4684	4684	4684	4684	4684
Pseudo R square	0.03718	0.03804	0.005544	0.05123	0.05317	0.007563	0.05678	0.06011	0.008340

This table presents how the textual features in responses to Questions 3 and 4 impact the settlement rate. The questions are listed in Table 1. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 10: Textual Features After Re-sampling

	Logit (1)	Probit (2)	CLogLog (3)	Logit (4)	Probit (5)	CLogLog (6)	Logit (7)	Probit (8)
Legal_representative	1.8436***	1.0995***	-1.5347***	1.8903***	1.1121***	-1.6393***	1.8211***	1.0711***
	(0.065)	(0.036)	(0.058)	(0.067)	(0.037)	(0.06)	(0.069)	(0.038)
Infringing_URL_count	0.0022	0.0016*	-0.0016	0.0017	0.0012	-0.0017	0.0024	0.0016
	(0.002)	(0.001)	(0.002)	c	(0.001)	(0.002)	(0.002)	(0.001)
Demand	-0.1093	-0.0905**	0.0903	-0.3422***	-0.2269***	0.1872***	-0.1914**	-0.1334***
	(0.073)	(0.044)	(0.058)	(0.071)	(0.043)	(0.057)	(0.076)	(0.045)
Ownership_fog	-0.0332***	-0.0216***	0.0241***				-0.0319***	-0.0208***
	(0.006)	(0.003)	(0.004)				(0.006)	(0.003)
Ownership_NER	0.311***	0.1758***	-0.2168***				0.3762***	0.2097***
	(0.027)	(0.015)	(0.02)				(0.029)	(0.017)
Ownership_word	-0.0254***	-0.0144***	0.0156***				-0.034***	-0.0186***
	(0.002)	(0.001)	(0.002)				(0.003)	(0.002)
Infringement_fog				0.0203***	0.011***	-0.0023	0.0138**	0.0073**
				(0.005)	(0.003)	(0.004)	(0.005)	(0.004)
Infringement_NER				0.0325***	0.0173**	-0.0189**	0.0273**	0.0147**
				(0.012)	(0.007)	(0.008)	(0.012)	(0.007)
Infringement_word				-0.0152***	-0.0083***	0.0068***	-0.0176***	-0.0096***
				(0.002)	(0.001)	(0.001)	(0.002)	(0.001)
Redundancy							0.0697***	0.0384***
							(0.01)	(0.006)
Intercept	0.5782***	0.3903***	-0.774***	0.4589***	0.2929***	-0.689***	0.7929***	0.5063***
	(0.096)	(0.057)	(0.075)	(0.088)	(0.053)	(0.069)	(0.108)	(0.066)
N	6925	6925	6925	6925	6925	6925	6925	6925
Pseudo R square	0.1512	0.1519	0.1711	0.1555	0.1544	0.1740	0.1940	0.1923

This table presents how the textual features in responses to Questions 3 and 4 impact the settlement rate using the re-sampled dataset. The questions are listed in Table 1. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 11: Fork and Interaction of Baseline Variables

	Logit	Probit	Logit	Probit	CLogLog
	(1)	(2)	(3)	(4)	(5)
Legal_representative	18.0508***	5.6426	1.0994***	0.4269***	-1.0914***
	(1.656)	(8.602)	(0.272)	(0.102)	(0.27)
Infringing_URL_count	-0.0036	-0.0013	-0.002	-0.0008	0.002
	(0.004)	(0.002)	(0.002)	(0.001)	(0.002)
Demand	-0.4512	-0.1802	-0.7853*	-0.3129**	0.779*
	(0.441)	(0.174)	(0.432)	(0.157)	(0.43)
Legal_representative$\times$*Demand*	-17.0642***	-5.2612			
	(1.603)	(8.596)			
Fork			-0.0009	-0.0003	0.0009
			(0.002)	(0.001)	(0.001)
Intercept	4.1801***	2.1673***	4.4718***	2.2808***	-4.478***
	(0.413)	(0.162)	(0.431)	(0.156)	(0.429)
N	4684	4684	4684	4684	4684
Pseudo R square	0.03658	0.03635	0.03105	0.03135	0.004643

This table presents the influence of the number of forks and the interaction between legal representation and demand on the settlement rate. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 12: Fork and Interaction of Baseline Variables After Re-sampling

	Logit	Probit	CLogLog	Logit	Probit	CLogLog
	(1)	(2)	(3)	(4)	(5)	(6)
Legal_representative	18.0617***	9.4164	-26.077	1.9053***	1.1295***	-1.614***
	(0.073)	(4890000.0)	(315.325)	(0.064)	(0.036)	(0.057)
Infringing_URL_count	0.0019	0.0014	-0.0013	0.0019	0.0014	-0.0013
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Demand	0.0309	0.0197	-0.0223	-0.2328***	-0.1678***	0.1423**
	(0.081)	(0.051)	(0.059)	(0.07)	(0.042)	(0.056)
Legal_representative$\times$*Demand*	-16.3393***	-8.3914	24.6251			
	(0.099)	(4890000.0)	(315.325)			
Fork				0.0007	0.0006	-0.0002
				(0.001)	(0.001)	(0.001)
Intercept	0.0258	0.015	-0.3862***	0.2386***	0.1659***	-0.5211***
	(0.074)	(0.049)	(0.054)	(0.066)	(0.04)	(0.053)
N	6925	6925	6925	6925	6925	6925
Pseudo R square	0.1342	0.1343	0.1570	0.1229	0.1235	0.1443

This table presents the influence of the number of forks and the interaction between legal representation and demand on the settlement rate using the re-sample dataset. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 13: Interaction of Legal Representative and Textual Features

	Logit (1)	Probit (2)	CLogLog (3)	Logit (4)	Probit (5)	CLogLog (6)	Logit (7)	Probit (8)	CLogLog (9)
Legal_representative	2.3155***	0.9058***	-2.2929***	0.8189	0.3147	-0.8117	1.1033***	0.4271***	-1.0958***
	(0.709)	(0.265)	(0.706)	(0.537)	(0.21)	(0.532)	(0.274)	(0.103)	(0.273)
Infringing_URL_count	-0.0023	-0.0008	0.0023	-0.0019	-0.0008	0.0019	-0.0021	-0.0008	0.002
	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)
Demand	-0.805*	-0.319**	0.7982*	-0.8862*	-0.3407**	0.931**	-0.795*	-0.3171**	0.7885*
	(0.433)	(0.157)	(0.431)	(0.47)	(0.165)	(0.466)	(0.433)	(0.157)	(0.431)
Ownership_fog	0.0021	0.0012	-0.0021						
	(0.03)	(0.012)	(0.029)						
Ownership_NER	0.0876	0.0329	-0.0871						
	(0.119)	(0.047)	(0.118)						
Ownership_word	-0.0065	-0.0024	0.0064						
	(0.01)	(0.004)	(0.01)						
Infringement_fog				0.0195	0.01	-0.0167			
				(0.024)	(0.01)	(0.024)			
Infringement_NER				0.027	0.0181	-0.0147			
				(0.055)	(0.026)	(0.049)			
Infringement_word				-0.0136**	-0.0072***	0.0115**			
				(0.006)	(0.003)	(0.005)			
Redundancy							-0.0103	-0.0052	0.0099
							(0.009)	(0.005)	(0.008)
Legal_representative×*Ownership_fog*	-0.0685	-0.0275	0.0676						
	(0.046)	(0.018)	(0.046)						
Legal_representative×*Ownership_NER*	0.1053	0.0481	-0.0993						
	(0.142)	(0.062)	(0.138)						
Legal_representative×*Ownership_word*	-0.0179	-0.0076	0.0172						
	(0.014)	(0.006)	(0.013)						
Legal_representative×*Infringement_fog*				0.0063	0.0012	-0.0084			
				(0.037)	(0.015)	(0.037)			
Legal_representative×*Infringement_NER*				-0.0397	-0.0229	0.0272			
				(0.078)	(0.034)	(0.073)			
Legal_representative×*Infringement_word*				0.0101	0.0054	-0.0081			
				(0.009)	(0.004)	(0.008)			
Legal_representative×*Redundancy*							0.0003	0.0009	-6.1E-05
							(0.013)	(0.007)	(0.012)
Intercept	4.4358***	2.2607***	-4.4426***	4.6493***	2.3414***	-4.7059***	4.5088***	2.2991***	-4.5136***
	(0.567)	(0.215)	(0.563)	(0.536)	(0.198)	(0.533)	(0.432)	(0.157)	(0.431)
N	4684	4684	4684	4684	4684	4684	4684	4684	4684
Pseudo R square	0.04514	0.04555	0.006735	0.05627	0.05831	0.008341	0.03168	0.03212	0.004735

This table presents how the interaction terms of the legal representative and the textual features in responses to Questions 3 and 4 impact the settlement rate. The questions are listed in Table 1. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 14: Interaction of Legal Representative and Textual Features After Re-sampling

	Logit (1)	Probit (2)	CLogLog (3)	Logit (4)	Probit (5)	CLogLog (6)	Logit (7)	Probit (8)
Legal_representative	2.9567***	1.7536***	-2.5118***	1.4099***	0.8233***	-1.211***	1.8507***	1.1002***
	(0.168)	(0.095)	(0.142)	(0.128)	(0.077)	(0.107)	(0.074)	(0.04)
Infringing_URL_count	0.0025	0.0018*	-0.0018	0.0017	0.0012	-0.0016	0.0019	0.0014
	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)
Demand	-0.1366*	-0.1062**	0.1029*	-0.3458***	-0.2267***	0.1993***	-0.2307***	-0.1659***
	(0.073)	(0.044)	(0.058)	(0.072)	(0.043)	(0.061)	(0.07)	(0.042)
Ownership_fog	-0.0193***	-0.0122***	0.0157***					
	(0.007)	(0.004)	(0.005)					
Ownership_NER	0.2838***	0.1696***	-0.1995***					
	(0.03)	(0.017)	(0.022)					
Ownership_word	-0.022***	-0.0129***	0.0137***					
	(0.003)	(0.002)	(0.002)					
Infringement_fog				0.0199***	0.0105***	-0.0056		
				(0.006)	(0.004)	(0.005)		
Infringement_NER				0.0798***	0.0448***	-0.0366***		
				(0.016)	(0.009)	(0.01)		
Infringement_word				-0.0228***	-0.0128***	0.0104***		
				(0.002)	(0.001)	(0.002)		
Redundancy							-0.0159	-0.0085*
							(0.011)	(0.005)
Legal_representative×*Ownership_fog*	-0.0668***	-0.0392***	0.0564***					
	(0.011)	(0.006)	(0.009)					
Legal_representative×*Ownership_NER*	0.091	0.0264	-0.0853*					
	(0.063)	(0.032)	(0.05)					
Legal_representative×*Ownership_word*	-0.012**	-0.0051*	0.0112***					
	(0.005)	(0.003)	(0.004)					
Legal_representative×*Infringement_fog*				0.0115	0.0083	-0.0198**		
				(0.01)	(0.006)	(0.008)		
Legal_representative×*Infringement_NER*				-0.1208***	-0.068***	0.0676***		
				(0.023)	(0.013)	(0.014)		
Legal_representative×*Infringement_word*				0.0227***	0.0127***	-0.0106***		
				(0.003)	(0.002)	(0.002)		
Legal_representative×*Redundancy*							0.0189	0.0101
							(0.013)	(0.006)
Intercept	0.393***	0.26***	-0.6566***	0.5633***	0.362***	-0.7377***	0.2841***	0.1905***
	(0.105)	(0.064)	(0.078)	(0.098)	(0.06)	(0.076)	(0.073)	(0.043)
N	6925	6925	6925	6925	6925	6925	6925	6925
Pseudo R square	0.1557	0.1566	0.1773	0.1644	0.1638	0.1812	0.1235	0.1240

This table presents how the interaction terms of the legal representative and the textual features in responses to Questions 3 and 4 impact the settlement rate using the re-sampled dataset. The questions are listed in Table 1. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 15: Interaction of Demand and Textual Features

	Logit	Probit	CLogLog	Logit	Probit	CLogLog	Logit	Probit	CLogLog
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Legal_representative	1.0245***	0.4062***	-1.0152***	1.1653***	0.4428***	-1.172***	1.0891***	0.4242***	-1.0808***
	(0.275)	(0.103)	(0.274)	(0.286)	(0.106)	(0.288)	(0.276)	(0.103)	(0.274)
Infringing_URL_count	-0.002	-0.0008	0.002	-0.0026	-0.001	0.0026	-0.002	-0.0008	0.0019
	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)
Demand	0.4732	0.214	-0.4635	-1.4931	-0.6039	1.4729	-0.5929	-0.2369	0.5884
	(0.802)	(0.33)	(0.792)	(1.106)	(0.4)	(1.105)	(0.58)	(0.214)	(0.577)
Demand × *Ownership_fog*	-0.0884	-0.0359	0.0876						
	(0.067)	(0.027)	(0.066)						
Demand × *Ownership_NER*	0.6956**	0.2735**	-0.6859**						
	(0.323)	(0.133)	(0.316)						
Demand × *Ownership_word*	-0.0921***	-0.0366**	0.0907***						
	(0.034)	(0.014)	(0.033)						
Demand × *Infringement_fog*				0.0588	0.0241	-0.0566			
				(0.071)	(0.027)	(0.07)			
Demand × *Infringement_NER*				-0.1423	-0.0525	0.1425			
				(0.287)	(0.079)	(0.293)			
Demand × *Infringement_word*				0.0086	0.003	-0.0091			
				(0.022)	(0.006)	(0.022)			
Demand × *Redundancy*							-0.0591	-0.0227	0.0586
							(0.127)	(0.043)	(0.127)
Intercept	3.5249***	1.8942***	-3.5418***	5.1698***	2.5612***	-5.1671***	4.313***	2.2212***	-4.32***
	(0.727)	(0.302)	(0.716)	(1.076)	(0.387)	(1.076)	(0.565)	(0.208)	(0.562)
N	4684	4684	4684	4684	4684	4684	4684	4684	4684
Pseudo R square	0.04427	0.04562	0.006588	0.05303	0.05515	0.007817	0.03206	0.03254	0.004790

This table presents how the interaction terms of the respondent's demand and the textual features in responses to Questions 3 and 4 impact the settlement rate. The questions are listed in Table 1. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 16: Interaction of Demand and Textual Features After Re-sampling

	Logit	Probit	CLogLog	Logit	Probit	Logit	Probit	CLogLog
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Legal_representative	1.823***	1.0825***	-1.5143***	1.9002***	1.1194***	1.899***	1.1245***	-1.6087***
	(0.066)	(0.037)	(0.058)	(0.068)	(0.037)	(0.065)	(0.036)	(0.058)
Infringing_URL_count	0.0014	0.0011	-0.0009	0.0015	0.0011	0.0018	0.0013	-0.0011
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Demand	1.0058***	0.6316***	-0.6798***	-0.2303	-0.182	-0.1025	-0.0745	0.066
	(0.202)	(0.123)	(0.142)	(0.197)	(0.12)	(0.093)	(0.058)	(0.071)
Demand × *Ownership_fog*	-0.0805***	-0.0508***	0.0561***					
	(0.016)	(0.01)	(0.011)					
Demand × *Ownership_NER*	0.4442***	0.2495***	-0.3013***					
	(0.063)	(0.038)	(0.053)					
Demand × *Ownership_word*	-0.0485***	-0.0288***	0.0309***					
	(0.007)	(0.004)	(0.005)					
Demand × *Infringement_fog*				-0.0312*	-0.0135			
				(0.017)	(0.011)			
Demand × *Infringement_NER*				-0.1103***	-0.0543***			
				(0.033)	(0.021)			
Demand × *Infringement_word*				0.0186***	0.0087**			
				(0.005)	(0.004)			
Demand × *Redundancy*						-0.0393**	-0.0272***	0.0245*
						(0.017)	(0.01)	(0.014)
Intercept	-0.3651**	-0.2245**	-0.1225	0.3603*	0.2534**	0.1439*	0.095*	-0.4653***
	(0.185)	(0.113)	(0.128)	(0.185)	(0.113)	(0.087)	(0.054)	(0.067)
N	6925	6925	6925	6925	6925	6925	6925	6925
Pseudo R square	0.1608	0.1614	0.1793	0.1583	0.1565	0.1237	0.1243	0.1449

This table presents how the interaction terms of the respondent's demand and the textual features in responses to Questions 3 and 4 impact the settlement rate using the re-sampled dataset. The questions are listed in Table 1. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 17: Platform's Actions

	Logit	Probit	CLogLog	Logit	Probit	CLogLog
	(1)	(2)	(3)	(4)	(5)	(6)
Legal_representative	1.1147***	0.4422***	-1.104***	1.1578***	0.4441***	-1.1502***
	(0.269)	(0.102)	(0.267)	(0.299)	(0.111)	(0.297)
Infringing_URL_count	-0.0019	-0.0008	0.0018	-0.0019	-0.0007	0.0019
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)
Demand	-0.34	-0.1538	0.3351	-0.7569*	-0.3078**	0.75*
	(0.497)	(0.177)	(0.498)	(0.434)	(0.157)	(0.433)
GitHub_revision	-0.5707*	-0.2354*	0.5621*	-0.4786	-0.2093	0.4711
	(0.338)	(0.14)	(0.333)	(0.417)	(0.18)	(0.411)
Chance_to_change	0.7712***	0.3072***	-0.7627***			
	(0.297)	(0.113)	(0.296)			
GitHub_verification	-0.1193	-0.0383	0.1209			
	(0.428)	(0.172)	(0.424)			
Legal_representative$\times$*GitHub_revision*				-0.361	-0.106	0.3638
				(0.699)	(0.283)	(0.691)
Intercept	3.8257***	2.0338***	-3.8373***	4.502***	2.2998***	-4.507***
	(0.529)	(0.192)	(0.528)	(0.423)	(0.154)	(0.422)
N	4684	4684	4684	4684	4684	4684
Pseudo R square	0.04657	0.04751	0.006948	0.03533	0.03576	0.005280

This table demonstrates the influence of the platform's actions on the settlement rate. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 18: Platform's Actions After Re-sampling

	Logit	Probit	CLogLog	Logit	Probit	CLogLog
	(1)	(2)	(3)	(4)	(5)	(6)
Legal_representative	1.9635***	1.1762***	-1.5131***	1.9296***	1.1503***	-1.618***
	(0.069)	(0.04)	(0.058)	(0.066)	(0.037)	(0.058)
Infringing_URL_count	-0.0006	-0.0006	0.0005	0.0011	0.0008	-0.0007
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Demand	0.8092***	0.4554***	-0.6896***	-0.2659***	-0.1868***	0.1651***
	(0.09)	(0.051)	(0.063)	(0.07)	(0.042)	(0.056)
GitHub_revision	1.623***	0.9065***	-1.2893***	1.7573***	1.0494***	-1.4618***
	(0.175)	(0.097)	(0.148)	(0.187)	(0.103)	(0.171)
Chance_to_change	2.5523***	1.4948***	-2.047***			
	(0.086)	(0.047)	(0.072)			
GitHub_verification	1.6499***	0.9742***	-1.2522***			
	(0.172)	(0.103)	(0.152)			
Legal_representative$\times$*GitHub_revision*				-0.7313**	-0.5361***	0.4843
				(0.345)	(0.17)	(0.33)
Intercept	-1.5333***	-0.896***	0.788***	0.1884***	0.1339***	-0.4833***
	(0.089)	(0.053)	(0.06)	(0.066)	(0.04)	(0.053)
N	6925	6925	6925	6925	6925	6925
Pseudo R square	0.2905	0.2921	0.3092	0.1386	0.1391	0.1612

This table demonstrates the influence of the platform's actions on the settlement rate using the re-sampled dataset. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 19: Interaction of Textual Features and GitHub's Revision

	Logit (1)	Probit (2)	CLogLog (3)	Logit (4)	Probit (5)	CLogLog (6)	Logit (7)	Probit (8)	CLogLog (9)
Legal_representative	1.0773***	0.4299***	-1.0641***	1.1223***	0.4312***	-1.1297***	1.1194***	0.4375***	-1.1105***
	(0.273)	(0.102)	(0.272)	(0.284)	(0.106)	(0.287)	(0.272)	(0.102)	(0.271)
Infringing_URL_count	-0.002	-0.0007	0.002	-0.002	-0.0006	0.002	-0.0019	-0.0007	0.0019
	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)
Demand	-0.8359*	-0.3169**	0.848*	-0.8086*	-0.3295**	0.7997*	-0.7649*	-0.3097**	0.7579*
	(0.471)	(0.162)	(0.48)	(0.447)	(0.161)	(0.441)	(0.434)	(0.157)	(0.433)
Ownership_fog	-0.0168	-0.0074	0.0166						
	(0.028)	(0.01)	(0.027)						
Ownership_NER	0.0851	0.0309	-0.0848						
	(0.091)	(0.034)	(0.09)						
Ownership_word	-0.0076	-0.0029	0.0075						
	(0.009)	(0.004)	(0.009)						
Infringement_fog				0.0356*	0.0168**	-0.032			
				(0.02)	(0.008)	(0.02)			
Infringement_NER				0.0212	0.0079	-0.0228			
				(0.055)	(0.021)	(0.053)			
Infringement_word				-0.0115*	-0.0054**	0.0105*			
				(0.006)	(0.002)	(0.006)			
Redundancy							-0.0043	-0.0018	0.0042
							(0.01)	(0.004)	(0.009)
GitHub_revision	-0.6062	-0.2864	0.5929	0.8343	0.4065*	-0.7994	-0.483	-0.1873	0.4822
	(0.854)	(0.35)	(0.845)	(0.561)	(0.237)	(0.553)	(0.342)	(0.142)	(0.338)
GitHub_revision×*Ownership_fog*	0.0218	0.0121	-0.0199						
	(0.06)	(0.024)	(0.059)						
GitHub_revision×*Ownership_NER*	0.0396	0.0461	-0.0179						
	(0.123)	(0.059)	(0.111)						
GitHub_revision×*Ownership_word*	-0.0136	-0.0086	0.0108						
	(0.013)	(0.006)	(0.011)						
GitHub_revision×*Infringement_fog*				-0.1128***	-0.052***	0.1072***			
				(0.034)	(0.015)	(0.033)			
GitHub_revision×*Infringement_NER*				0.0849	0.0341	-0.0814			
				(0.17)	(0.067)	(0.166)			
GitHub_revision×*Infringement_word*				0.0009	0.0009	-0.0002			
				(0.012)	(0.005)	(0.012)			
GitHub_revision×*Redundancy*							-0.027	-0.0143	0.0252*
							(0.017)	(0.01)	(0.015)
Intercept	4.8153***	2.4128***	-4.8351***	4.3979***	2.2501***	-4.4088***	4.5331***	2.3093***	-4.538***
	(0.569)	(0.203)	(0.576)	(0.496)	(0.184)	(0.491)	(0.425)	(0.155)	(0.424)
N	4684	4684	4684	4684	4684	4684	4684	4684	4684
Pseudo R square	0.04387	0.04550	0.006494	0.06148	0.06404	0.009027	0.03670	0.03751	0.005477

This table presents how the interaction terms between the platform's actions and the textual features of answers to Questions 3 and 4 impact the settlement rate. The questions are in 1. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 20: Interaction of Textual Features and GitHub's Revision After Re-sampling

	Logit	Probit	Logit	Probit	CLogLog	Logit	Probit
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Legal_representative	1.8485***	1.0995***	1.8738***	1.097***	-1.5853***	1.9168***	1.1331***
	(0.065)	(0.037)	(0.068)	(0.037)	(0.059)	(0.065)	(0.036)
Infringing_URL_count	0.0013	0.0009	0.0008	0.0007	-0.0007	0.0011	0.0008
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Demand	-0.1442**	-0.1106**	-0.3875***	-0.2518***	0.2189***	-0.2689***	-0.1879***
	(0.074)	(0.044)	(0.071)	(0.043)	(0.059)	(0.07)	(0.042)
Ownership_fog	-0.0323***	-0.0208***					
	(0.006)	(0.003)					
Ownership_NER	0.3103***	0.176***					
	(0.027)	(0.015)					
Ownership_word	-0.0251***	-0.0142***					
	(0.003)	(0.001)					
Infringement_fog			0.0231***	0.0129***	-0.0053		
			(0.005)	(0.003)	(0.004)		
Infringement_NER			0.0421***	0.0239***	-0.029***		
			(0.013)	(0.007)	(0.01)		
Infringement_word			-0.0176***	-0.0098***	0.0085***		
			(0.002)	(0.001)	(0.001)		
Redundancy						-0.0061	-0.0035
						(0.005)	(0.003)
GitHub_revision	2.4631***	1.385***	2.699***	1.5689***	-2.2508***	1.8192***	1.0197***
	(0.434)	(0.231)	(0.28)	(0.155)	(0.253)	(0.204)	(0.104)
GitHub_revision×*Ownership_fog*	-0.0275	-0.0185					
	(0.029)	(0.016)					
GitHub_revision×*Ownership_NER*	-0.0435	-0.0206					
	(0.13)	(0.075)					
GitHub_revision×*Ownership_word*	-0.0102	-0.0047					
	(0.012)	(0.007)					
GitHub_revision×*Infringement_fog*			-0.1124***	-0.0664***	0.0792***		
			(0.02)	(0.011)	(0.017)		
GitHub_revision×*Infringement_NER*			0.222	0.0802	-0.2201*		
			(0.138)	(0.067)	(0.118)		
GitHub_revision×*Infringement_word*			0.0043	0.0043	0.0042		
			(0.008)	(0.004)	(0.007)		
GitHub_revision×*Redundancy*						-0.0514	-0.0292**
						(0.032)	(0.013)
Intercept	0.5122***	0.3493***	0.4279***	0.2729***	-0.6581***	0.2123***	0.1505***
	(0.097)	(0.058)	(0.09)	(0.054)	(0.071)	(0.069)	(0.042)
N	6925	6925	6925	6925	6925	6925	6925
Pseudo R square	0.1685	0.1686	0.1756	0.1738	0.1944	0.1395	0.1395

This table presents how the interaction terms between the platform's actions and the textual features of answers to Questions 3 and 4 impact the settlement rate using the re-sampled dataset. The questions are in 1. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 21: Interaction of Legal Representative and the Investigation Efforts

	Logit	Probit	Logit	Probit	CLogLog
	(1)	(2)	(3)	(4)	(5)
Legal_representative	0.8494***	0.3335***	1.074***	0.4164***	-1.0664***
	(0.282)	(0.107)	(0.271)	(0.101)	(0.269)
Infringing_URL_count	-0.0137***	-0.006**	-0.002	-0.0008	0.002
	(0.005)	(0.003)	(0.002)	(0.001)	(0.002)
Demand	-0.8058*	-0.3193**	-0.7851*	-0.312**	0.779*
	(0.431)	(0.159)	(0.432)	(0.157)	(0.43)
Fork			-0.0025	-0.0013	0.0024
			(0.002)	(0.001)	(0.002)
Legal_representative$\times$*Infringing_URL_count*	0.0687**	0.026**			
	(0.035)	(0.013)			
Legal_representative$\times$*Fork*			0.0071	0.0031	-0.0069
			(0.005)	(0.002)	(0.005)
Intercept	4.5569***	2.3134***	4.4771***	2.2832***	-4.4831***
	(0.437)	(0.161)	(0.431)	(0.157)	(0.43)
N	4684	4684	4684	4684	4684
Pseudo R square	0.03860	0.03830	0.03167	0.03210	0.004733

This table presents how the interaction terms between the legal representative and the investigation efforts impact the settlement rate. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 22: Interaction of Legal Representative and the Investigation Efforts After Re-sampling

	Logit	Probit	Logit	Probit
	(1)	(2)	(3)	(4)
Legal_representative	1.5603***	0.9442***	1.8697***	1.1105***
	(0.075)	(0.042)	(0.065)	(0.036)
Infringing_URL_count	-0.0026	-0.0017	0.0019	0.0014
	(0.002)	(0.001)	(0.001)	(0.001)
Demand	-0.2273***	-0.1625***	-0.2294***	-0.1659***
	(0.07)	(0.043)	(0.07)	(0.042)
Fork			-0.0032	-0.0015
			(0.005)	(0.002)
Legal_representative$\times Infringing_URL_count$	0.1529***	0.0797***		
	(0.022)	(0.011)		
Legal_representative$\times Fork$			0.0165***	0.0086***
			(0.006)	(0.002)
Intercept	0.2558***	0.1767***	0.2449***	0.1696***
	(0.066)	(0.041)	(0.066)	(0.04)
N	6925	6925	6925	6925
Pseudo R square	0.1302	0.1308	0.1240	0.1246

This table presents how the interaction terms between the legal representative and the investigation efforts impact the settlement rate after re-sampling. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 23: Interaction of Precaution and Legal Representative/Platform's Actions

	Logit (1)	Probit (2)	CLogLog (3)	Logit (4)	Probit (5)	CLogLog (6)	Logit (7)	Probit (8)	CLogLog (9)	Logit (10)	Probit (11)	CLogLog (12)
Legal_representative	1.2281***	0.4761***	-1.2193***	1.0184***	0.3926***	-1.0115***	1.0947***	0.4299***	-1.0859***	1.0504***	0.4093***	-1.0421***
	(0.304)	(0.113)	(0.303)	(0.283)	(0.106)	(0.282)	(0.269)	(0.101)	(0.268)	(0.271)	(0.102)	(0.269)
Infringing_URL_count	-0.002	-0.0008	0.002	-0.0021	-0.0008	0.0021	-0.0019	-0.0006	0.0019	-0.002	-0.0007	0.002
	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)
Demand	-0.7929*	-0.3179**	0.7863*	-0.8367*	-0.3431**	0.8269*	-0.7646*	-0.3115**	0.7573*	-0.8118*	-0.3398**	0.7993*
	(0.432)	(0.158)	(0.431)	(0.428)	(0.157)	(0.427)	(0.434)	(0.158)	(0.433)	(0.419)	(0.153)	(0.421)
Anti_circumvention	0.3633	0.1506	-0.3592				-0.0452	-0.0303	0.043			
	(0.476)	(0.19)	(0.472)				(0.367)	(0.144)	(0.363)			
License				-1.0464**	-0.4617**	1.0248**				-0.9003**	-0.3654*	0.89**
				(0.42)	(0.193)	(0.408)				(0.442)	(0.191)	(0.435)
GitHub_revision							-0.6988**	-0.2991**	0.6875**	-0.5365	-0.217	0.5312
							(0.356)	(0.149)	(0.351)	(0.365)	(0.15)	(0.36)
Legal_representative × *Anti_circumvention*	-0.7622	-0.2969	0.7564									
	(0.705)	(0.271)	(0.701)									
Legal_representative × *License*				0.4778	0.2505	-0.459						
				(1.116)	(0.442)	(1.105)						
GitHub_revision × *Anti_circumvention*							0.7824	0.361	-0.7664			
							(1.116)	(0.435)	(1.107)			
GitHub_revision × *License*										-0.1323	-0.1378	0.096
										(0.923)	(0.424)	(0.888)
Intercept	4.4311***	2.2654***	-4.4375***	4.6183***	2.3498***	-4.62***	4.5344***	2.3126***	-4.5389***	4.6501***	2.3657***	-4.6495***
	(0.432)	(0.158)	(0.431)	(0.427)	(0.158)	(0.426)	(0.426)	(0.155)	(0.425)	(0.422)	(0.155)	(0.423)
N	4684	4684	4684	4684	4684	4684	4684	4684	4684	4684	4684	4684
Pseudo R square	0.03267	0.03301	0.004884	0.03836	0.03895	0.005728	0.03582	0.03662	0.005349	0.04157	0.04228	0.006202

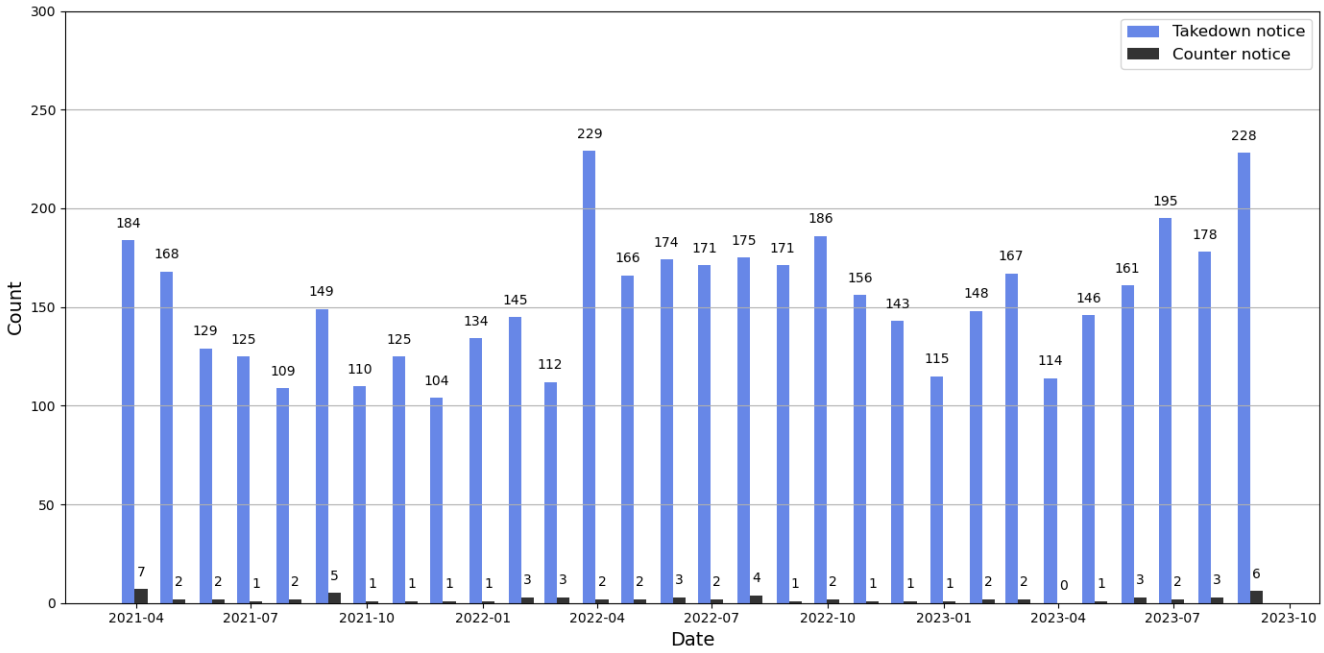
This table reports the effect of the interaction terms between the precautionary measures and the legal representative or GitHub's actions on the settlement rate. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Table 24: Interaction of Precaution and Legal Representative/Platform's Actions After Re-sampling

	Logit (1)	Probit (2)	CLogLog (3)	Logit (4)	Probit (5)	CLogLog (6)	Logit (7)	Probit (8)	CLogLog (9)	Logit (10)	Probit (11)	CLogLog (12)
Legal_representative	1.8639***	1.1203***	-1.5441***	1.9196***	1.1406***	-1.6189***	1.8845***	1.1205***	-1.5576***	1.9266***	1.1406***	-1.6166***
	(0.065)	(0.037)	(0.058)	(0.064)	(0.036)	(0.058)	(0.065)	(0.037)	(0.057)	(0.065)	(0.036)	(0.057)
Infringing_URL_count	0.002	0.0014	-0.0012	0.002	0.0015*	-0.0014	0.0011	0.0008	-0.0006	0.0012	0.0009	-0.0007
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Demand	-0.3453***	-0.2369***	0.221***	-0.2117***	-0.1539***	0.1257**	-0.3796***	-0.2564***	0.2422***	-0.2479***	-0.1741***	0.1507***
	(0.07)	(0.042)	(0.056)	(0.07)	(0.042)	(0.056)	(0.07)	(0.043)	(0.056)	(0.07)	(0.042)	(0.056)
Anti_circumvention	1.9086***	1.1461***	-1.5789***				2.0224***	1.165***	-1.7166***			
	(0.161)	(0.088)	(0.148)				(0.15)	(0.079)	(0.141)			
License				1.0541***	0.6437***	-0.8478***				1.0772***	0.6539***	-0.865***
				(0.18)	(0.106)	(0.155)				(0.178)	(0.103)	(0.155)
GitHub_revision							1.5658***	0.8787***	-1.3116***	1.5628***	0.8681***	-1.3277***
							(0.167)	(0.093)	(0.149)	(0.167)	(0.092)	(0.15)
Legal_representative × *Anti_circumvention*	0.8121*	0.0912	-1.0595**									
	(0.481)	(0.193)	(0.474)									
Legal_representative × *License*				1.3043	0.4269	-1.4467						
				(1.024)	(0.404)	(1.013)						
GitHub_revision × *Anti_circumvention*							0.2296	-0.0279	-0.4344			
							(1.031)	(0.433)	(1.018)			
GitHub_revision × *License*										-0.2133	-0.156	0.0761
										(0.783)	(0.394)	(0.741)
Intercept	0.2045***	0.1435***	-0.4963***	0.1816***	0.1301***	-0.4776***	0.1558**	0.1173***	-0.4578***	0.1399**	0.1058***	-0.4455***
	(0.066)	(0.04)	(0.053)	(0.066)	(0.04)	(0.053)	(0.066)	(0.04)	(0.053)	(0.066)	(0.04)	(0.053)
N	6925	6925	6925	6925	6925	6925	6925	6925	6925	6925	6925	6925
Pseudo R square	0.1558	0.1565	0.1790	0.1288	0.1293	0.1507	0.1698	0.1703	0.1934	0.1430	0.1430	0.1660

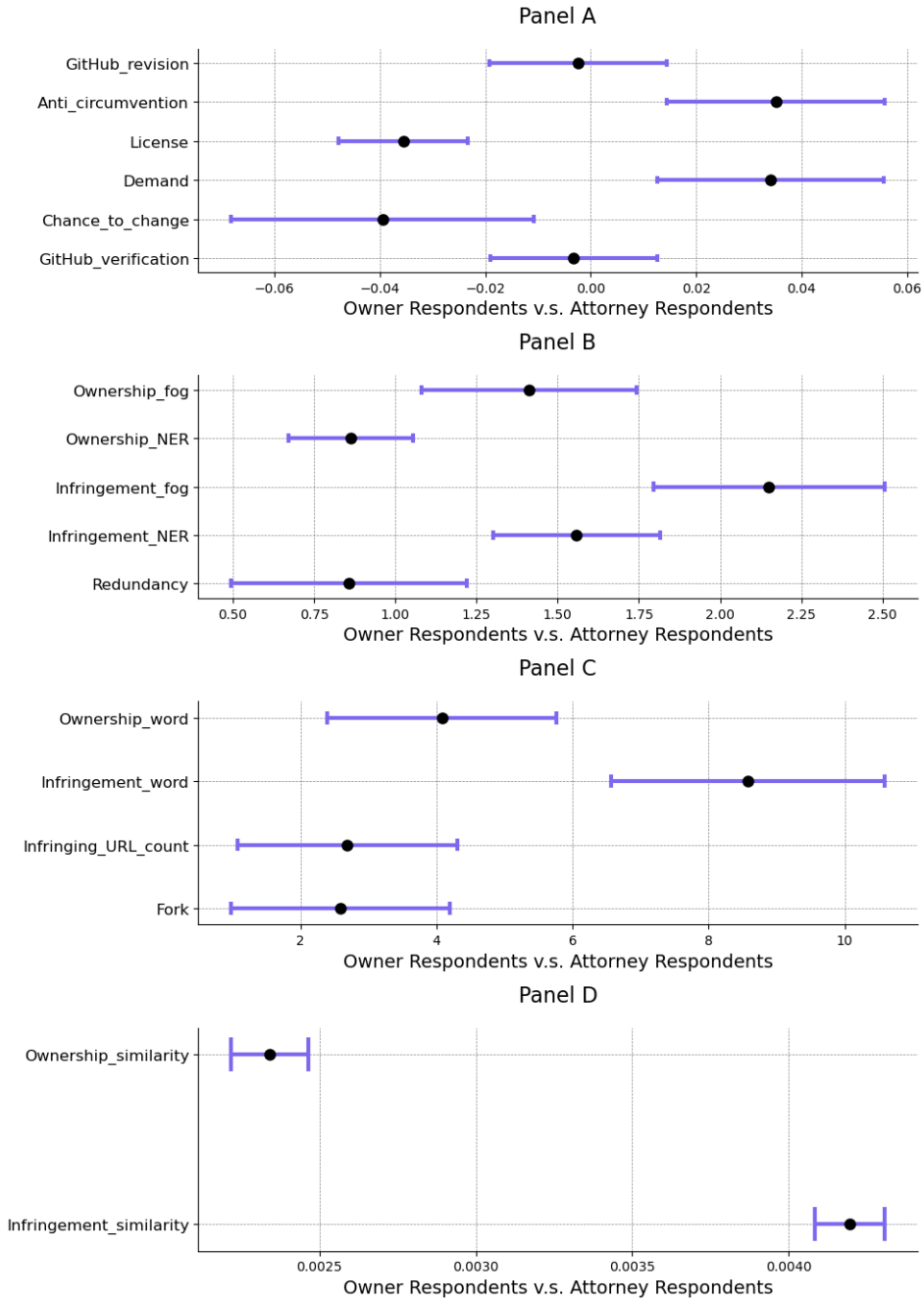
This table reports the effect of the interaction terms between the precautionary measures and the legal representative or GitHub's actions on the settlement rate after re-sampling. Detailed definitions for all variables can be found in Table 2. Robust standard errors are indicated in parentheses. *, **, and *** denote significance levels at 1%, 5%, and 10% confidence levels, respectively.

Figure 1: Number of the Takedown Notices and Counter Notices



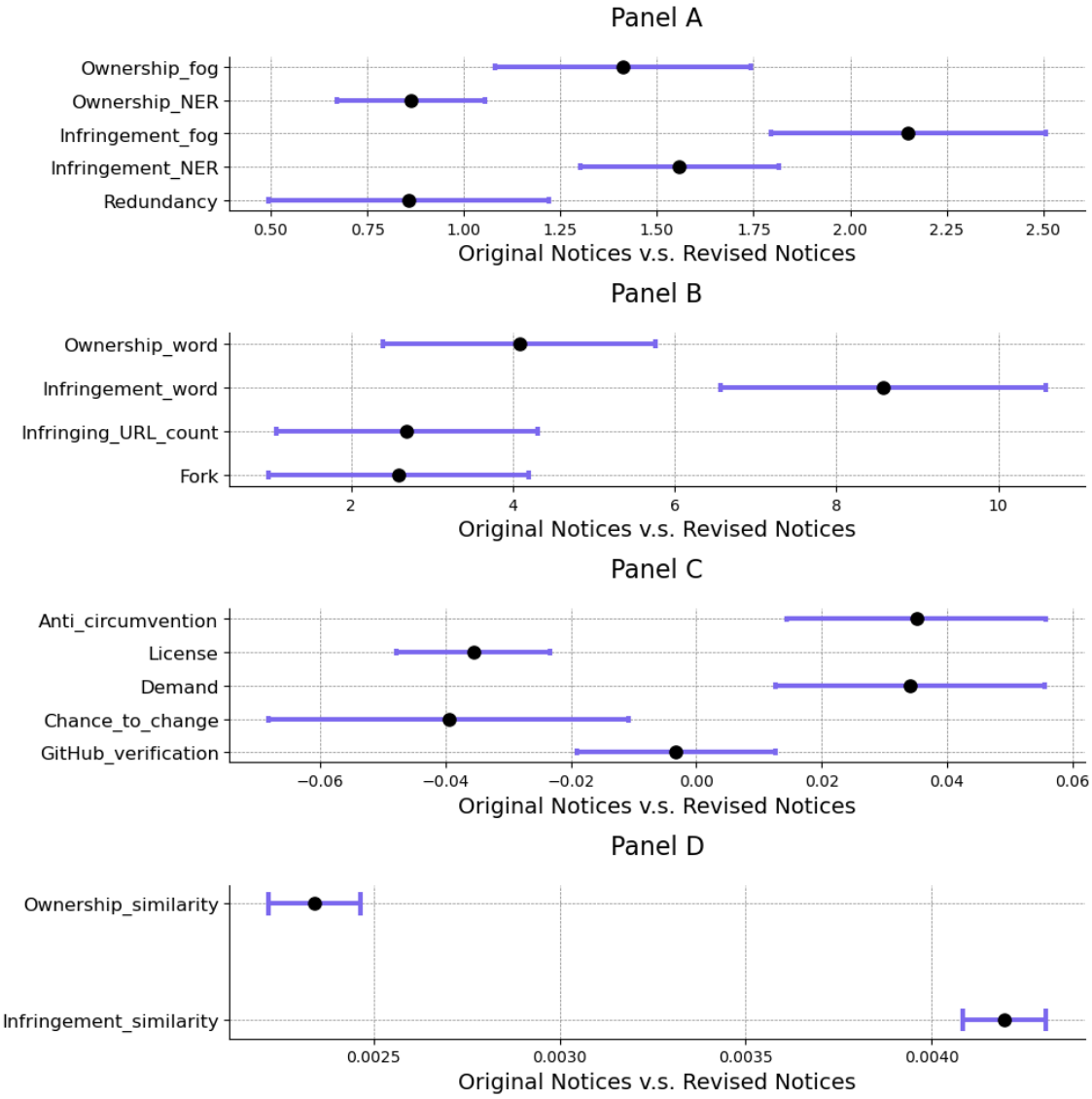
This figure demonstrates the monthly count of takedown notices and counter notices from March 2021 to August 2023.

Figure 2: Comparing Attorney-Written and Owner-Written Notices: Welch's t-Tests



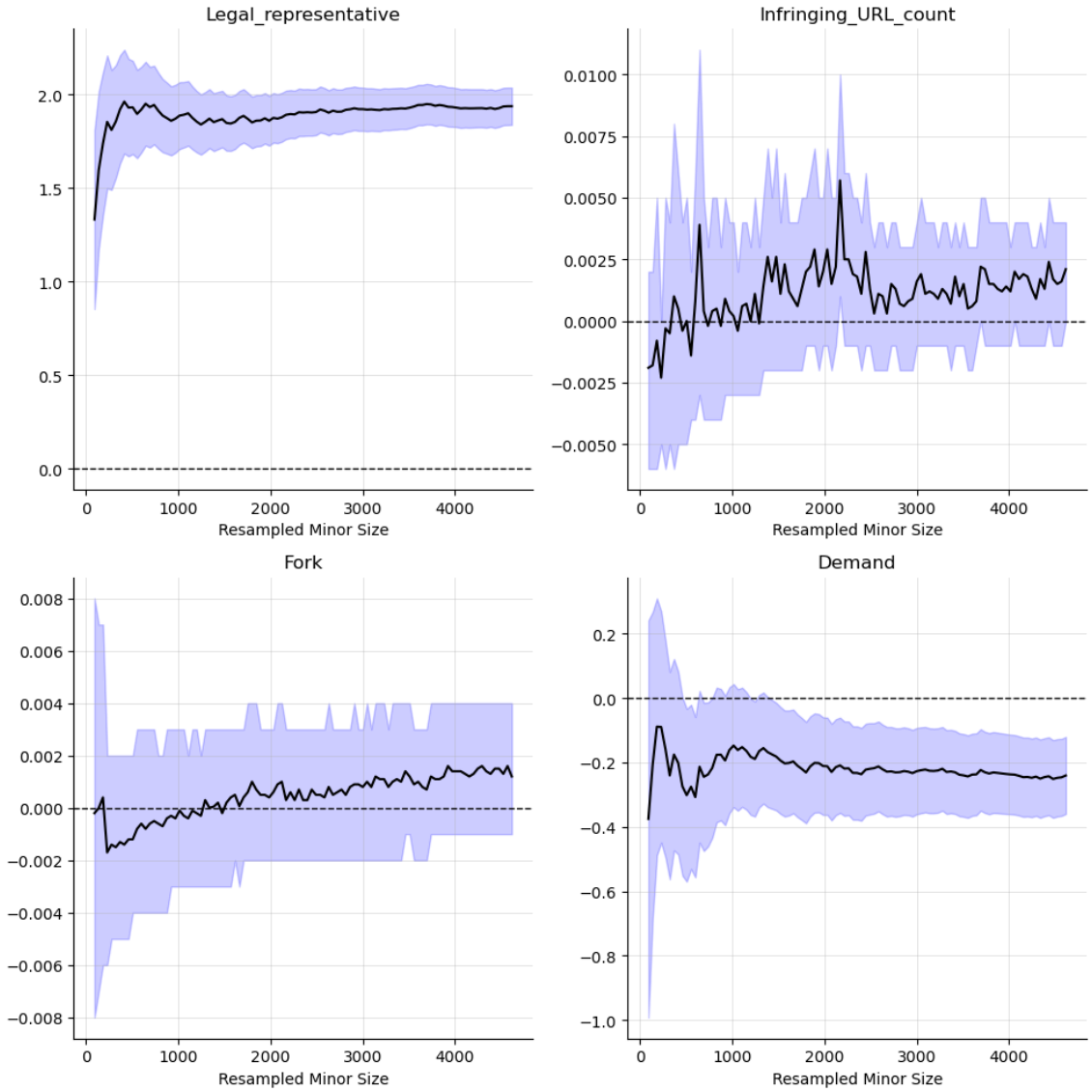
This figure depicts the results of Welch's t-test comparing the features between owner-written takedown notices and attorney-written takedown notices. The black dot represents the difference value, while the blue interval denotes the 95% confidence interval of the difference. The difference is significant when the interval encompasses 0.

Figure 3: Comparing Revised and Original Notices: Welch's t-Tests



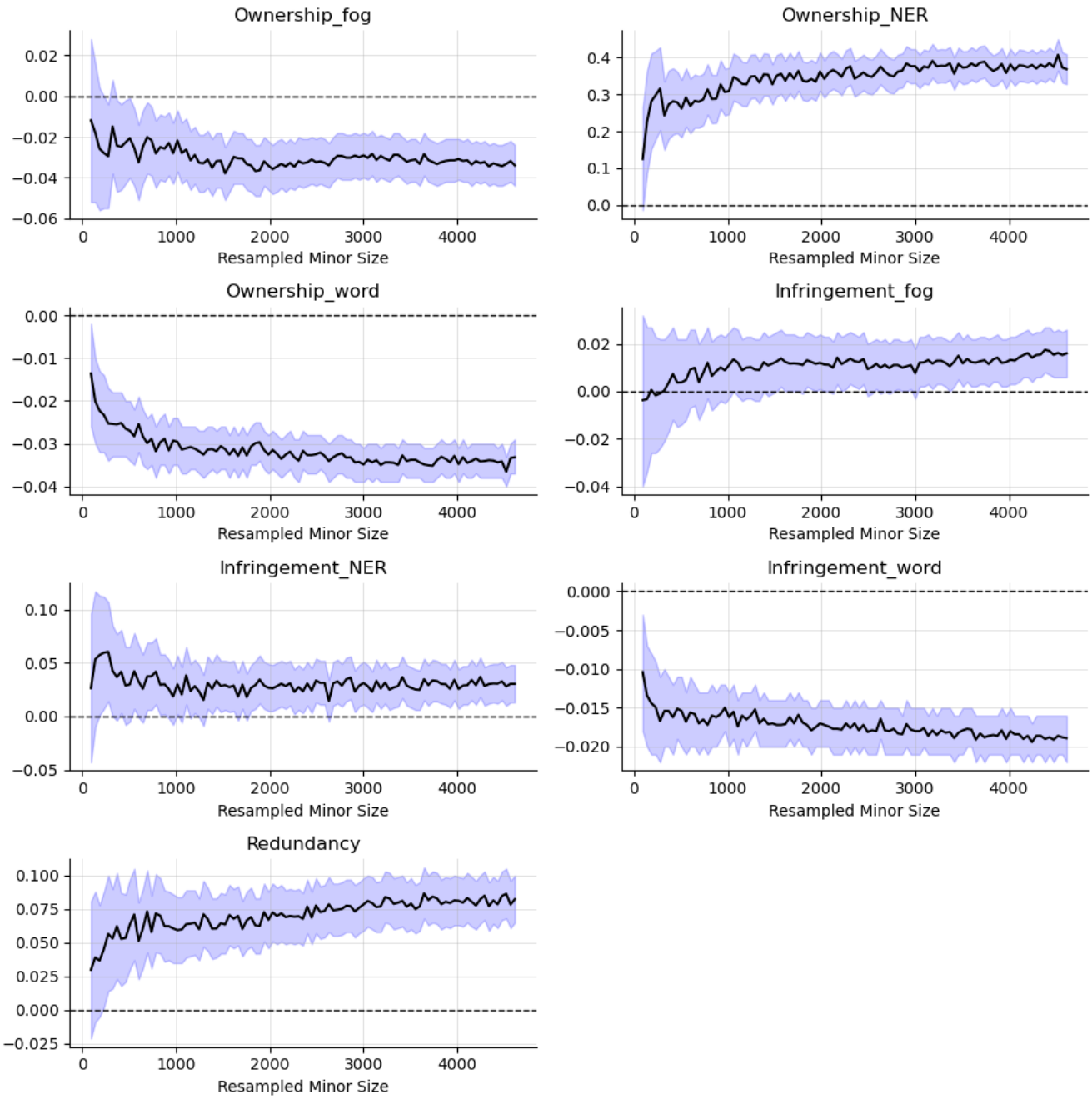
This figure depicts the results of Welch's t-test comparing the features between the original takedown notices and the original takedown notices. The black dot represents the difference value, while the blue interval denotes the 95% confidence interval of the difference. The difference is significant when the interval encompasses 0.

Figure 4: Coefficients from Different Minority Sample Sizes - Baseline Variables



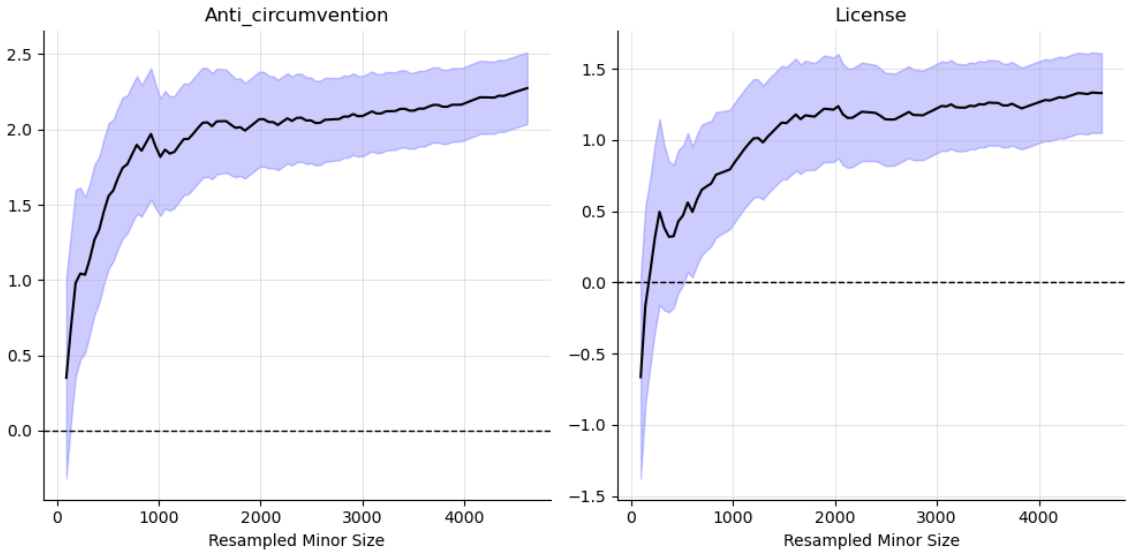
This figure illustrates the coefficients on the baseline variables along with their corresponding 95% confidence intervals, as the size of re-sampling increases. We have employed SMOTE to augment the sample size of takedown notices that did not receive a counter notice, which is represented on the x-axis. The coefficient is significant when the confidence interval encompasses 0.

Figure 5: Coefficients from Different Minority Sample Sizes - Textual Features



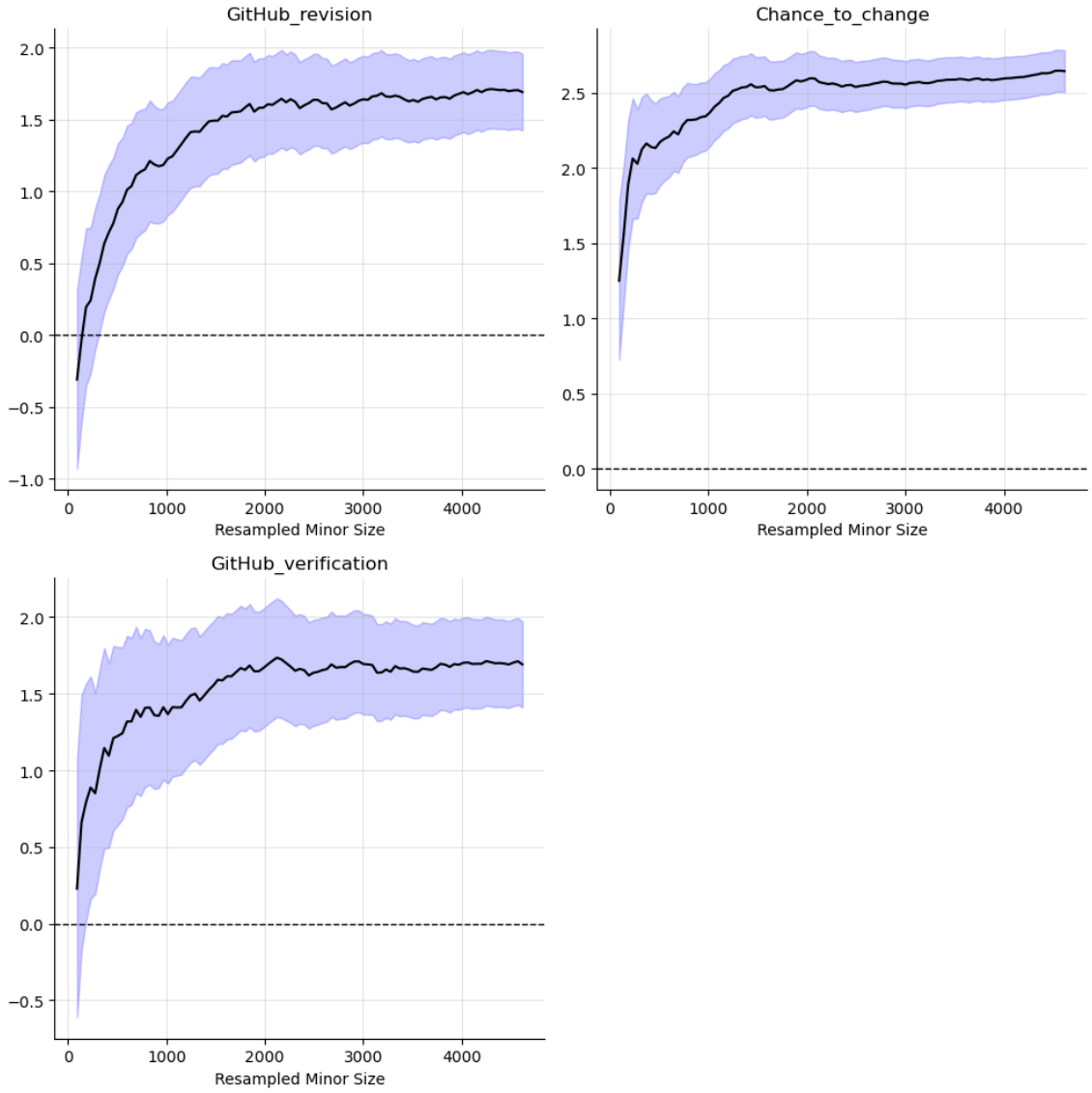
This figure illustrates the coefficients on the textual features along with their corresponding 95% confidence intervals, as the size of re-sampling increases. We have employed SMOTE to augment the sample size of takedown notices that did not receive a counter notice, which is represented on the x-axis. The coefficient is significant when the confidence interval encompasses 0.

Figure 6: Coefficients from Different Minority Sample Sizes - Precaution



This figure illustrates the coefficients on the precautionary measures along with their corresponding 95% confidence intervals, as the size of re-sampling increases. We have employed SMOTE to augment the sample size of takedown notices that did not receive a counter notice, which is represented on the x-axis. The coefficient is significant when the confidence interval encompasses 0.

Figure 7: Coefficients from Different Minority Sample Sizes - Platform's Mediation



This figure illustrates the coefficients on the platform's mediation along with their corresponding 95% confidence intervals, as the size of re-sampling increases. We have employed SMOTE to augment the sample size of takedown notices that did not receive a counter notice, which is represented on the x-axis. The coefficient is significant when the confidence interval encompasses 0.