Deleting Information:
Evidence from the Digital Service Act

ABSTRACT

We study how the European Union's Digital Services Act (DSA), which requires platforms to act expeditiously on notices flagging content as allegedly illegal, affects the deletion of employee reviews on Glassdoor. Exploiting variation in firms' EU exposure in a difference-in-differences design, we find that deletions increase significantly post-DSA for EU-exposed firms relative to non-exposed firms. Using large language model classification to evaluate each review against Glassdoor's content guidelines, we find that the majority of deleted reviews contain no policy violation identifiable from the review text, and the flagged-review increase is concentrated in high-discretion categories where enforcement is subjective. The deletion increase is concentrated in negative and critical reviews, with no statistically significant change for positive reviews, and is larger for firms with stronger strategic incentives. Displayed average ratings increase significantly post-DSA and become stronger predictors of both worker inflows and outflows, consistent with selectively curated ratings attracting workers who subsequently discover conditions that diverge from displayed scores.

Keywords: Content moderation, employee reviews, strategic deletion, platform regulation, labor market information
JEL Classification: K22, M41, J20, L86, G38

1. Introduction

Online platforms such as Glassdoor aggregate employee disclosures that are otherwise difficult to observe, providing job seekers, investors, and regulators with information about firm-level working conditions. The value of this information depends on whether the content available on these platforms accurately reflects the underlying population of employee experiences. In this paper, we study how the European Union's Digital Services Act (DSA) affects the deletion of employee reviews on Glassdoor. The DSA, fully enforced for all hosting services from February 17, 2024, requires online platforms to implement a notice-and-action mechanism through which any party can flag content as allegedly illegal, and conditions platforms' liability exemption on acting expeditiously once a qualifying notice is received. We use this regulatory change to examine whether and how the nature of review deletion changes under a regime that mandates a formal channel for content removal requests.

We focus on Glassdoor because employer reviews on the platform have been shown to convey economically meaningful information. Glassdoor ratings predict stock returns (Green et al. 2019), employee satisfaction is linked to long-run firm value (Edmans 2012), and job seekers rely on employer reviews in their search decisions (Marinescu et al. 2021). Given the economic significance of these reviews, the integrity of review content on Glassdoor has come under increasing scrutiny. Employers have pursued legal action to unmask anonymous reviewers (*Glassdoor, Inc. v. Superior Court* 2017; *Glassdoor, Inc. v. Andra Group, LP* 2019; *Zuru, Inc. v. Glassdoor, Inc.* 2022), and a growing reputation management industry advertises services to help firms remove unfavorable reviews through strategic flagging or third-party intermediaries. While prior studies have examined manipulation through legal pressure on platforms, the addition of fabricated positive reviews (Luca and Zervas, 2016; Mayzlin et al. 2014), and coordinated rating

campaigns (He et al. 2022), the complementary channel, selective deletion of authentic negative reviews, has received less attention. The distinction matters economically. Unlike fake positives, which add noise to the information environment, deletion of genuine negatives removes precisely the information most relevant to prospective employees and investors.

To help ensure the integrity of the information on the platform, Glassdoor publishes community guidelines that prohibit profanity, threats, spam, discriminatory language, confidential information, and reviews that appear to be written by someone other than a current or former employee. Removal of reviews that violate these guidelines serves as quality control that benefits all users by maintaining the credibility of the review ecosystem. But the same moderation infrastructure that enables quality control also creates a channel through which firms can pursue a different objective: suppressing reviews that damage their reputation, regardless of whether those reviews violate any content policy. Under this framework, platform-initiated and firm-initiated removal operated through largely distinct channels, as the platform retained editorial control over the moderation queue and firms had limited formal mechanisms to compel removal.

The DSA changes this dynamic by introducing a mandatory notice-and-action mechanism (Article 16) through which any party can formally notify the platform of content they consider illegal. Because the platform's liability exemption is conditioned on acting expeditiously once a notice is received (Article 6), the expected cost of retaining flagged content can exceed the cost of removing it. This regulatory architecture alters the balance between platform-initiated and firm-initiated moderation, as formal removal requests now carry processing obligations and potential liability that constrain the platform's editorial discretion.

We construct a balanced panel of 96,395 firms observed monthly in a twelve-month window on each side of the DSA enforcement date, using data from Revelio Labs through the

Wharton Research Data Services (WRDS) platform. Because Revelio Labs collects snapshots of Glassdoor at regular intervals, we can identify deleted reviews by comparing consecutive snapshots, yielding an extensive record of deletion activity, although reviews posted and removed between consecutive snapshots are not observed (Section 3.1). A central challenge is determining whether deleted reviews actually violate Glassdoor's content policies or whether they are substantively legitimate reviews that happen to be unfavorable.

Because review text is unstructured, spans multiple languages, and requires contextual judgment to evaluate, we deploy a large language model (LLM), specifically GPT-5.4 Nano, to classify each review against eleven categories of content policy violations drawn from Glassdoor's published guidelines. The multilingual capability of the model is particularly relevant in our setting, as Glassdoor operates globally and our sample includes reviews written in dozens of languages across countries in Europe, the Americas, and Asia. We complement the LLM classification with RoBERTa-based sentiment scores to capture the overall tone of each review. We validate the classifier against five independent human graders on a random sample of 500 reviews and find an overall agreement rate of 83.6%.

Our empirical approach exploits the DSA's enforcement as a natural experiment. We construct a balanced firm-by-month panel and estimate a difference-in-differences model comparing firms with EU exposure (headquartered in the EU or with EU-based reviewers pre-DSA) to firms without. We first estimate the effect on deletion volume. We then decompose the deletions by violation status and review characteristics. Finally, we test whether these deletions reduce the informativeness of visible ratings.

Descriptively, deleted reviews are systematically lower-rated and more likely to be negative. Monthly deletion volume shows a sharp spike immediately following the DSA

enforcement date, with a sustained elevation in deletion activity throughout the post-DSA period. Decomposing deletions by content, we show that the post-DSA surge is driven overwhelmingly by reviews with no violation identifiable from the review text.

Our LLM classification reveals that deleted reviews are nearly three times as likely as surviving reviews to contain at least one content policy violation (lift of 2.93). At the same time, 86.17% of deleted reviews have no violation identifiable from the review text. Among reviews deleted before the DSA, violations are often readily identifiable: a one-star review titled "They will slit your throat" contains explicit violent language; a five-star review titled "Application as Deck Cadet" is a job application rather than an employee review, representing obvious spam. These pre-DSA deletions are largely consistent with platform-initiated quality control.

By contrast, the post-DSA deletion increase is concentrated in reviews with no identifiable policy violation, and the deleted reviews of EU-exposed firms have lower violation rates than those of non-EU-exposed firms. A two-star review states: “Morale has remained low since the layoffs. Chaotic restructuring and poor communication.” Another describes management dysfunction: “Everything is mostly show, there is no equality, no inclusion, managers are just looking to advance on the backs of those below them.” Neither review contains profanity, named individuals, proprietary information, or any other observable content that would warrant removal under Glassdoor's published guidelines. At the extreme, some firms delete virtually all their reviews, boosting their visible score by several stars. The Online Appendix provides additional examples.

Having documented what gets deleted on average, we turn to how the DSA changes the nature of deletion. Our difference-in-differences estimates show that deletions increase significantly post-DSA for EU-exposed firms relative to non-exposed firms. To assess the composition of this increase, we decompose the treatment effect along three dimensions: review

tone, review content balance, and the severity of criticism. Across all three components, the point estimates suggest that reviews that cast the firm in a negative light are more likely to be deleted. Harsh reviews (those with the lowest possible score on at least one sub-rating dimension) show a significant increase, while the corresponding estimate for universally perfect reviews is statistically insignificant. Negatively toned reviews drive the increase, while the estimate for positively toned reviews is statistically insignificant. Reviews with proportionately larger "cons" sections account for the increase along this dimension, while reviews with proportionately larger "pros" sections show no change.

We complement these decomposition tests with cross-sectional analyses that examine which firm characteristics predict a larger deletion response. If the DSA-induced increase reflects strategic behavior, it should be concentrated among firms with greater exposure to content moderation on the platform, stronger competitive pressure, deteriorating ratings, and more negative review profiles. The point estimates confirm this: the deletion increase is significantly larger among firms with higher pre-DSA deletion activity, firms facing more competitive local markets, firms whose ratings were declining, and firms with more negative pre-DSA review tone. While none of these cross-sectional patterns is individually conclusive, and each moderator may also proxy for the supply of removable content or baseline exposure to content moderation, taken together they are consistent with the view that the firms most likely to benefit from reputation management are those driving the post-DSA increase. Review-level determinant tests point in the same direction: after the DSA, clear policy violations became less predictive of deletion for treated firms, while negative tone and subjective content judgments became more predictive, suggesting a shift from content-based enforcement toward reputation-based removal.

An important question is whether these findings reflect broader trends in content moderation rather than a causal effect of the DSA. The estimates are stable across progressively demanding specifications, shorter estimation windows, and geographic subsamples. Event-study estimates provide further support: treated and control firms follow similar deletion trajectories prior to the DSA but diverge at the enforcement date. The absence of differential pre-trends, combined with a placebo permutation test that randomly reassigns treatment status across firms, suggests that the patterns we document are unlikely to reflect pre-existing differences between the two groups.

Finally, we assess consequences for labor market information. If strategic deletion selectively removes negative reviews, the visible rating becomes an upward-biased signal of workplace quality. We find that post-DSA, visible Glassdoor ratings become significantly stronger predictors of both worker inflows and outflows for EU-exposed firms relative to non-exposed firms. This pattern is consistent with selectively curated ratings attracting workers who respond to the posted signal and who, upon discovering conditions that diverge from the curated rating, are more likely to leave. We interpret this evidence as suggestive: it reflects changes in the association between visible ratings and subsequent labor flows rather than direct evidence that deletion-induced rating changes caused those flows.

First, we contribute to work on the economics of content moderation and platform regulation (Jiménez Durán et al. 2024) by providing evidence that notice-and-action mechanisms can serve as a channel for strategic content removal by economic actors with reputational incentives, an unintended consequence of regulation designed to address illegal content. As content moderation regulation expands across jurisdictions, our findings suggest that the design of

notice-and-action systems has direct implications for whether these frameworks protect or degrade the quality of online information.

Second, we contribute to the literature on corporate disclosure management (Healy and Palepu 2001; Leuz and Wysocki 2016) by documenting that firms can manage not only their own disclosures but also information produced by third parties on independent platforms. Prior work has examined factors that affect whether employees post reviews in the first place, including nondisclosure agreements or anti-SLAPP protections (Holstead et al., 2025, Böke et al. 2024), and whether employees use review platforms strategically during labor disputes (Brau and Jeong 2023). Our paper focuses on a distinct margin: whether reviews that have already been posted are subsequently removed. The two margins have different implications for information quality. The posting margin determines the volume and composition of new information entering the platform; the deletion margin determines which existing information is selectively withdrawn.

Third, we contribute to the literature on employer review platforms and labor market information (Sockin and Sojourner 2023; Holstead et al. 2025) by showing that strategic deletion degrades the informativeness of visible ratings, creating conditions under which posted scores attract workers who subsequently discover that workplace quality diverges from what the ratings implied.

2. Background and Hypothesis Development

2.1 The Digital Services Act and Content Moderation

The Digital Services Act (DSA) is the European Union's comprehensive framework governing how online platforms moderate user-generated content (Regulation (EU) 2022/2065 2022). The DSA was adopted in October 2022 and became fully applicable to all hosting services

on February 17, 2024 (Eur. Comm. 2022).[1] The DSA incorporates the existing liability exemption rules of the previous regulation, the e-Commerce Directive (Directive 2000/31/EC 2000), which establishes a safe harbor regime shielding platforms from liability for user content as long as they acted as passive hosts. Over two decades, the e-Commerce Directive's notice-and-takedown approach came under increasing criticism for creating inconsistent enforcement across member states, providing insufficient transparency into platform moderation decisions, and failing to address the scale of content moderation challenges posed by modern platforms (Eur. Comm. 2024).

The DSA was designed to clarify when online platforms obtain "actual knowledge" of illegal content and are therefore required to act to remove the content. Before the DSA, firms seeking to remove a negative review, such as from Glassdoor, could flag content for review, but whether that constituted "actual knowledge" of illegal content for the platform was not precisely defined. After the DSA, content flagged as illegal is processed through a standardized notice-and-action mechanism, which clarifies when platforms are deemed to have "actual knowledge" and introduces enforceable penalties for non-compliance.

An unintended effect of the e-Commerce Directive was that platforms were reluctant to engage in self-initiated moderation. Pre-DSA, proactive monitoring could be interpreted as creating "knowledge" of illegal content, thereby jeopardizing a platform's liability exemption (Kuczerawy 2021). The DSA addresses this by providing explicit liability protections for voluntary content moderation (Article 7), removing the disincentive for platforms to police their own content.

The DSA relies on member state and union laws to determine what constitutes illegal activity.[2] The preamble of the DSA clarifies that the concept of illegal content mirrors “the existing

[1] To see the timeline of implementation refer to Figure OA1: Timeline of Digital Services Act

[2] Article 3(h) of the DSA defines ‘illegal content’ as “any information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law.”

rules in the offline environment," thereby incorporating pre-existing legal standards that may not have been designed specifically for online interactions.[3]

The DSA imposes obligations on all "hosting services," defined broadly to include any platform that stores information provided by any European Union users at their request. This definition encompasses social media platforms, marketplaces, cloud services, and review platforms alike, regardless of country of incorporation. Glassdoor, as a platform that hosts employee-generated reviews of employers, falls within this definition.

Six provisions working in combination are central to our setting[4]. Article 6 codifies a conditional liability exemption: platforms are shielded from liability for user content only as long as they lack knowledge of illegal content or act expeditiously upon gaining such knowledge. Article 7 clarifies that voluntary content moderation does not forfeit platforms' liability exemption. Article 14 requires platforms to disclose the policies and procedures governing content moderation, including the categories under which content may be removed. Article 16 creates a mandatory notice-and-action mechanism through which any individual or entity can formally notify a platform of content they consider illegal; under Article 16(3), a qualifying notice is deemed

[3] This reliance on existing legal framework is further reflected in Article 34 of the DSA, which requires certain platforms to assess systemic risks arising from their services, including: (i) the dissemination of illegal content and (ii) any actual or foreseeable negative effects on fundamental rights under the Charter of Fundamental Rights of the European Union, such as human dignity (Art. 1), privacy (Art. 7), data protection (Art. 8), freedom of expression (Art. 11), non-discrimination (Art. 21), children's rights (Art. 24), and consumer protection (Art. 38). This risk framework is reflected in the content guidelines of online platforms, such as Glassdoor, as platforms seek to reduce these risks and remain in compliance with union law.

[4] In addition to the articles and new obligations outlined here, the DSA also introduces broader transparency, accountability, and risk-mitigation requirements. Platforms must publish detailed information on their content moderation activities, including through public transparency reporting and centralized databases of statements of reasons. Users are granted enhanced rights to challenge moderation decisions, including access to out-of-court dispute resolution mechanisms. The regulation further imposes obligations on Very Large Online Platforms (VLOPs) to assess and mitigate systemic risks (e.g., misinformation, electoral interference), increases transparency and user control over recommender systems and targeted advertising, restricts the use of sensitive data and targeting of minors in advertising, and strengthens enforcement through coordinated EU oversight and substantial financial penalties, among other new requirements (Eur. Comm. 2026b). The European Commission provides the annual transparency reports for the VLOPs and VLOSEs here: https://digital-strategy.ec.europa.eu/en/policies/dsa-brings-transparency. The European Commission provides summary statistics on the statements of reasons submitted by providers of all relevant online platforms here: https://transparency.dsa.ec.europa.eu/.

to create "actual knowledge" for purposes of Article 6. Article 17 requires platforms to provide a statement of reasons for each content moderation decision. Article 20 introduces an internal complaint-handling system through which users can appeal removal decisions. Together, these provisions create a low-cost, scalable channel through which firms can request the removal of employee reviews, and a strong incentive for platforms to comply. The cost of under-removal (potential liability, fines of up to 6% of total global revenue) may greatly exceed the cost of over-removal (one review disappears; the reviewer may appeal but faces effort costs).[5]

The enforcement infrastructure behind these provisions is active and growing. Although no company has been fined the penalty of 6% of global revenue, in December 2025 X (formerly Twitter) was fined €120 million under the DSA for violations to specific articles of the regulation regarding marketing, consumer protection, and data access. The fine marked the first enforcement action under the DSA (Eur. Comm. 2025a). That same month, the European Commission published its first annual systemic risk report covering a designated set of online platforms and search engines (Eur. Comm. 2025b).[6] These developments illustrate that the DSA is not merely a statutory framework but an active regulatory apparatus with ongoing oversight.

Under the formal notice-and-action mechanism outlined by the DSA in Article 16, firms can submit dozens of removal notices in a single day at approximately zero marginal cost. Online platforms have increasing freedoms in removing content. The breadth of "illegal content" under EU law amplifies this channel. For example, reviews naming specific managers can be flagged as

[5] The DSA provides for disputing content moderation internally and an out-of-court dispute settlement. Dispute settlement will be adjudicated by out-of-court dispute settlement (ODS) bodies certified by their respective national Digital Services Coordinator. Dispute settlement is intended to be free of charge or at a low cost. If the ODS body settles the dispute in favor of the user, the online platform bears all the fees (Eur. Comm. 2026a).

[6] Another example of the regulation's scope being tested involves Amazon. In November 2025, the EU General Court dismissed Amazon's challenge to its designation as a Very Large Online Platform, ruling that even marketplaces can pose systemic risk through the dissemination of content that may infringe fundamental rights (Meijer 2025). Amazon had argued that as a marketplace it merely sells goods and does not disseminate or amplify information, but the court rejected this distinction.

privacy violations under the GDPR, and reviews alleging misconduct can be flagged as defamatory. This regulatory design creates conditions under which the notice-and-action system can be exploited for strategic purposes as a byproduct to the regulation's objective of addressing genuinely harmful content.

2.2 Employer Reviews and Information Quality

A substantial literature documents that employer reviews on platforms like Glassdoor convey economically meaningful information. Green et al. (2019) show that Glassdoor ratings predict stock returns, Edmans (2012) links employee satisfaction to long-run firm value, and Huang et al. (2015) demonstrate that employee satisfaction predicts corporate performance. Marinescu et al. (2021) and Sockin and Sojourner (2023) establish that job seekers rely on employer reviews in their search decisions, and Barrios et al. (2026) document the rapid growth of research using labor market data, including Glassdoor reviews, at the intersection of labor economics and financial disclosure. Together, this evidence establishes that the information content of employer reviews matters for both capital and labor market efficiency.

The reliability of this information, however, depends on the integrity of the review population visible to platform users. Glassdoor maintains a Trust and Safety team that reviews content before publishing, enforces community guidelines, and states that employers cannot pay to have reviews removed (Glassdoor 2025a). Yet Glassdoor's support pages warn users against third-party reputation management services that claim to be able to remove reviews from the platform, an acknowledgement that an industry exists to manipulate content moderation processes

on behalf of employers.[7] The tension between Glassdoor's stated neutrality and the industry that has emerged around manipulating its content is central to our paper: the DSA may have provided a new, legally sanctioned channel through which this industry operates.

Mayzlin et al. (2014) and Luca and Zervas (2016) document strategic manipulation of consumer reviews through fake positive postings. Our paper examines the complementary channel: manipulation through selective deletion of authentic negative reviews. While the addition of fake positives increases the volume of uninformative content, the deletion of genuine negatives removes precisely the reviews that are most informative about employer quality. Anderson and Magruder (2012) show that even small changes in visible ratings can have economically meaningful effects on outcomes, suggesting that the marginal deletion of a few negative reviews per firm could distort the information environment faced by job seekers and investors.

2.3 Content Moderation Regulation and Strategic Behavior

A small but growing literature examines the behavioral consequences of content moderation regulation. Jiménez Durán et al. (2024) studies Germany's "Network Enforcement Act" (NetzDG), finding effects on political speech volume and composition. Jiménez-Durán (2022) provides experimental evidence on how content moderation rules affect user behavior on Twitter. Gorwa et al. (2020) survey the technical and political challenges in algorithmic content moderation[8].

[7] Companies such as Removify, runner-up winner to 2021 Deloitte Australia's Technology Fast 50, offer services to remove negative Glassdoor content (Deloitte 2026). In response to how content is deleted, the company states: "Most cases will involve us creating, submitting and managing a detailed application for the removal of the offending content by the publishing website. We liaise with most of these sites on a daily basis, and it is our experience from analyzing thousands of these applications that we have determined which process is likely to be successful for each type of content" (Removify 2026). Lawful business services such as this, as well as individual firm behavior, could increase removals post-DSA.

[8] Recent evidence also suggests potential unintended consequences of policies like mandatory wage disclosures (Jeong , 2024), suggesting the need to understand the effects of regulations.

What is largely absent from this literature is evidence on whether content moderation regulation is exploited by economic actors with reputational incentives for purposes unrelated to the regulation's stated objectives. The DSA's notice-and-action system, characterized by low filing costs and liable knowledge attribution, creates an institutional environment in which such exploitation is plausible. We contribute to this literature by examining whether the DSA's infrastructure is associated with changes in the nature and determinants of review deletion on Glassdoor.

2.4 Empirical Predictions

The effect of the DSA on review deletion is theoretically ambiguous, and the direction of each prediction depends on which economic forces dominate. The notice-and-action mechanism created by Article 16 lowers the cost of filing removal requests, giving firms a scalable tool for targeting reviews they find objectionable regardless of whether those reviews violate any content policy. The liability asymmetry created by Articles 6 and 16(3) means that platforms bear legal risk for retaining flagged content but face only a potential user initiated deletion-appeal, tilting the moderation calculus toward compliance with removal requests. Both forces push toward increased deletion of reviews, in particular unfavorable ones.

Several countervailing forces, however, could limit strategic exploitation or even improve the information environment. The notice-and-action mechanism (Article 16) effectively crowdsources detection. Employees, firms, and third parties can flag content that Glassdoor's internal systems may miss. Furthermore, the DSA imposes structured procedures and safeguards. Platforms must disclose moderation policies (Article 14), provide statements of reasons for each decision (Article 17), and offer users both internal appeals (Article 20) and access to out-of-court

dispute resolution (Article 21), creating ex post checks on erroneous removals. More broadly, transparency and reporting requirements increase scrutiny over moderation outcomes. If most notices target genuine violations, these features will increase the removal of problematic reviews while leaving compliant reviews largely unaffected, improving the informativeness of the review pool. Glassdoor's business model depends on credible content. Widespread strategic removal would erode user trust and platform value, creating incentives to resist low-quality requests despite legal risk. It is therefore an empirical question whether the DSA's cost-lowering effects dominate or whether these countervailing forces preserve or improve the pre-DSA information environment.

A related question is which types of content violations are associated with removal. Our classification distinguishes between low-discretion violations, where the presence of a violation can be relatively straightforwardly verified (e.g., explicit threats, spam, naming individuals), and high-discretion violations, which require subjective judgment about context and intent (e.g., whether a review is "trolling" or merely critical, whether content is "low quality" or simply unfavorable). Under a strategic exploitation account, firms filing removal requests may gravitate toward high-discretion categories precisely because these labels are harder for the platform to reject and harder for the reviewer to contest. If so, high-discretion violations should become more predictive of deletion post-DSA, while low-discretion violations, which are more readily verifiable and less amenable to strategic framing, should become less predictive. This grouping reflects how readily a violation can be verified from the review text, not the plausibility of a strategic notice; some low-discretion categories, such as privacy claims based on naming individuals, can also serve as the basis for strategic removal requests. Alternatively, if the DSA simply improves platform enforcement, all violation categories should become more predictive of deletion, reflecting better

detection across the board. The within-category pattern thus helps distinguish between these accounts.

Finally, we consider how visible ratings relate to labor market outcomes. If firms selectively remove negative reviews, the visible average rating becomes an upward-biased signal of employer quality. The behavioral consequences depend on how workers respond to that signal. Under a benchmark in which ratings accurately reflect quality, higher ratings should predict more inflows and fewer outflows. If deletion inflates ratings and workers take the posted signal at face value, the association between ratings and inflows should strengthen, and the association between ratings and outflows can shift toward positive as workers who joined on the basis of curated ratings reassess their match. Alternatively, if workers recognize the possibility of strategic deletion after the DSA, they may discount Glassdoor employer ratings and supplement them with other information sources, leaving the informativeness of ratings unchanged. The joint behavior of inflows and outflows therefore helps distinguish whether workers respond to the curated signal or see through the distortion. These are predictions about associations between visible ratings and subsequent flows; we discuss the limits of a causal interpretation in Section 5.2.

3. Research Design

3.1 Data and Sample

Our data comes from Revelio Labs, a workforce intelligence provider that systematically collects and archives Glassdoor review data. We access the data through the Wharton Research Data Services (WRDS) platform. The dataset covers the near-universe of Glassdoor reviews with rich metadata including star ratings (overall and sub-categories), free-text fields (pros, cons, and advice), reviewer role, firm identifiers, and country information. Detailed variable definitions are provided in Appendix A.

A key feature of the Revelio data for our identification is that deleted reviews are retained in the dataset. Revelio scrapes Glassdoor approximately once per month (with a subset of firms scraped weekly) and stores all reviews it encounters, regardless of whether those reviews are later removed from the platform. When a review is deleted from Glassdoor, Revelio simply stops observing it in subsequent scrapes, but the review remains in the historical record. Because WRDS does not include information on when each review was last observed, we obtained a supplemental file from Revelio containing the last-scraped date for each review. We identify deleted reviews by comparing each review's last-scraped date to the most recent scrape date for its firm (defined as the maximum last-scraped date across all of the firm's reviews). If a review was last observed more than 45 days before its firm's most recent scrape, we classify it as deleted, since the firm's page was visited at least once after that date and the review was no longer present.[9] For reviews classified as deleted, we use the last-scraped date as a proxy for the approximate timing of deletion.

This approach has an important limitation: reviews that are posted and removed between two consecutive scrapes will never appear in our data, because the review has entered and exited the platform environment before Revelio next visits that firm's page. Our deletion counts therefore represent a lower bound on actual deletion activity, and this measurement error is likely biased against finding results, since the fastest deletions, potentially the most aggressive cases of strategic removal, are the ones most likely to escape detection.[10]

[9] Because Revelio scrapes all of Glassdoor approximately once per month, a surviving review should be refreshed at least monthly, and its last-scraped date should fall close to its firm's most recent scrape date. The 45-day threshold provides a comfortable margin above the monthly scraping cycle, ensuring that reviews are not misclassified as deleted due to minor variation in scraping timing. It also ensures that reviews near the end of our data collection window are not misclassified as deleted simply because insufficient time has elapsed for subsequent scrapes to confirm their presence. The last-scraped date proxies deletion timing with error of up to the scraping interval, so deletions occurring close to the enforcement date may be assigned to either side of it. This timing noise affects month-level dynamics more than the twelve-month comparison.

[10] Glassdoor's community guidelines also provide that if an employee submits a new review of the same employer within 365 days, the earlier review may be removed. Because employees may use different accounts across

We construct our sample by selecting all reviews within a twelve-month window on each side of the DSA enforcement date (February 17, 2023 through February 17, 2025). The symmetric window is long enough to provide statistical power while limiting exposure to confounding events, and it balances seasonal variation in posting and deletion across the pre- and post-DSA periods. We construct a balanced firm-by-month panel by retaining firms with at least one identified deletion during the event window, yielding 96,395 firms observed over 25 months (2,409,875 firm-month observations). We zero-fill firm-months with no deletion activity to form a complete balanced panel. We classify each review as pre-DSA or post-DSA based on the event date: for deleted reviews, the event date is the last-scraped date (proxy for deletion date); for surviving reviews, it is the review posting date.

3.2 LLM Violation Classification

A central challenge in our setting is determining whether deleted reviews contain genuine content policy violations or whether they are substantively legitimate reviews that happen to be unfavorable.

Because review text is unstructured, spans multiple languages, and requires contextual judgment to evaluate, we deploy GPT-5.4 Nano, a large language model, to classify each review text against eleven violation categories derived from Glassdoor's published content guidelines, terms of use, and discussions of content moderation (Appendix D provides the mapping): (1) Abuse/profanity, (2) Violence/threats, (3) Discrimination/hate speech, (4) Trolling/inflammatory content, (5) Privacy/naming individuals, (6) Proprietary/confidential information, (7) Spam/promotional content, (8) Low quality/unintelligible text, (9) Irrelevant/off-topic content,

submissions, we cannot reliably identify duplicate reviews from the same individual. To the extent that this mechanism is stable over time, it would not affect our pre- versus post-DSA comparisons.

(10) Fraudulent/inauthentic content, and (11) Salary/compensation disclosure (Glassdoor 2025b)[11]. The model evaluates reviews in any language and returns binary flags for each category. Throughout the paper, terms such as clean and no identifiable violation refer to the absence of a violation identifiable from the review text under our reading of Glassdoor's published guidelines. They do not rule out removal grounds that are not observable from the text, such as the truth of factual claims underlying a defamation notice, and they do not capture non-moderation removals such as user-initiated deletion or review replacement.

We aggregate the eleven categories into two higher-level groupings based on the degree of discretion involved in adjudicating a violation. Low-Discretion violations involve claims whose presence can be relatively straightforwardly verified from the review text alone—violence/threats, privacy/naming individuals, spam/promotional content, proprietary/confidential information, and fraudulent/inauthentic content. High-Discretion violations require more subjective judgment about context and intent—abuse/profanity, discrimination/hate speech, trolling/inflammatory content, low quality/unintelligible text, and irrelevant/off-topic content. We classify salary/compensation disclosure separately, as it reflects workplace information sharing rather than a content policy violation in the traditional sense. We also compute lift ratios, the prevalence of each violation among deleted reviews divided by prevalence among surviving reviews, as a measure of how strongly each violation type predicts deletion.

[11] In communication with Glassdoor's Trust & Safety team and as stated in its community guidelines, reviews may be removed outside of content moderation through two mechanisms: (i) user-initiated deletion, and (ii) replacement when a user submits a new review for the same company within 365 days, in which case only the most recent review remains visible. We do not expect user-initiated deletions to vary systematically with the introduction of the DSA. To address potential bias from review replacement by newer posts, we restrict the sample through February 2025, allowing sufficient time for within-year updates to occur. This ensures that the probability of review replacement is comparable in both the pre- and post-DSA periods. Glassdoor also removes reviews for companies who offer special incentives for positive reviews and for the same individual who posts multiple reviews across different accounts, where detected (Glassdoor 2025c).

Two features of our research design mitigate concerns about LLM classification accuracy. First, we do not expect to reproduce the Glassdoor proprietary algorithmic classification but rely on violation classifications based on our interpretation of the Glassdoor community guidelines. Our primary inferences rely on changes in violation prevalence between the pre- and post-DSA periods and between EU-exposed and non-EU-exposed firms, rather than on the absolute level of violations in any single period or group. Any constant measurement error in the LLM classification, including language-specific bias, category-specific false negative rates, or systematic differences in model performance across violation types, attenuates rather than reverses these comparisons; our inferences require only that classification error not vary differentially across periods and treatment groups. Furthermore, the country-by-industry fixed effects absorb cross-sectional variation in language composition that could affect classification accuracy, and the industry-by-year-month fixed effects absorb time-varying sector-wide shifts in the types of reviews posted or the language patterns the classifier encounters.

Second, the threat to our inferences is not that the classifier is noisy but that its error rate changes differentially across treatment groups and periods. For our findings to be an artifact of misclassification, GPT-5.4 Nano would need to systematically miss more violations in post-DSA deleted reviews from EU-exposed firms than in pre-DSA deleted reviews or reviews from non-EU-exposed firms. Because reviews on both sides of the enforcement date are drawn from the same platform, written in the same languages, and follow the same format, there is no obvious reason for differential classifier performance across periods or treatment groups.

Table OA1 provides validation statistics for the LLM classification. We randomly sample 500 reviews and have five independent human graders evaluate each review for content policy violations. Panel A reports an overall agreement rate of 83.6% between the LLM and the human

benchmark, under which a review is treated as flagged if at least one of the five graders identifies a violation. The LLM identifies more violations than human graders, flagging 108 reviews versus 76, a 42% increase. To the extent that the classifier overflags violations on average, our estimate that 86.17% of deleted reviews contain no text-identifiable violation is unlikely to overstate the share of clean deletions. Panel B stratifies agreement by the strength of human consensus. When four or more of five graders independently identify a violation, the LLM agrees in every case. Disagreements are concentrated in borderline reviews where the human graders themselves do not reach consensus, suggesting that the classifier's errors arise in precisely the cases where reasonable evaluators also disagree. Panel C shows that restricting attention to higher confidence flags increases accuracy to 90.2%; our main analyses use no confidence threshold, which is the most conservative approach.

3.3 Key Variables

Our primary dependent variables operate at two levels. At the review level, we study the characteristics of deleted reviews, including star ratings, sub-ratings, tone, violation flags, and content structure, as well as the binary deletion indicator. At the firm-month level, our main dependent variable is the count of deletions. We decompose deletions by violation status (clean versus flagged), by discretion level (low-discretion versus high-discretion), and by review characteristics including tone extremity, rating harshness, and the balance of negative versus positive content. We also examine effects on the visible information environment by testing whether the average displayed rating, computed over surviving reviews only, changes differentially for treated firms post-DSA.[12]

[12] The Glassdoor Help Center states: "Glassdoor ratings are calculated using a proprietary algorithm based on employee feedback, with recent reviews carrying more weight" (Glassdoor 2025c). We are unable to reflect the *actual* visible rating as seen on the Glassdoor site so calculate company ratings based on the surviving reviews. Because the

Our identification relies on a difference-in-differences design. The key independent variable is the interaction of Treated and Post, where Post equals one for events occurring after the DSA enforcement date (February 17, 2024) and Treated equals one for firms with EU exposure, defined as those headquartered in an EU member state or with pre-DSA EU-based reviewers. The coefficient on this interaction captures the differential change in deletion behavior for EU-exposed firms after DSA enforcement, relative to non-EU firms over the same period. Because any individual or entity can submit an Article 16 notice regardless of location, and because Glassdoor may implement compliance changes platform-wide, non-EU-exposed firms are unlikely to be fully insulated from the regulation. To the extent that control firms are partially exposed, our estimates understate the total effect of the DSA and should be read as the differential response associated with EU exposure.

Several of the review characteristic variables require construction beyond the raw Glassdoor fields. Tone is measured using the multilingual XLM-RoBERTa-base model (Barbieri et al. 2022), which produces positive and negative probability scores for the review's cons text[13]. We assign each review a primary tone direction based on which probability is higher, then identify the most extreme five percent within each directional group as Extreme Negative Tone or Extreme Positive Tone. Because tone is measured on the cons text, Extreme Positive Tone captures

weighting algorithm is unobserved, our surviving-review average approximates the displayed rating. If recent reviews carry more weight and deletions are concentrated in recent negative reviews, a pattern we observe in the data, our estimate would understate the effect on the displayed rating; because we cannot verify the algorithm, we treat the direction of this approximation error as uncertain.

[13] The multilingual XLM-RoBERTa base model is trained on approximately 198 million tweets and fine-tuned for sentiment analysis. Glassdoor content closely resembles user-generated, opinion-based text on Twitter, making this setting well aligned with the model's training domain. The model is fine-tuned on eight languages (Arabic, English, French, German, Hindi, Italian, Spanish, and Portuguese) while still producing predictions for other languages present in our Glassdoor dataset, albeit with lower expected accuracy.

unusually benign content in the field where criticism is expressed, rather than overall review positivity.

We also exploit the structure of Glassdoor reviews, which require separate free-text entries for positive aspects ("pros") and negative aspects ("cons") of the employer. The relative length of these two fields provides a simple measure of whether a review leans toward praise or criticism. Reviews where the cons text dominate are classified as *Cons-Heavy*, and those where the pros text dominates as *Pros-Heavy*.

In addition to the overall star rating, Glassdoor collects sub-ratings across five workplace dimensions: compensation and benefits, culture and values, senior leadership, work-life balance, and career opportunities. A reviewer who rates the company five stars overall but assigns one star on senior leadership is expressing pointed dissatisfaction that the overall score alone would mask. We use these sub-ratings to construct two boundary indicators. *Any Harsh* identifies reviews where the reviewer assigned the lowest possible score on any dimension, capturing cases of severe dissatisfaction with at least one aspect of the employer. *Universally Perfect* identifies reviews where every dimension received the highest possible score, a pattern more consistent with promotional intent than a balanced assessment.

Each of these review-level indicators is then aggregated to the firm-month level by counting the number of deleted reviews with each characteristic. These counts form the dependent variables in our decomposition analysis, allowing us to test which types of reviews are disproportionately targeted for removal after DSA enforcement.

3.4 Descriptive Statistics

Table 1 reports the review-level descriptive sample restricted to firms in the main regression universe. The sample contains 7,984,833 reviews posted within the twelve-month

window on each side of the DSA enforcement date. Of these reviews, 369,234 are deleted and 7,615,599 survive, implying an overall deletion rate of 4.62%. The review-level sample covers 96,330 represented firms, drawn from the 96,395 firms in the balanced firm-month regression panel. Panel B shows that deleted reviews are lower rated (mean 3.310 versus 3.687 for surviving reviews) and nearly twice as likely to be negative (34.9% versus 18.3%).

These patterns are visible in the time-series and cross-sectional figures. Figure 1 plots total monthly deletion volume relative to the pre-DSA average, revealing a sharp spike in the month immediately following the DSA enforcement date and a sustained elevation in deletion activity throughout the post-DSA period. Figure OA2 decomposes these deletions into those with and without LLM-detected content violations, showing that the post-DSA increase is driven overwhelmingly by reviews with no identifiable policy violation.

Figure 2 displays deletion rates by star rating, revealing a U-shaped pattern. The elevated deletion rate at the top of the rating distribution is notable: five-star reviews are deleted at roughly twice the rate of three- and four-star reviews. This is consistent with the platform removing reviews that appear inauthentic or promotional, such as employer-solicited posts, even when they contain no explicit policy violation. The U-shape suggests that deletion reflects two distinct forces: removal of harsh criticism at the low end, and removal of suspected manipulation at the high end.

Beyond ratings, Table 2 compares the content characteristics of deleted and surviving reviews along several dimensions. Panel A shows that deleted reviews score lower not only on the overall rating but across all sub-rating dimensions, with all differences significant at the one percent level.

Panel B presents the prevalence of LLM-detected violations among surviving and deleted reviews. The violation classifier is a meaningful predictor of deletion: 13.83% of deleted reviews

are flagged for any policy violation, compared to 4.72% of surviving reviews, a lift of 2.93. Similarly, 3.86% of deleted reviews are flagged for low-discretion violations, relative to 0.83% of surviving reviews, a lift of 4.62. Lift captures overrepresentation in deletions where values above one indicate that flagged reviews are disproportionately more likely to be deleted. Despite this overrepresentation, 86.17% of deleted reviews contain no identifiable policy violation. Table 3 reports descriptive statistics for the balanced firm-month panel used in the main regression analyses.

Figure 3 visualizes the violation prevalence gap, comparing surviving and deleted reviews across no-violation, low-discretion, and high-discretion categories. Deleted reviews are more likely to contain identifiable violations, particularly in high-discretion categories such as abuse or profanity and discrimination. However, the overwhelming share of deletions appears to be driven by factors other than identifiable policy violations.[14]

Appendix E provides illustrative examples consistent with the preceding findings. Among reviews deleted before the DSA, content violations are often readily identifiable. For example, a one-star review titled "They will slit your throat" contains explicit violent language (flagged for Abuse/Profanity and Violence/Threats), and a five-star review titled "Application as Deck Cadet" is a job application rather than an employee review (flagged as Spam).

Finally, Tables OA2 through OA3 in the Online Appendix provide additional detail on the geographic, and occupational dimensions of deletion activity. Table OA2 reports the distribution of reviews and deletions by country, showing how deletion rates vary across jurisdictions. Table OA2 reports the distribution of reviews and deletions by reviewer role, allowing comparison of deletion rates across seniority levels.

[14] Table OA4 reports prevalence and lift ratios for each of the eleven individual violation categories.

3.5 Empirical Specification

Having established that deleted reviews differ systematically from surviving reviews, we turn to whether the DSA altered deletion behavior. Our primary specification is a difference-in-differences model estimated on a balanced firm-by-month panel:

$$Y_{i,t} = \beta * (Treat_i \times Post_t) + X_i'\gamma + \mu_{c\times j} + \delta_{j\times t} + \varepsilon_{i,t} \quad (1)$$

where $Y_{i,t}$ is a firm-month level outcome (number of deletions, number of clean deletions, number of flagged deletions) for firm *i* in month *t*, *Treat* equals one for firms with EU exposure (headquartered in an EU member state or with EU-based reviewers pre-DSA), *Post* equals one for months after the DSA enforcement date (February 17, 2024), $\mu_{c\times j}$ denotes country-by-industry fixed effects, $\delta_{j\times t}$ denotes industry-by-year-month fixed effects, and X_i is a vector of time-invariant control variables. Standard errors are clustered at the country level.

The control variables capture pre-DSA firm characteristics that may predict deletion activity independent of the DSA. We control for the average pre-DSA firm rating to try to capture that firms with lower baseline ratings may attract more critical reviews and thus face more deletion opportunities. We control for geographic complexity (log number of distinct reviewer countries), because firms with a more internationally dispersed workforce may be more likely to receive reviews that trigger content moderation across jurisdictions. We control for occupational breadth (log number of distinct job roles), because firms with more heterogeneous workforces may generate reviews covering a wider range of experiences. We control for competitive density (log number of competitors in the same country-industry cell) and the leave-one-out mean competitor rating, because firms in more crowded or higher-rated competitive environments may face stronger reputational incentives to manage their review profiles. The country-by-industry fixed effects absorb some of this variation. In particular, competitive density is measured at the country-by-

industry level and is therefore absorbed in our preferred specification; the remaining controls capture within-cell heterogeneity.[15]

4. Results

4.1 Main Effects

Table 4 reports estimates from the balanced firm-by-month panel, with the dependent variable being the count of reviews deleted per firm-month. We report the coefficient on *Post* × *Treat* across three progressively demanding fixed effect structures: no fixed effects (Column 1), country and year-month fixed effects (Column 2), and country-by-industry and industry-by-year-month fixed effects (Column 3). The coefficient is stable across specifications: 0.0286 ($p < 0.01$) in Columns (1) and (2), and 0.0300 ($p < 0.01$) in Column (3). The preferred specification in Column (3) implies an increase of 0.030 deletions per firm-month, or approximately 22% relative to the sample mean of 0.132, within the deletion-active sample described in Section 3.1. The stability of the estimate across increasingly demanding fixed effects suggests the result is not driven by unobserved country-level or industry-level trends.

A concern with any difference-in-differences design is that the treatment and control groups may not be comparable, or that they may follow different underlying trends that could explain the patterns we observe. Our fixed effect structure addresses part of this concern: country-by-industry fixed effects absorb time-invariant differences across groups, and industry-by-year-month fixed effects absorb sector-wide shocks common to all firms in a given industry and month.

[15] Our results are robust to using firm fixed effects (Table OA8). Although firm fixed effects absorb the Treat indicator and all time-invariant firm characteristics, the Treat × Post interaction remains identified, as Table OA8 shows. We prefer country-by-industry fixed effects in the baseline because they allow us to report pre-DSA firm characteristics as controls and to compare firms facing similar local industry environments while retaining the cross-sectional variation used in the heterogeneity tests.

To try to further assess the robustness of our analyses, we also estimate a dynamic difference-in-differences specification that replaces the single *Post* × *Treat* interaction with a full set of month-by-month interactions between the treatment indicator and event-time dummies. Figure 4 plots the resulting coefficients. The pre-treatment estimates are individually and jointly indistinguishable from zero, providing no evidence that EU-exposed and non-EU-exposed firms were on different deletion trajectories before the DSA. The post-enforcement coefficients show an immediate and sustained increase in deletion activity among EU-exposed firms, with the effect appearing in the first month after enforcement and persisting through the end of the sample window.

4.2 Clean vs. Flagged Deletions

Having established that EU-exposed firms increased deletions after the DSA, we next examine what types of reviews drive the increase. This decomposition is central to interpretation. Concentration among reviews with identifiable policy violations would suggest stricter enforcement. Concentration among reviews without identifiable violations, especially negative reviews, would suggest strategic removal.

We separately estimate the DiD model using the count of clean deletions and flagged deletions as dependent variables. Table 5, Panel A decomposes the estimated 0.030 increase in deletions. Clean reviews account for 0.027 of the increase, or 89%, while flagged reviews account for only 0.003. Table 5 Panel B shows that among flagged reviews, the increase is concentrated in high-discretion violation categories, while low-discretion violations show no significant change. The overwhelming concentration of the effect in clean reviews suggests that the deletion increase is not primarily driven by improved enforcement of existing content policies.

4.3 Which Reviews and Firms Are Targeted

We next examine whether the deletion increase is selective along two additional dimensions: the content of deleted reviews (Table 6) and the firm characteristics associated with larger treatment effects (Table 7).

Table 6 tests whether the deletion increase disproportionately targets reviews with negative content. If the DSA facilitated strategic removal, we would expect firms to target reviews that are most damaging to their reputation while leaving favorable reviews untouched. Each column uses a different dependent variable, measured as the number of deleted reviews in a firm-month that fall into a specific content category. We pair each negative category with its positive counterpart to assess whether the deletion increase is asymmetric.

Table 6 Panel A compares reviews classified as *harsh* (any user-submitted sub-rating dimension equal to one) against reviews classified as *universally perfect* (all rating dimensions equal to five). The DiD coefficient for *harsh* reviews is 0.015 and statistically significant at the 1% level, whereas the coefficient for *universally perfect* reviews is statistically insignificant. Panel B compares reviews with extremely negative tone against reviews with extremely positive tone. The increase is again more precisely estimated on the negative side: 0.003 ($p < 0.01$) for negative tone versus 0.002 ($p > 0.10$) for positive tone. Panel C compares *cons-heavy* reviews (where the cons section dominates the review text) against *pros-heavy* reviews. *Cons-heavy* reviews increase by 0.005 ($p < 0.05$), while *pros-heavy* reviews show no change (-0.000, $p > 0.10$). Across the three tests, the statistically significant increases are concentrated in reviews that cast the firm in a negative light, while the estimates for reviews with positive characteristics are statistically insignificant. In Panels A and B, the positive-category point estimates are similar in magnitude to their negative counterparts, and we do not directly test the difference between the paired

coefficients, so we view those two comparisons as suggestive. The clearest asymmetry appears in Panel C, where cons-heavy deletions increase and the pros-heavy estimate is essentially zero.

Table 7 shifts from review characteristics to firm characteristics, asking whether certain types of firms respond more strongly to the DSA. If the observed increase in deletions reflects strategic behavior, we would expect larger effects among firms with greater exposure to content moderation on the platform, firms facing stronger local competition, firms whose ratings were deteriorating, and firms receiving more negative reviews.

Each column of Table 7 interacts the Treat × Post coefficient with a different firm-level moderator. The deletion increase is significantly larger among firms with higher pre-DSA deletion activity ($p < 0.01$), firms in more competitive local markets ($p < 0.01$), firms whose average monthly ratings were declining prior to the DSA ($p < 0.01$), and firms with more negative pre-DSA review tone ($p < 0.01$). While none of these cross-sectional patterns is individually conclusive, taken together they are consistent with firms that stand to benefit most from strategic review management.

4.4 Strategic Removal vs. Improved Enforcement

Two accounts can explain the post-DSA increase in deletions. Under an improved enforcement account, Glassdoor expanded its moderation infrastructure around the enforcement date, and the additional deletions reflect the platform identifying more genuine violations. Under a strategic removal account, firms, or reputation management intermediaries acting on their behalf (Section 2.2), exploit the DSA's notice-and-action mechanism to request removal of critical but policy-compliant reviews. These accounts are not mutually exclusive: Glassdoor may well have upgraded its compliance systems in response to the DSA while firms simultaneously used the resulting infrastructure to pursue removals that go beyond content enforcement. We cannot

observe who initiates each deletion, so we rely on the composition of the incremental effect to assess the relative importance of each channel.

Several features of the results are difficult to reconcile with improved enforcement alone. Eighty-nine percent of the additional deletions involve reviews with no identifiable policy violation (Section 4.2), the increase is concentrated in reviews with negative content while positive reviews show no significant change (Section 4.3), and the effect is larger among firms with stronger reputational incentives to manage their review presence (Section 4.3). If improved enforcement were the primary channel, we would expect the additional deletions to contain more violations, not fewer.

Table OA5 provides further support. Panel A shows that, among treated firms, deleted reviews become lower-rated post-DSA across the overall rating and all sub-rating dimensions. The largest decline is for Culture and Values, followed by Career Opportunities, Compensation and Benefits, Senior Leadership, and Work-Life Balance. This pattern is consistent with firms targeting reviews critical of workplace environment and management rather than reviews with identifiable content problems.

Panel B compares violation patterns between EU-exposed and non-EU-exposed deleted reviews. Among EU-exposed firms, 11.9% of deleted reviews are flagged for any violation, compared to 15.4% for non-EU-exposed firms, a gap of 3.5 percentage points. The gap is present for both low-discretion violations and high-discretion violations. The lower violation prevalence among EU-exposed deletions is suggestive of the DSA enabling removal of reviews that lack identifiable policy violations.

The qualitative examples in Appendix E illustrate what these clean deletions look like in practice. Post-DSA deletions frequently contain measured workplace criticism with no detectable

violations: a two-star review from the United States describes workplace restructuring and communication failures, and a review from Israel reports concerns about management conduct, both with no profanity, no named individuals, and no identifiable policy breach. While no single piece of evidence rules out improved enforcement entirely, the overall pattern across Sections 4.1 through 4.3 and Table OA5 is more consistent with strategic removal playing a meaningful role in the post-DSA deletion increase.

4.5 What Predicts Deletion?

The prior analysis focuses on firm-month aggregates. To examine which review characteristics predict deletion at the individual review level, and how those predictors shift after the DSA, we estimate a linear probability model on a matched sample. The unit of observation is the individual review. We retain all 369,234 deleted reviews and randomly select an approximately equal number of surviving reviews, yielding 702,694 observations after excluding reviews with missing covariates. This design preserves variation in the deleted population while ensuring that the low base rate of deletion in the full sample does not obscure the determinants of interest.[16]

Table OA6 reports the triple interaction (*Determinant* × *Post* × *Treat*) from a specification that includes all determinants simultaneously. *Low-Discretion Violation* is the strongest predictor of deletion overall, but its interaction with *Post* × *Treat* is negative (−0.027, $p < 0.01$), indicating that clear policy violations became less predictive of deletion after the DSA. In contrast, *High-Discretion Violation* × *Post* × *Treat* is positive (0.013, $p < 0.05$), suggesting that subjective

[16] Because the matched sample retains all deleted reviews but only a subsample of surviving reviews, estimates from an unweighted linear probability model are conditional on the sampling design; both level effects and interaction coefficients can differ from their full-sample counterparts. We therefore interpret the coefficients in Table OA6 as describing changes in the conditional association between review characteristics and deletion within this matched sample, rather than as population marginal effects. deletion rate rather than the population rate of 4.62%, the estimated intercept and level effects should not be interpreted as population marginal effects. Our inferences rely on interaction terms (*Post* × review characteristics), which capture changes in the conditional association between review characteristics and deletion and are less sensitive to the sampling ratio.

infractions became relatively more important in post-DSA deletion decisions. *Extreme Negative Tone* × *Post* × *Treat* is positive (0.020, $p < 0.05$), consistent with a shift toward tone-based removal. *Any Harsh* × *Post* × *Treat* is essentially zero and statistically insignificant, as is *Extreme Positive Tone* × *Post* × *Treat* (0.016, $p > 0.10$).

Taken together, these results are consistent with a shift from content-based enforcement toward reputation-based removal. Before the DSA, reviews with clear policy violations were the primary targets for deletion. After the DSA, treated firms increasingly remove reviews that are negative in tone or involve subjective content judgments, regardless of whether they contain identifiable violations. This pattern is difficult to reconcile with the expectation that platform incentives to maintain content integrity or the DSA's own transparency requirements (Regulation (EU) 2022/2065, art. 17) would prevent strategic exploitation of the notice-and-action mechanism.

5. Consequences for Ratings and Labor Markets

5.1 Effects on Displayed Ratings

The preceding sections document that post-DSA, EU-exposed firms selectively increase deletions of negative, violation-free reviews. If this pattern reflects strategic removal, it should have a mechanical consequence: by removing low-rated reviews from the visible pool, firms can inflate the average rating that prospective employees observe. We test this prediction in Table 8 using two dependent variables: the total number of visible reviews (*N Reviews*) and the average displayed rating, defined as the cumulative mean of all reviews that remain on the platform.

Table 8 reports the *Post* × *Treat* coefficient under two fixed effect structures for each outcome. For review volume (Columns 1 and 2), the coefficient is small and statistically insignificant, indicating that the DSA did not change the total stock of visible reviews for EU-exposed firms relative to non-EU-exposed firms. This is consistent with the deletion volumes

documented in Section 4 being small relative to the total review stock. For the average displayed rating (Columns 3 and 4), the coefficient is positive and significant across both specifications, ranging from 0.007 to 0.010 ($p < 0.01$). Although the magnitude is modest in absolute terms, it confirms that the selective removal of negative reviews produces a measurable upward shift in the rating signal that job seekers observe.[17] The effect is identified from differential changes across treated and control firms, ruling out platform-wide trends in rating inflation.

This result connects the deletion findings to a concrete informational consequence: visible ratings become a less accurate summary of the underlying distribution of employee experiences. The remaining question is whether this distortion has implications for actual labor market behavior.

5.2 Labor Market Flows

To assess whether inflated ratings affect real outcomes, we examine whether the relationship between visible Glassdoor ratings and subsequent labor market flows changes after the DSA. If strategic deletion makes ratings less informative, the post-DSA association between ratings and labor market outcomes could shift in ways that reflect the distortion. Specifically, we estimate a triple-interaction model:

$$LaborFlow_{i.\,t+h} = \alpha + \beta_1(VisRating_{it} \times Post_t \times Treat) + X\Gamma + \gamma_{cj} + \varepsilon_{it} \qquad (2)$$

where $LaborFlow_{i.\,t+h}$ is a labor market outcome at a six-month horizon (outflow rate, inflow rate, or net flow) for firm *i* in period *t* through *h*, *VisRating* is the average rating computed over surviving reviews, and γ_{cj} denotes country-by-industry fixed effects. The vector X includes all constituent terms of the triple interaction (the main effects of *VisRating*, *Post, and Treat,* and

[17] Our surviving-review average approximates the rating displayed on Glassdoor, which weights recent reviews more heavily under a proprietary algorithm (Glassdoor 2025c). If deletions are concentrated in recent negative reviews, our measure would understate the effect on the displayed rating, although we cannot verify this directly.(Glassdoor 2025c)

the two-way interactions: *VisRating × Post, VisRating × Treat, and Post × Treat*) as well as the time-invariant firm-level controls from the main specification in Equation (1). The coefficient of interest is β_1, which captures the change in the association between visible ratings and labor flows post-DSA for treated firms relative to control firms.[18]

Table 9 reports the results. Coefficients are scaled by 100 for readability and represent percentage-point changes in labor flow rates. The triple interaction (*Visible Rating × Post × Treat*) is positive and significant for both outflows (4.60, $p < 0.01$) and inflows (7.32, $p < 0.01$). This indicates that post-DSA, a given increase in the visible rating for an EU-exposed firm became associated with a larger subsequent increase in both worker arrivals and worker departures. The inflow result is consistent with job seekers responding to the inflated rating signal: higher visible ratings attract more workers. The outflow result suggests that workers who join firms with curated ratings may subsequently discover that workplace conditions do not match the impression conveyed by the displayed reviews.

Table 10 decomposes these flows into external and internal components. External flows capture workers moving between firms; internal flows capture transfers within multi-establishment organizations. Coefficients are again scaled by 100. For outflows, the external component is 2.07 ($p < 0.10$) and the internal component is 2.45 ($p < 0.01$). For inflows, the external component is 5.35 ($p < 0.01$) and the internal component is 2.12 ($p < 0.01$). The inflow effect is driven primarily by external hiring, consistent with outside job seekers relying on visible ratings when evaluating potential employers. The outflow decomposition shows effects in both channels, suggesting that

[18] This specification uses a smaller sample than the main deletion tests because it requires matched employer-level labor flow data from Revelio Labs, which tracks individual worker transitions across firms and covers a subset of employers in our panel with sufficient transition records to construct reliable flow measures.

the downstream consequences of inflated ratings operate through both voluntary departures and internal reassignment.[19]

6. Robustness Tests

We subject the main findings to several robustness checks across alternative specifications, inference procedures, and sample definitions.

The main DiD result is stable across alternative functional forms (Table 11). Replacing the count of deletions with an extensive-margin indicator (any deletion in a firm-month) yields a coefficient of 0.007 ($p < 0.01$). Using a log specification yields 0.008 ($p < 0.01$). In both cases the sign, significance, and economic interpretation are consistent with the baseline count model. We also verify robustness to alternative inference procedures and find similar results clustering the standard errors at the firm level rather than the country level or including firm fixed effects (Table OA8). Inference is also unchanged under industry-level clustering, country-by-industry clustering, and two-way clustering by firm and time (Cameron et al. 2008).[20]

Our baseline specification uses a twelve-month window on each side of the DSA enforcement date. A wider window increases statistical power but also increases the risk that unrelated events contaminate the estimate. Narrowing the window to three months on each side isolates the period closest to the enforcement date, where the treatment effect should be sharpest and confounding events least likely. The coefficient under the narrower window is 0.119 ($p <$

[19] Two caveats apply to these results. First, they reflect changes in the association between visible ratings and subsequent flows rather than direct evidence that deletion-induced rating changes caused the flows; visible ratings remain endogenous in this specification. Second, contemporaneous firm-level changes, such as growth, restructuring, or shifts in workforce composition, could affect both ratings and flows. We therefore interpret the labor market evidence as suggestive and consistent with the mismatch interpretation, rather than as identifying the worker-learning mechanism.

[20] Table OA7 reports additional cross-sectional heterogeneity tests, including public sector status and firm size.

0.01), confirming that the effect is present in the immediate aftermath of the DSA and is not driven by trends that emerge only later in the sample period.

The results also hold across geographic subsamples. The United States accounts for approximately half of all reviews and deletions (Table OA2), raising the concern that a single country could drive the aggregate result. The deletion increase is present when excluding U.S. observations (0.036, $p < 0.01$). We also verify that the result is not driven by specific EU countries with other country-specific legislation (Germany and Austria); the coefficient remains significant (0.072, $p < 0.01$). The breadth of the effect across regions suggests that Glassdoor may have implemented compliance changes globally rather than restricting them to EU-based users.

To rule out the possibility that the result is an artifact of the panel structure or of chance, we conduct a placebo permutation test. We randomly reassign treatment status across firms 500 times, holding constant the number of treated and control firms, and re-estimate the DiD for each permutation. This generates a null distribution of coefficients under the assumption of no treatment effect. None of the 500 placebo coefficients exceeds the actual estimate in absolute value (permutation p-value = 0.002, computed as $(0 + 1)/(500 + 1)$), indicating that the result is unlikely to arise from mechanical features of the data or from random variation in treatment assignment.

We verify that the informativeness results (Tables 9 and 10) are robust to alternative flow horizons of three and twelve months (untabulated), with the pattern of significant inflow and outflow coefficients preserved across windows. This suggests that the labor market consequences of inflated ratings are not sensitive to the specific horizon used to measure worker transitions.

7. Conclusion

This paper examines how the European Union's Digital Services Act is associated with changes in the deletion of employee reviews on Glassdoor. Using a balanced panel in a twelve-

month window around the DSA enforcement date, we document that deletions increase significantly post-DSA for EU-exposed firms, with 89% of the incremental effect concentrated in reviews with no violation identifiable from the review text. The deletion increase is disproportionately concentrated in reviews with negative tone, harsh ratings, and content critical of workplace culture and senior leadership, while the corresponding estimates for reviews with positive characteristics are statistically insignificant. Cross-sectional tests indicate that the effect is concentrated among firms with higher pre-DSA deletion activity, greater local competition, declining rating trajectories, and more negative review tone.

These findings are more consistent with strategic exploitation of the DSA's notice-and-action mechanism than with improved content enforcement. If the post-DSA increase in deletions reflected better detection of policy violations, we would expect the additional deletions to contain more violations, not fewer. Review-level determinant tests reinforce this interpretation. After the DSA, clear policy violations became less predictive of deletion, while negative tone and subjective content judgments became more predictive, consistent with a shift from content-based enforcement toward reputation-based removal.

The consequences extend beyond the review platform itself. Selective deletion of negative reviews produces a measurable increase in average displayed ratings, and post-DSA, visible ratings become stronger predictors of both worker inflows and outflows. The inflow result is consistent with job seekers responding to inflated rating signals. The outflow result suggests that workers who arrive based on curated ratings may subsequently discover that workplace conditions do not match the impression conveyed by displayed reviews.

Our findings have implications for the design of content moderation regulation. The DSA's notice-and-action framework lowers the cost of filing removal requests while imposing

asymmetric liability on platforms, creating incentives that extend well beyond the removal of genuinely harmful content. Sanatizadeh et al. (2025) show that regulation can improve review authenticity when it curbs suppression; our evidence suggests that regulation can also degrade authenticity when it lowers the cost of removal. Platform regulators may wish to consider mechanisms that restore balance to the notice-and-action system, such as penalties for frivolous removal requests, mandatory disclosure of removal volumes by entity type, or audit requirements that evaluate the content justification for removals triggered by Article 16 notices. More broadly, well-intentioned content moderation regulation can produce opacity when economic actors have incentives to exploit the regulatory machinery for self-interested purposes.

REFERENCES

Anderson, Michael, and Jeremy Magruder. 2012. "Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database." *The Economic Journal* 122 (563): 957–89. https://doi.org/10.1111/j.1468-0297.2012.02512.x.

Barbieri, Francesco, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. "XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond." arXiv:2104.12250. Preprint, arXiv, May 11. https://doi.org/10.48550/arXiv.2104.12250.

Barrios, John Manuel, Jung Ho Choi, Carolyn Deller, Joseph Pacelli, and Heidi A. Packard. 2026. "Labor and the Corporate Information Environment." SSRN Scholarly Paper No. 6390718. Social Science Research Network, February 25. https://doi.org/10.2139/ssrn.6390718.

Böke, Jens, Brigham Brau, Stephen Glaeser, and Junyoung Jeong. 2024. "Freedom of Speech and Employee Disclosure." SSRN Scholarly Paper No. 4904664. Social Science Research Network, July 24. https://papers.ssrn.com/abstract=4904664.

Brau, B., & Jeong, J. (2023). Digital Picket Lines: Strategic Employee Disclosure. Available at SSRN 4527388.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." Review of Economics and Statistics 90 (3): 414–27. https://doi.org/10.1162/rest.90.3.414.

Deloitte (Deloitte). 2026. "2026 Tech Fast 50." https://www.deloitte.com/au/en/Industries/tmt/about/technology-fast-50.html.

Directive 2000/31/EC, CONSIL, EP, 178 OJ L (2000). http://data.europa.eu/eli/dir/2000/31/oj.

"DSA Transparency Database." 2026. https://transparency.dsa.ec.europa.eu/.

Edmans, Alex. 2012. "The Link Between Job Satisfaction and Firm Value, With Implications for Corporate Social Responsibility." *Academy of Management Perspectives* 26 (4): 1–19. https://doi.org/10.5465/amp.2012.0046.

Eur. Comm. (European Commission). 2022. "DSA: Landmark Rules for Online Platforms Enter into Force." Press Release. Brussels, November 15. https://ec.europa.eu/commission/presscorner/detail/en/ip_22_6906.

Eur. Comm. (European Commission). 2024. "Implementing Regulation Laying down Templates Concerning the Transparency Reporting Obligations of Providers of Online Platforms." November 4. https://digital-strategy.ec.europa.eu/en/library/implementing-regulation-laying-down-templates-concerning-transparency-reporting-obligations.

Eur. Comm. (European Commission). 2025a. "Commission Fines X €120 Million under the Digital Services Act." Press Release. Brussels, December 5. https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2934.

Eur. Comm. (European Commission). 2025b. "Digital Services Act Report Lays out Landscape of Systemic Risks Online." Press Release. November 20. https://digital-strategy.ec.europa.eu/en/news/digital-services-act-report-lays-out-landscape-systemic-risks-online.

Eur. Comm. (European Commission). 2026a. "Out-of-Court Dispute Settlement Bodies under the Digital Services Act (DSA)." March 10. https://digital-strategy.ec.europa.eu/en/policies/dsa-out-court-dispute-settlement.

Eur. Comm. (European Commission). 2026b. "The Impact of the Digital Services Act on Digital Platforms." March 10. https://digital-strategy.ec.europa.eu/en/policies/dsa-impact-platforms.

Glassdoor (Glassdoor). 2025a. "Can Paid Services Remove Glassdoor Reviews?" December 8. https://help.glassdoor.com/s/article/I-ve-seen-ads-for-services-that-say-they-can-get-negative-reviews-removed-from-Glassdoor-Do-they-work?language=en_US.

Glassdoor (Glassdoor). 2025b. "Community Guidelines." April 15. https://help.glassdoor.com/s/article/Community-Guidelines?language=en_US.

Glassdoor (Glassdoor). 2025c. "Ratings on Glassdoor." December 4. https://help.glassdoor.com/s/article/Ratings-on-Glassdoor?language=en_US.

Glassdoor (Glassdoor). n.d. "Trust Glassdoor to Keep Your Data Safe and Secure." Accessed March 28, 2026. https://about-us.glassdoor.com/site-us/security/.

Glassdoor, Inc. v. Andra Group, LP, 575 SW 3d 523 (Supreme Court 2019).

Glassdoor, Inc. v. Superior Court, 9 Cal. App. 5th 623 (Court of Appeal, 6th Appellate Dist. 2017).

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7 (1): 2053951719897945. https://doi.org/10.1177/2053951719897945.

Green, T. Clifton, Ruoyan Huang, Quan Wen, and Dexin Zhou. 2019. "Crowdsourced Employer Reviews and Stock Returns." *Journal of Financial Economics* 134 (1): 236–51. https://doi.org/10.1016/j.jfineco.2019.03.012.

He, Sherry, Brett Hollenbeck, and Davide Proserpio. 2022. "The Market for Fake Reviews." *Marketing Science* 41 (5): 896–921. https://doi.org/10.1287/mksc.2022.1353.

Healy, Paul M., and Krishna G. Palepu. 2001. "Information Asymmetry, Corporate Disclosure, and the Capital Markets: A Review of the Empirical Disclosure Literature." *Journal of Accounting and Economics* 31 (1): 405–40. https://doi.org/10.1016/S0165-4101(01)00018-0.

Holstead, C., J. Huang, and Jedson Pinto. 2025. "Employee Non-Disclosure Agreements and Corporate News." *Unpublished Working Paper, University of Texas at Dallas*.

Huang, Minjie, Pingshu Li, Felix Meschke, and James P. Guthrie. 2015. "Family Firms, Employee Satisfaction, and Corporate Performance." *Journal of Corporate Finance* 34 (October): 108–27. https://doi.org/10.1016/j.jcorpfin.2015.08.002.

Jeong, J. (2024). Salaries on Display: Unintended Consequences of Wage Disclosure. Available at SSRN.

Jiménez Durán, Rafael, Karsten Muller, and Carlo Schwarz. 2024. "The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany's NetzDG." *CAGE Online Working Paper Series*, CAGE Online Working Paper Series, 701. https://ideas.repec.org//p/cge/wacage/701.html.

Jiménez-Durán, Rafael. 2022. "The Economics of Content Moderation: Theory and Experimental Evidence from Hate Speech on Twitter." *Working Papers*, 324. https://ideas.repec.org//p/zbw/cbscwp/324.html.

Kaushal, Rishabh, Jacob van de Kerkhof, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi. 2024. "Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database." In The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), 1121–1132. https://doi.org/10.1145/3630106.3658960.

Kuczerawy, Aleksandra. 2021. "The Good Samaritan That Wasn't: Voluntary Monitoring under the (Draft) Digital Services Act." *Verfassungsblog*, ahead of print, January 12. https://doi.org/10.17176/20210112-181758-0.

Leuz, Christian, and Peter D. Wysocki. 2016. "The Economics of Disclosure and Financial Reporting Regulation: Evidence and Suggestions for Future Research." *Journal of Accounting Research* 54 (2): 525–622. https://doi.org/10.1111/1475-679X.12115.

Luca, Michael, and Georgios Zervas. 2016. "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud." *Management Science* 62 (12): 3412–27. https://doi.org/10.1287/mnsc.2015.2304.

Marinescu, Ioana, Andrew Chamberlain, Morgan Smart, and Nadav Klein. 2021. "Incentives Can Reduce Bias in Online Employer Reviews." *Journal of Experimental Psychology. Applied* 27 (2): 393–407. https://doi.org/10.1037/xap0000342.

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier. 2014. "Promotional Reviews: An Empirical Investigation of Online Review Manipulation." *American Economic Review* 104 (8): 2421–55. https://doi.org/10.1257/aer.104.8.2421.

Meijer, Bart H. 2025. "EU Court Dismisses Amazon's Request to Scrap EU Tech Label." Retail & Consumer. *Reuters*, November 19. https://www.reuters.com/business/retail-consumer/eu-court-dismisses-amazons-request-scrap-eu-tech-label-2025-11-19/.

Regulation (EU) 2022/2065, 277 OJ L (2022). http://data.europa.eu/eli/reg/2022/2065/oj.

Removify (Removify). 2026. “FAQ Archive.” https://removify.com/faq/.

Sanatizadeh, Aida, Gordon Burtch, Yili Hong, and Yuheng Hu. 2025. “Can Regulation Restore Authenticity? Evidence from Policies Against Online Review Suppression.” *Information Systems Research*, ahead of print, December 22. https://doi.org/10.1287/isre.2023.0436.

Sockin, Jason, and Aaron Sojourner. 2023. “What’s the Inside Scoop? Challenges in the Supply and Demand for Information on Employers.” *Journal of Labor Economics* 41 (4): 1041–79. https://doi.org/10.1086/721701.

Zuru, Inc. v. Glassdoor, Inc., 614 F. Supp. 3d 697 (Dist. Court 2022).

Appendix A: Variable Definitions

Variable	Definition
Treatment and Timing	
Treated	Indicator equal to one for firms headquartered in an EU member state or with pre-DSA EU-based reviewers.
Post	Indicator equal to one if event date falls on or after February 17, 2024. For deleted reviews, the event date is the last-scraped date (proxy for deletion timing). For surviving reviews, the event date is the review posting date.
Firm-Month Dependent Variables	
N Deleted	Count of reviews deleted in a given firm-month.
N Clean Deleted	Count of deleted reviews with no LLM-detected violation
N Flagged Deleted	Count of deleted reviews with at least one LLM-detected violation
N High-Disc Deleted	Count of deleted reviews flagged for at least one high-discretion violation
N Low-Disc Deleted	Count of deleted reviews flagged for at least one low-discretion violation
N Reviews	Total number of visible (surviving) reviews for a firm
Avg Displayed Rating	Mean rating_overall across surviving reviews only.
LLM Violation Flags	
Any Violation	Indicator equal to one if any of eleven LLM violation flags is triggered. Categories: (1) Abuse/profanity, (2) Violence/threats, (3) Discrimination/hate speech, (4) Trolling/inflammatory, (5) Privacy/naming individuals, (6) Proprietary/confidential, (7) Spam/promotional, (8) Low quality/unintelligible, (9) Irrelevant/off-topic, (10) Fraudulent/inauthentic, (11) Salary/compensation disclosure.
Low-Discretion Violation	Indicator equal to one if any low-discretion flag is triggered. Low-discretion violations involve claims that can be relatively straightforwardly verified from the review text: violence/threats, privacy/naming individuals, spam/promotional content, proprietary/confidential information, and fraudulent/inauthentic content.
High-Discretion Violation	Indicator equal to one if any high-discretion flag is triggered. High-discretion violations require subjective judgment: abuse/profanity, discrimination/hate speech, trolling/inflammatory content, low quality/unintelligible text, and irrelevant/off-topic content.
Lift	Prevalence of a violation flag among deleted reviews divided by prevalence among surviving reviews. Values above one indicate flagged reviews are disproportionately more likely to be deleted.
Any Harsh	
Universally Perfect	
Control Variables	
Pre-DSA Rating	Firm's average rating_overall computed over all reviews posted before February 17, 2024.
Geographic Complexity	Log of the number of distinct reviewer countries for a firm in the pre-DSA period.
Occupational Breadth	Log of the number of distinct job roles represented among a firm's reviews in the pre-DSA period.
Competitive Density	Log of the number of firms in the same country-industry cell.

Variable	Definition
Competitor Rating	Leave-one-out mean rating of all other firms in the same country-industry cell, computed over the pre-DSA period.
Additional variables	
Tone	Review sentiment measured using the multilingual XLM-RoBERTa-base model. Each review is assigned a primary tone direction based on the higher probability score.
Labor Flow Variables	
Outflow Rate	Departures divided by headcount, measured at a six-month horizon. Sourced from Revelio Labs via WRDS. Multiplied by 100.
Inflow Rate	New hires divided by headcount, measured at a six-month horizon. Multiplied by 100 in Tables 9–10.
Net Flow	Inflow rate minus outflow rate. Multiplied by 100.
External / Internal	External flows capture workers moving between firms; internal flows capture transfers within multi-establishment organizations. Reported separately in Table 10.
Visible Rating	Average rating computed over surviving reviews. Interacted with Post and Treated in the triple-interaction specification.

Appendix B: Conceptual Channels of Glassdoor Content Moderation

Figure B.1 provides a visual representation of the content moderation channels on Glassdoor before and after the DSA. Before the DSA, Glassdoor's internal moderation team was the primary gatekeeper: reviews were removed based on automated filters and manual review against the platform's content guidelines. Firms could request removal through informal channels, but the absence of a mandatory response mechanism gave Glassdoor discretion to reject requests. The DSA introduces a formal notice-and-action channel that operates in parallel with the platform's internal moderation. Any party can submit a structured notice under Article 16, and the platform must process it, take a decision, and provide a statement of reasons. Because the platform's liability exemption under Article 6 is conditioned on acting expeditiously, the incentive structure favors removal over retention when the legal basis of a notice is even marginally colorable.

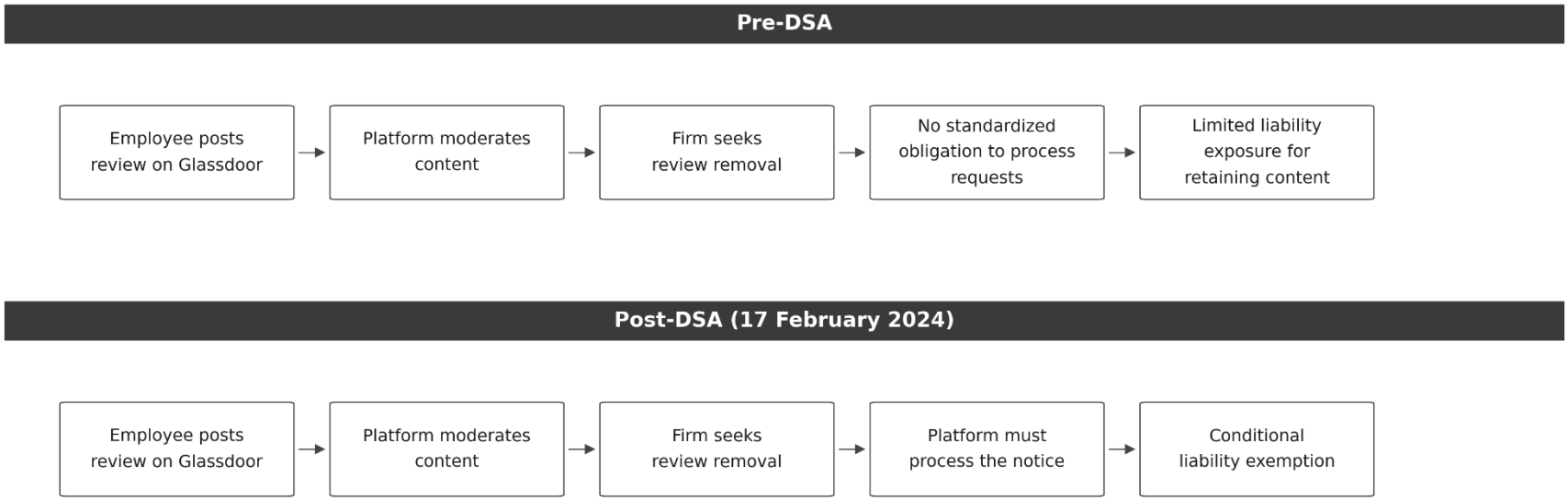
Figure B.1: Conceptual Channels of Glassdoor Content Moderation



Appendix C: Content Moderation Framework Before and After the Digital Services Act

This appendix illustrates the content moderation process before and after the Digital Services Act. Panel A depicts the pre-DSA environment, in which firms seeking review removal had no standardized mechanism; removal requests required informal engagement, legal action, or reputation management services, all at higher cost and with limited platform obligation to respond. Panel B depicts the post-DSA environment, in which Article 16 establishes a standardized, low-cost notice mechanism. Prior informal channels persist, but the new formal channel creates a mandatory obligation for the platform to process notices and a conditional liability exemption (Article 6) that incentivizes removal over retention.

Figure C.1: Content Moderation Framework Before and After the DSA



Appendix D: Glassdoor Community Guidelines and LLM Classification Mapping

Glassdoor publishes Community Guidelines that define the content standards for employee reviews on the platform. Reviews that violate these guidelines are subject to removal through Glassdoor's content moderation process. The guidelines are publicly available at *help.glassdoor.com/s/article/Community-Guidelines* (Glassdoor 2025b). We also draw on the Glassdoor Terms of Use (§4.2) and from content information provided, which enumerate additional prohibited behaviors including defamatory, fraudulent, harassing, threatening, and discriminatory content.

Table D1: Mapping of Glassdoor Community Guidelines to LLM Classification Categories

#	Glassdoor Guideline	LLM Category
1	Abuse & Insults	Abuse/Profanity
2	Violent Threats & Wishes of Harm	Violence/Threats
3	Hateful Content & Slurs	Discrimination/Hate
4	Trolling	Trolling/Inflammatory
5	Naming Non-Executives & Privacy	Privacy/Naming
6	Proprietary Information	Proprietary/Confidential
7	Spam & Advertising	Spam/Promotional
8	Content Quality	Low Quality/Unintelligible
9	Professional Irrelevance	Irrelevant/Off-topic
10	Fraudulent & Inauthentic Content	Fraudulent/Inauthentic
11	Salary Discussion	Salary/Comp. Disclosure

Appendix E: Examples of Deleted Reviews

Pre-DSA Examples of Deleted Reviews

Example 1

Rating: 1.0★ | Flags: Abuse/Profanity, Violence/Threats | Deleted: Pre-DSA | Country: United States

Summary: They will slit your throat

Pros: The employees are great people except for KC=evil

Cons: Management is slimy. A person can be a model employee and be terminated if they are sabotaged by KC Payroll doesn't pay hours. Not depositing 401k deductions

Advice: Get rid of rats

Example 2

Rating: 1.0★ | Flags: Slurs/Hate, Garbled | Deleted: Pre-DSA | Country: Ukraine

Summary: stay away from this company

Pros: I didn't see any pros

Cons: Total mess in all processes, huge turnover, toxic environment, hazing, abuse and bulling

Advice: Think about people, not just money-money-sales-sales, stop avoiding problems and bury your head in the sand when others (oldies) bully normal people

Example 3

Rating: 5.0★ | Flags: Spam | Deleted: Pre-DSA | Country: United States

Summary: APPLICATION AS DECK CADET

Pros: I'm a diligent person in work and ready to explore and learn something new every day and as well willing to move my career to the pick I would be great ful if I'm considered for the above position on Royal Caribbean group I'm available for work anywhere anytime thanks .

Cons: It has always been my dream serving onboard this ship and I never see any downside of it keep moving Royal Caribbean.

Advice: I advise that the management keep up the good work .

Example 4

Rating: 5.0★ | Flags: Privacy/Naming | Deleted: Pre-DSA | Country: United States

Summary: "A fantastic and enjoyable working atmosphere that makes working a pleasure"

Pros: It's no surprise that the MINTH Group has a good staff retention rate. Working is enjoyable, the compensation is competitive, coworkers are kind and polite, and management puts out their best effort. Employees are also highly rewarded and motivated. Jeremy, the plant manager, probably is the best colleague and senior I have met in recent years of working.

Cons: I understand that we all need to socialize at work, but it should be done with real enthusiasm on the side of every employee, not bringing some home and personal discussions to work.

Advice to Mgmt: The management should concentrate on amending the activities that workers are required to participate in, including many options so no employee will be left out.

Example 5

Rating: 2.0★ | Flags: NONE | Deleted: Pre-DSA | Country: Unknown

Summary: I don't recommend

Pros: -fully remote -random days off

Cons: -bad pay -multiple rounds of layoffs -leadership lied to us and said they weren't going to take a PPP loan after firing 30% of the company. they lied and actually received a 2 million dollar PPP loan (check Google)

Advice: -pay more competitively -stop lying

Appendix E: Examples of Deleted Reviews (Continued)

Post-DSA Examples of Deleted Reviews

Example 1

Rating: 2.0★ | Flags: NONE | Deleted: Post-DSA | Country: United States

Summary: Not great, not the worst

Pros: The people (most of them) are nice.

Cons: Morale has remained low since the layoffs. Chaotic restructuring and poor communication. I don't know anyone who actually likes their job here anymore, but there are worse places to work.

Advice: <NA>

Example 2

Rating: 2.0★ | Flags: NONE | Deleted: Post-DSA | Country: Unknown

Summary: High turnover

Pros: Nice consultants and most staff

Cons: Management needs to think if its time to move some of the furniture on

Example 3

Rating: 2.0★ | Flags: NONE | Deleted: Post-DSA | Country: Israel

Summary: Toxic managment

Pros: Paycheck, health care, dentist care.

Cons: Everything is mostly show, there is no equality, no inclusion, managers are just looking to advance on the backs of those below them. There is no equality in promotions and distribution of bonuses, things come down to personal reasons.

Advice: Senior managers abuse their power and discriminate based on geography and political opinions\Inequality in promotions and dismissals No oversight

Example 4

Rating: 5.0★ | Flags: Proprietary | Deleted: Post-DSA | Country: Unknown

Summary: Good workplace benefits

Pros: I like working at Inventys as there are many benefits like: Friendly work culture with skill development opportunities. Fantastic work-life balance and travel facility. Exposure to big clients and continuous learning chances. More than 20 large multinationals have shared their tech packs for contract manufacturing done at this company. Great Place to improve your life.

Cons: Be open to constructive criticism and learning new things.

Advice: Nothing major. Management is always implementing the feedback given.

Example 5

Rating: 5.0★ | Flags: Abuse/Profanity | Deleted: Post-DSA | Country: Brazil

Summary: Bom

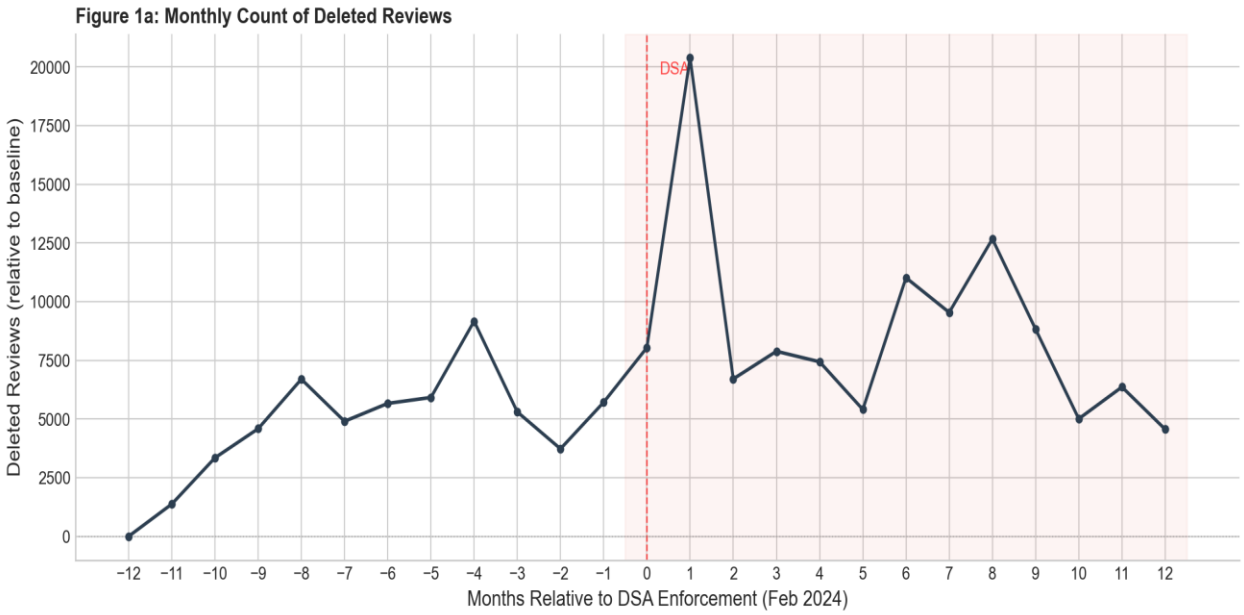
Pros: Localização. Beneficio e salário ajustados ao mercado. Quanto a gerência, alguns são bem solicitos, simpáticos. O restante são abusivos, arrogantes.

Cons: Refeitório distante dos galpões. Você não pode sair do galpão, portanto, não há horário de café. Café somente pra quem é de "confiança". Intenso sistema de segurança. (dificulta o ambiente de trabalho, ex: 3h pra sair com o lixo do galpão.)

Advice: Entendo que há colaboradores mal intencionados, e, já houve muitos no passado. Por isso entendo o intenso sistema de segurança. Creio que o certo seria punir os colaboradores que não estão dispostos a seguir as regras da empresa e filtrar os próximos candidatos para que as medidas rígidas não possam respingar em quem trabalha corretamente

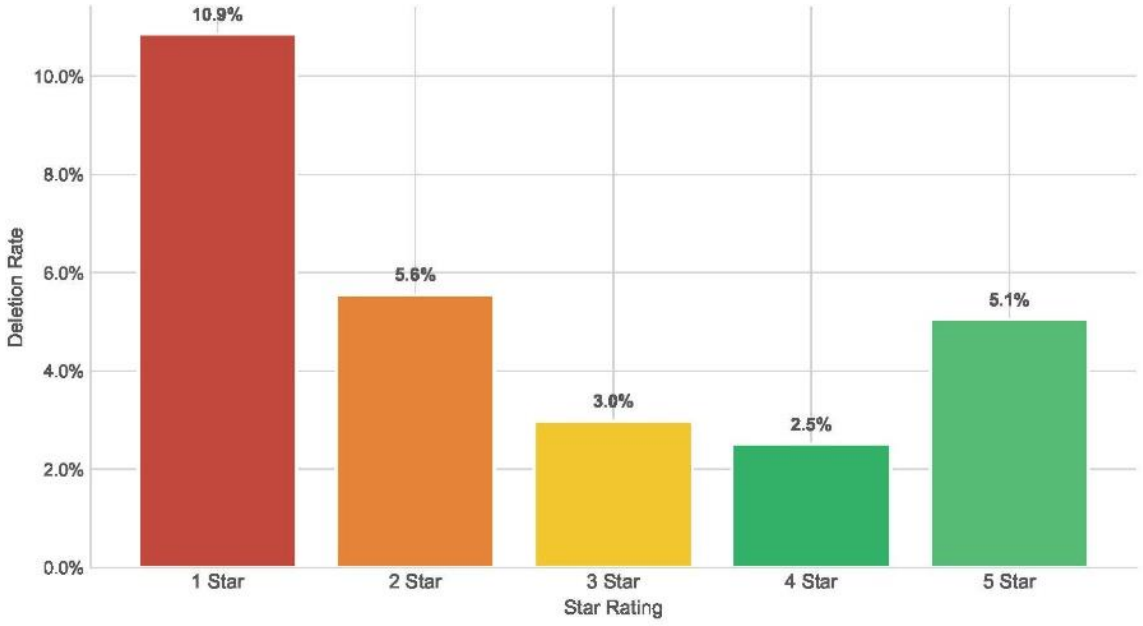
Figure 1 - Monthly Deletion Volume

Monthly Count of Deleted Reviews



This figure plots total monthly deletion volume relative to the pre-DSA average. The vertical dashed line indicates the DSA enforcement date (February 17, 2024). The sample includes all firms in the balanced regression panel observed in the 12-month window on each side of the enforcement date.

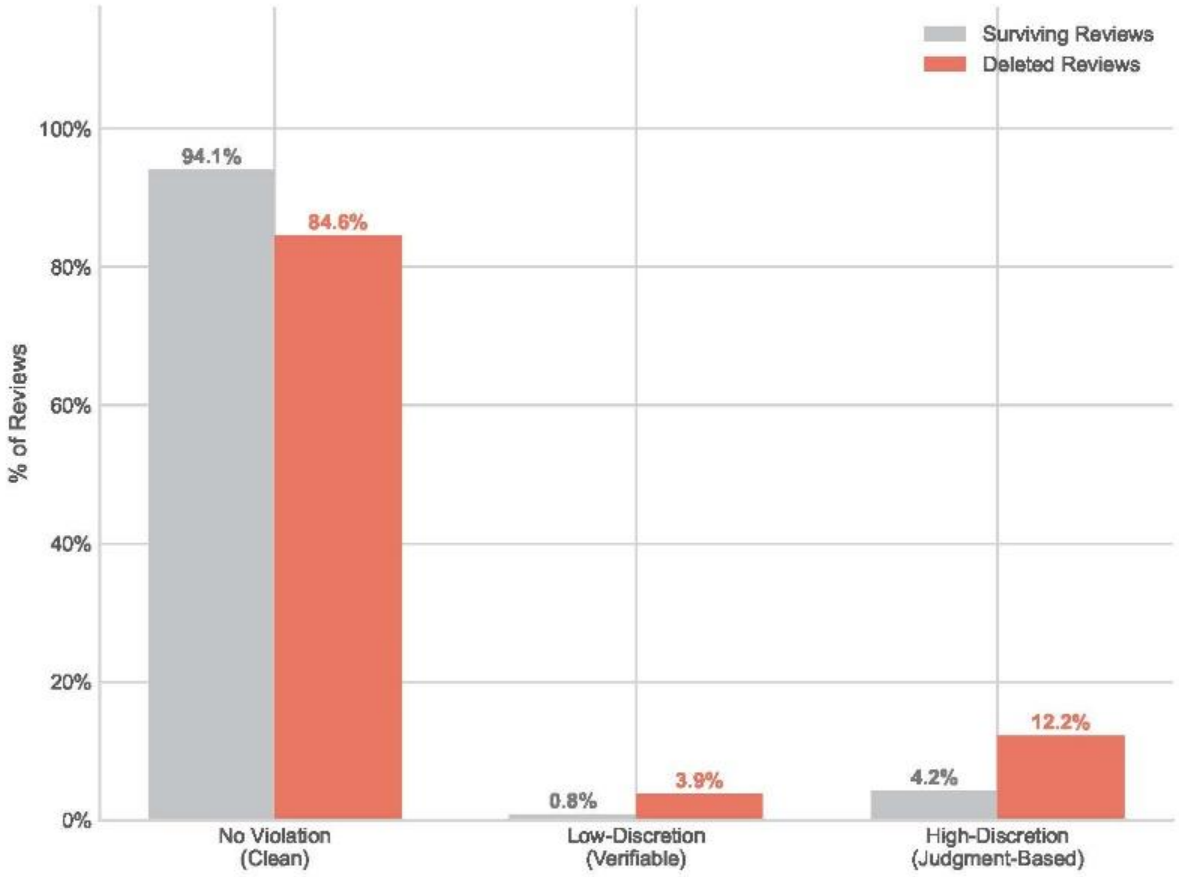
Figure 2 - Deletion Rate by Star Rating



This figure displays the deletion rate (number of deleted reviews divided by total reviews) for each star rating level. The U-shaped pattern shows that one-star and five-star reviews are deleted at the highest rates, consistent with removal of harsh criticism at the low end and suspected fake positives at the high end.

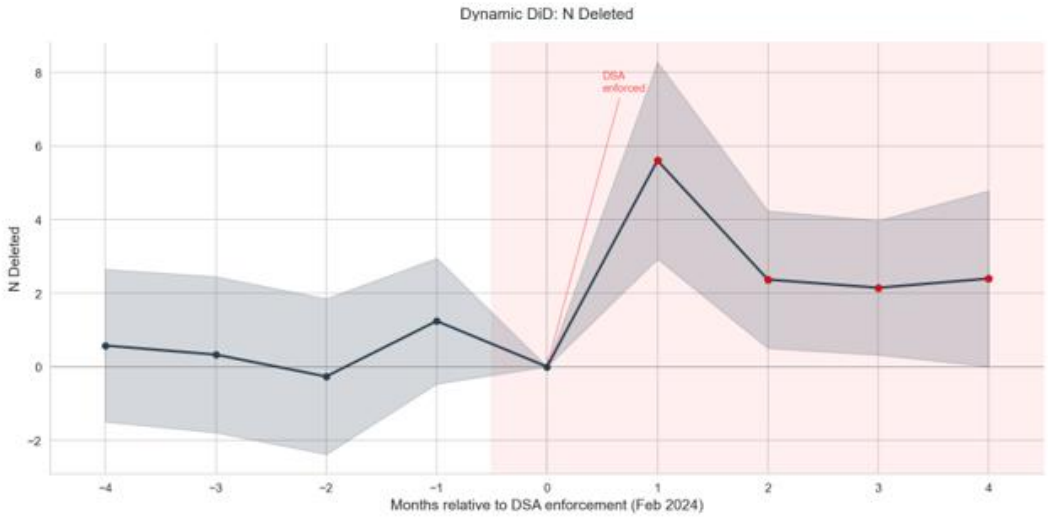
Figure 3 - LLM Violation Prevalence: Surviving vs. Deleted Reviews

LLM Violation Prevalence by Category Group



This figure visualizes the prevalence of LLM-detected content policy violations among surviving and deleted reviews. Surviving review violation rates are computed on a random subsample due to classifier cost constraints. See Section 3.2 for details. Violation flags are assigned by GPT-5.4 Nano.

Figure 4 – Event-Study Estimates of DSA Enforcement on Review Deletion Rates



This figure plots month-by-month difference-in-differences estimates for the first months following DSA enforcement. Each coefficient represents the treatment effect for a single event-time month. The dependent variable is the count of deletions per firm-month. The specification includes country-by-industry and industry-by-year-month fixed effects and the full set of pre-DSA firm-level controls. Coefficients are plotted with 99% confidence intervals. The vertical dashed line indicates the DSA enforcement date.

Table 1 – Sample Description

This table reports descriptive statistics for the review-level sample within the 12-month window on each side of the DSA enforcement date (February 17, 2024). The sample contains 7,984,833 reviews across 96,330 firms in the balanced regression panel. Panel A reports sample composition. Panel B compares the characteristics of surviving and deleted reviews. Panel C reports deletion timing relative to the DSA enforcement date. Deletion is identified by comparing consecutive Revelio Labs data snapshots; a review is classified as deleted if it was last observed more than 45 days before its firm's most recent scrape. All variables are defined in Appendix A.

Panel A: Sample Overview

	N
Total reviews	7,984,833
Deleted reviews	369,234
Surviving reviews	7,615,599
Deletion rate	4.62%

Panel B: Surviving vs. Deleted

	N	**Share (%)**	**Mean Rating**	**% Negative**
Surviving	7,615,599	95.4%	3.687	18.3%
Deleted	369,234	4.6%	3.310	34.9%

Panel C: Deletion Timing

	N
Surviving reviews	7,615,599
Deleted pre-DSA	116,396
Deleted post-DSA	252,838

Table 2: Content Characteristics of Deleted and Surviving Reviews

This table compares the content characteristics of deleted and surviving reviews. Panel A reports mean values for the overall rating and five sub-rating dimensions, along with the difference and statistical significance based on Welch two-sample t-tests. Panel B reports the prevalence of LLM-detected content policy violations among surviving and deleted reviews, along with the lift ratio (prevalence in deleted divided by prevalence in surviving). Violation flags are assigned by GPT-5.4 Nano based on eleven categories derived from Glassdoor's published content guidelines. All variables are defined in Appendix A.

Panel A: Review Characteristics

	Surviving	Deleted	Difference	t-stat
Overall Rating	3.6870	3.3102	-0.3769***	-138.38
Comp & Benefits	3.4510	3.2972	-0.1538***	-55.45
Culture & Values	3.5793	3.2556	-0.3237***	-105.91
Senior Leadership	3.2975	3.0510	-0.2465***	-79.16
Work-Life Balance	3.5151	3.3248	-0.1903***	-64.97
Career Opportunities	3.4747	3.2059	-0.2698***	-90.57

Panel B: Violation Rates

	Surviving (%)	Deleted (%)	Lift
Any Violation	4.72%	13.83%	2.93x
Low-Discretion	0.83%	3.86%	4.62x
High-Discretion	4.23%	12.22%	2.89x

Table 3 - Firm-Month Panel Descriptive Statistics

This table reports descriptive statistics for the balanced firm-month panel used in the main regression analyses. The panel contains 96,395 firms observed over 25 months (2,409,875 firm-month observations). Firm-months with no deletion activity are zero-filled. All variables are defined in Appendix A.

Regression Sample Descriptives	***(Firm-Month Level)***					
Variable	**N**	**Mean**	**SD**	**P25**	**P50**	**P75**
N Deleted	2,409,875	0.132	0.461	0.000	0.000	0.000
N Reviews	2,409,875	2.780	7.172	0.000	1.000	2.000
I(Any Deleted)	2,409,875	0.097	0.297	0.000	0.000	0.000
N Del Clean	2,409,875	0.111	0.413	0.000	0.000	0.000
N Del Flagged	2,409,875	0.021	0.144	0.000	0.000	0.000
Control Variables:						
Avg Firm Rating	2,409,875	3.574	0.680	3.158	3.632	4.027
Log Geo Complexity	2,409,875	1.357	0.452	1.099	1.099	1.386
Log N Roles	2,409,875	2.597	0.890	1.946	2.485	3.091
Log N Competitors	2,326,875	4.963	1.913	3.611	5.568	6.528
Avg Rating Competitors	2,326,875	3.678	0.242	3.512	3.686	3.838

Table 4 – Effect of DSA on Number of Deletions

This table examines the effect of the DSA on the number of reviews deleted per firm-month. *Treat* equals one for firms headquartered in an EU member state or with pre-DSA EU-based reviewers, and *Post* equals one for months after the DSA enforcement date (February 17, 2024). Column (1) includes no fixed effects, Column (2) includes country and year-month fixed effects, and Column (3) includes country-by-industry and industry-by-year-month fixed effects. Controls include average pre-DSA rating, geographic complexity, occupational breadth, competitive density, and leave-one-out competitor rating. The coefficient of interest is β, which captures the differential change in deletion volume for EU-exposed firms after DSA enforcement. All variables are defined in Appendix A. Standard errors clustered by country are in parentheses. ***, **, and * indicate significance at the 0.01, 0.05, and 0.10 levels, respectively.

	N Deleted		
	(1)	**(2)**	**(3)**
Treat × Post	**+0.0286*** **	**+0.0286*** **	**+0.0300*** **
	(0.0075)	(0.0075)	(0.0072)
Avg Firm Rating	−0.0172***	−0.0166***	−0.0182***
	(0.0008)	(0.0006)	(0.0008)
Log Geo Complexity	+0.1454***	+0.1425***	+0.1411***
	(0.0188)	(0.0180)	(0.0130)
Log N Roles	+0.1098***	+0.1104***	+0.1217***
	(0.0034)	(0.0035)	(0.0040)
Log N Competitors	+0.0110***	+0.0014	-
	(0.0009)	(0.0013)	-
Avg Rating Competitors	+0.0115**	+0.0179***	+0.0530***
	(0.0058)	(0.0062)	(0.0150)
Country FE		✓	
Year-Month FE		✓	
Country × Industry FE			✓
Industry × Year-Month FE			✓
Main Effects	✓	✓	✓
Controls	✓	✓	✓
Observations	2,326,875	2,326,875	2,326,875
Mean DV	0.132	0.132	0.132

Table 5 - Deletion Decomposition

This table decomposes the effect of the DSA on deletion volume by violation status. Panel A separates deleted reviews into clean reviews with no LLM-detected violation (Column 1) and flagged reviews with at least one violation (Column 2). Panel B separates flagged deletions into low-discretion violations (Column 1) and high-discretion violations (Column 2). Low-discretion and high-discretion categories are not mutually exclusive; a single review may contain violations of both types. All specifications include country-by-industry and industry-by-year-month fixed effects and the full set of pre-DSA firm-level controls. All variables are defined in Appendix A. Standard errors clustered by country are in parentheses. ***, **, and * indicate significance at the 0.01, 0.05, and 0.10 levels, respectively.

Panel A: Clean vs. Flagged

	N Deleted	
	No Violation (Clean)	**Flagged**
	(1)	**(2)**
Treat × Post	**+0.0267*** **	**+0.0027*** **
	(0.0062)	(0.0009)
Country × Industry FE	✓	✓
Industry × Year-Month FE	✓	✓
Main Effects	✓	✓
Controls	✓	✓
Observations	2,326,875	2,326,875
Mean DV	0.111	0.021

Panel B: Low-Discretion (Verifiable) vs. High-Discretion (Judgment-Based)

	N Deleted	
	Low Discretion	**High Discretion**
	(1)	**(2)**
Treat × Post	**+0.0005**	**+0.0032*** **
	(0.0005)	(0.0008)
Country × Industry FE	✓	✓
Industry × Year-Month FE	✓	✓
Main Effects	✓	✓
Controls	✓	✓
Observations	2,326,875	2,326,875
Mean DV	0.006	0.020

Table 6 - Decomposition of Deletion Increase by Review Characteristics

This table tests whether the DSA-induced deletion increase disproportionately targets reviews with negative content. Each column uses a different dependent variable: the count of deleted reviews in a given firm-month falling into a specific content category. Panel A compare reviews classified as harsh (any sub-rating equals 1) against reviews classified as universally perfect (all sub-ratings equal 5). Panel B compares reviews with extremely negative tone against reviews with extremely positive tone. Panel C compares cons-heavy reviews (where the cons text dominates) against pros-heavy reviews (where the pros text dominates). All specifications include country-by-industry and industry-by-year-month fixed effects and the full set of pre-DSA firm-level controls. All variables are defined in Appendix A. Standard errors clustered by country are in parentheses. ***, **, and * indicate significance at the 0.01, 0.05, and 0.10 levels, respectively.

Panel A: Review Harshness

	N Deleted	
	Any Harsh (Any dimension with score = 1)	**Universally Perfect** (All dimensions = 5)
	(1)	**(2)**
Treat × Post	**+0.0150*** **	**+0.0130**
	(0.0043)	(0.0079)
Observations	2,326,875	2,326,875
Mean DV	0.060	0.034

Panel B: Review Tone

	Extreme Neg Tone	**Extreme Pos Tone**
	(1)	**(2)**
Treat × Post	**+0.0026*** **	**+0.0024**
	(0.0009)	(0.0016)
Observations	2,326,875	2,326,875
Mean DV	0.010	0.008

Panel C: Content Balance

	Cons-Heavy	**Pros-Heavy**
	(1)	**(2)**
Treat × Post	**+0.0049****	**−0.0001**
	(0.0019)	(0.0019)
Observations	2,326,875	2,326,875
Mean DV	0.028	0.013

Table 7 - Cross-Sectional Heterogeneity by Firm Characteristics

This table examines whether the effect of the DSA on review deletion varies with firm characteristics associated with strategic incentives. Each column interacts Treat × Post with a different firm-level moderator: pre-DSA deletion activity (Column 1), local competitive density (Column 2), pre-DSA rating trend (Column 3), and pre-DSA review tone (Column 4). Pre-DSA deletion and competitive density are split at the sample median. Rating trend equals one if the firm's average monthly rating was declining over the pre-DSA window. Negative firm tone equals one if the firm's average pre-DSA sentiment score is below the sample median. All specifications include country-by-industry and industry-by-year-month fixed effects and the full set of pre-DSA firm-level controls. All variables are defined in Appendix A. Standard errors clustered by country are in parentheses. ***, **, and * indicate significance at the 0.01, 0.05, and 0.10 levels, respectively.

	N Deleted			
Moderator (Mod):	Pre-DSA Del.	Local Comp.	Rating Trend	Neg. Firm Tone
	(1)	**(2)**	**(3)**	**(4)**
Treat × Post	+0.0119***	+0.0195***	+0.0187***	+0.0215***
	(0.0035)	(0.0032)	(0.0055)	(0.0046)
Treat × Post × Mod	**+0.0439*** **	**+0.0311*** **	**+0.0191*** **	**+0.0168*** **
	(0.0083)	(0.0059)	(0.0046)	(0.0060)
Country × Industry FE	✓	✓	✓	✓
Industry × Year-Month FE	✓	✓	✓	✓
Main Effects	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Observations	2,326,875	2,326,875	2,326,875	2,326,875

Table 8 - Effect on Visible Review Volume and Displayed Ratings

This table examines whether the DSA affected the visible information environment on Glassdoor. Columns (1) and (2) use the total number of surviving reviews as the dependent variable. Columns (3) and (4) use the average displayed rating, defined as the cumulative mean of all surviving reviews. Columns (1) and (3) include country and year-month fixed effects; Columns (2) and (4) include country-by-industry and industry-by-year-month fixed effects. All specifications include the full set of pre-DSA firm-level controls. All variables are defined in Appendix A. Standard errors clustered by country are in parentheses. ***, **, and * indicate significance at the 0.01, 0.05, and 0.10 levels, respectively.

	N Reviews	N Reviews	Avg. Displayed Rating	Avg. Displayed Rating
	(1)	**(2)**	**(3)**	**(4)**
Treat × Post	**−0.0360**	**−0.0758**	**+0.0065*** **	**+0.0099*** **
	(0.1778)	(0.1930)	(0.0012)	(0.0018)
Country FE	✓		✓	
Year-Month FE	✓		✓	
Country × Industry FE		✓		✓
Industry × Year-Month FE		✓		✓
Main Effects	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Observations	2,326,875	2,326,875	2,137,212	2,137,212
Mean DV	2.780	2.780	3.618	3.618

Table 9 - Rating Informativeness

This table examines whether the association between visible Glassdoor ratings and subsequent labor market flows changed differentially for EU-exposed firms after the DSA. The dependent variable is *LaborFlow*, a labor market outcome measured at a six-month horizon (outflow rate in Column 1, inflow rate in Column 2, net flow in Column 3), *VisRating* is the average rating computed over surviving reviews. All columns include country-by-industry and industry x year-month fixed effects. Coefficients are scaled by 100 and represent percentage-point changes in labor flow rates. Labor flow data are sourced from Revelio Labs via WRDS. All variables are defined in Appendix A. Standard errors clustered by country are in parentheses. ***, **, and * indicate significance at the 0.01, 0.05, and 0.10 levels, respectively.

	Outflow	Inflow	Net Flow
	(1)	**(2)**	**(3)**
Vis × Post × Treat	**+4.60*** **	**+7.32*** **	**+2.27**
	(1.37)	(1.05)	(1.66)
Country × Industry FE	✓	✓	✓
Industry × Year-Month FE	✓	✓	✓
Main Effects	✓	✓	✓
Double Interactions	✓	✓	✓
Controls	✓	✓	✓
Observations	163,859	163,859	163,859
R^2	0.027	0.035	0.002

Table 10 - Decomposition of Labor Flows into External and Internal Components

This table decomposes the labor flow results from Table 9 into external flows (workers moving between firms) and internal flows (transfers within multi-establishment organizations). Columns (1) and (2) decompose outflows; Columns (3) and (4) decompose inflows. The specification follows Equation (2). Coefficients are scaled by 100 and represent percentage-point changes in labor flow rates. All variables are defined in Appendix A. Standard errors clustered by country are in parentheses. ***, **, and * indicate significance at the 0.01, 0.05, and 0.10 levels, respectively.

	Ext. Outflow	**Int. Outflow**	**Ext. Inflow**	**Int. Inflow**
	(1)	**(2)**	**(3)**	**(4)**
Vis × Post × Treat	**+2.07***	**+2.45*****	**+5.35*****	**+2.12*****
	(1.04)	(0.55)	(1.08)	(0.47)
Country × Industry FE	✓	✓	✓	✓
Industry × Year-Month FE	✓	✓	✓	✓
Main Effects	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Observations	163,859	163,859	163,859	163,859
R^2	0.018	0.022	0.029	0.022

Table 11 - Robustness: Alternative Specifications

This table reports the baseline deletion result alongside two alternative functional forms. Column (1) reproduces the baseline count specification from Table 4, Column (3). Column (2) replaces the count with an extensive-margin indicator equal to one if the firm deleted any review in the month. Column (3) uses a log specification, log(1 + N Deleted). All specifications include country-by-industry and industry-by-year-month fixed effects and the full set of pre-DSA firm-level controls. All variables are defined in Appendix A. Standard errors clustered by country are in parentheses. ***, **, and * indicate significance at the 0.01, 0.05, and 0.10 levels, respectively.

	N Deleted	**I(Any Del)**	**log(1+N Del)**
	(1)	**(2)**	**(3)**
Treat × Post	**+0.0300*** **	**+0.0074*** **	**+0.0083*** **
	(0.0072)	(0.0013)	(0.0023)
Country × Industry FE	✓	✓	✓
Industry × Year-Month FE	✓	✓	✓
Main Effects	✓	✓	✓
Controls	✓	✓	✓
Observations	2,326,875	2,326,875	2,326,875
Mean DV	0.132	0.097	0.080

ONLINE APPENDIX

Christensen, Lang, and Pinto

Online Appendix

Online Appendix Figures:

Figure OA1: Timeline of Digital Services Act
Figure OA2: Clean vs. Flagged Deletions

Online Appendix Tables:

Table OA1: LLM Classification Validation
Table OA2: Reviews and Deletions by Country (Top 10)
Table OA3: Distribution of Reviews and Deletions by Reviewer Role
Table OA4: Individual Violation Prevalence
Table OA5: Composition and Violation Patterns in Deleted Reviews
Table OA6: Determinants of Deletion
Table OA7: Additional Cross-Sectional Heterogeneity
Table OA8: Main Results with Firm FE

Figure OA1 - Timeline of Digital Services Act

. The DSA entered into force on November 16, 2022. For online platforms not designated as a VLOP or VLOSE, its obligations became applicable on February 17, 2024.

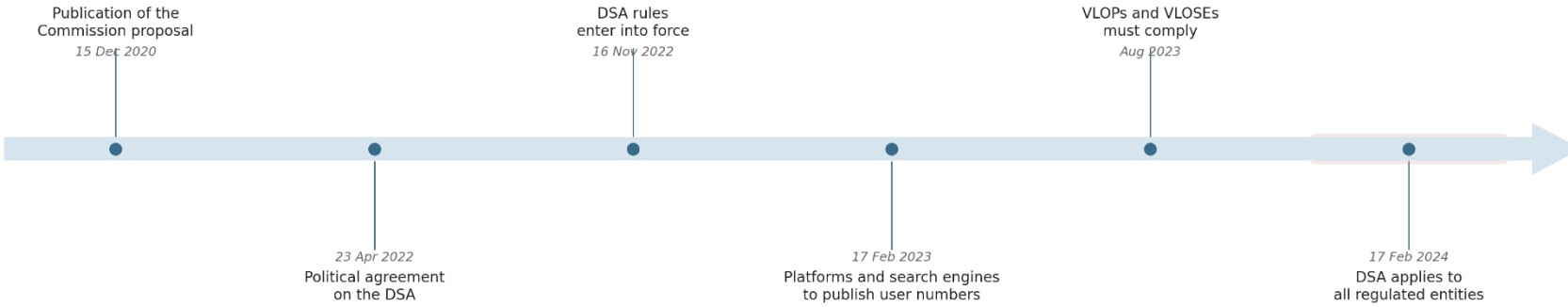
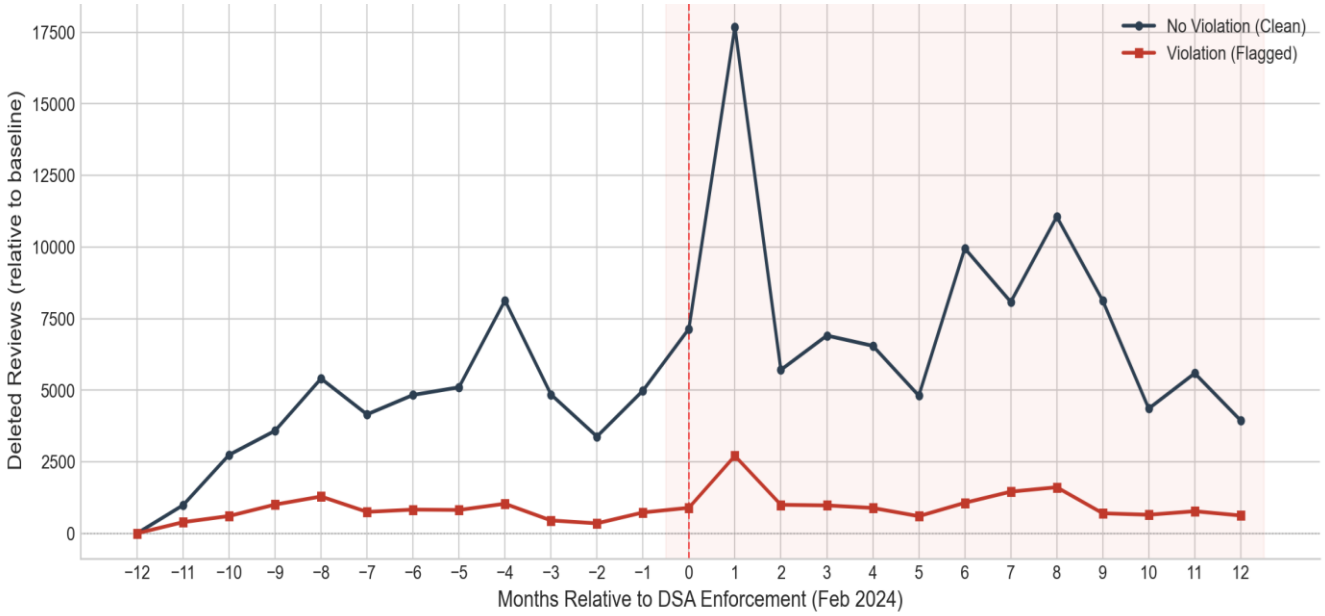


Figure OA2 - Clean vs. Flagged Deletions

Clean vs. Flagged Deletions



This figure plots monthly deletion volume relative to the pre-DSA average, separately for reviews with no LLM-detected content policy violation and reviews with at least one LLM-detected violation. The vertical dashed line indicates the DSA enforcement date (February 17, 2024). The sample includes all firms in the balanced regression panel observed in the 12-month window on each side of the enforcement date.

Table OA1 -LLM Classification Validation: GPT-5.4 Nano vs. Human Graders

This table reports validation statistics for the LLM violation classifier. A random sample of 500 reviews is independently evaluated by five human graders and GPT-5.4 Nano. A review is flagged as having a violation if at least one of five graders identifies one. Panel A reports overall classification agreement between the LLM and the human benchmark. Panel B stratifies LLM-human agreement by the number of graders who independently identified a violation. Panel C reports accuracy at alternative confidence score thresholds. All variables are defined in Appendix A.

Panel A: Classification Agreement

	Human Graders	**GPT-5.4 Nano**		
Reviews evaluated	500	500		
Flagged as violation	76 (15.2%)	108 (21.6%)		
	Both Flag	*Both Clean*	*Only LLM*	*Only Human*
Agreement breakdown	51	367	57	25
	(10.2%)	(73.4%)	(11.4%)	(5.0%)
Overall agreement rate		83.6%		

Panel B: LLM–Human Agreement by Grader Consensus

Human Grader Consensus	**Reviews**	**LLM Agrees**	**LLM Disagrees**	**Match Rate**
5 of 5 graders (unanimous)	14	14	0	100%
4 of 5 graders	7	7	0	100%
3 of 5 graders	10	6	4	60%
2 of 5 graders	16	11	5	69%
1 of 5 graders (single grader)	29	13	16	45%
Total	**76**	**51**	**25**	**67%**

Panel C: Confidence Score Calibration

Confidence Threshold	Reviews Flagged	Accuracy
Any flag (no threshold)	108	83.6%
≥ 70%	81	87.8%
≥ 80%	47	90.2%
≥ 90%	44	90.0%
≥ 95%	25	89.4%

Table OA2 – Reviews and Deletions by Country (Top 10)

This table reports the ten largest countries by review count in the regression sample. Country is the firm headquarters country from Revelio Labs. *Del. Rate* is the fraction of reviews deleted. The sample covers ±12 months around the DSA enforcement date, restricted to firms with at least one identified deletion.

Country (HQ)	N Reviews	N Deleted	Del. Rate	N Firms
United States	3,977,347	190,056	4.78%	43,865
United Kingdom	647,761	32,343	4.99%	8,627
India	585,162	30,573	5.22%	7,242
Brazil	546,773	22,146	4.05%	7,949
France	354,926	14,869	4.19%	2,842
Canada	277,239	12,609	4.55%	4,199
Germany	173,190	6,172	3.56%	1,618
Australia	127,701	6,610	5.18%	2,042
Singapore	109,867	6,506	5.92%	1,599
Ireland	106,273	3,174	2.99%	481

Table OA3 - Distribution of Reviews and Deletions by Reviewer Role

This table reports the distribution of reviews and deletions by reviewer role, allowing comparison of deletion rates across seniority levels. All variables are defined in Appendix A.

Role	N Reviews	N Deleted	Del. Rate
Retail Customer Advisor	609,053	47,898	7.86%
Software Engineering	314,900	15,987	5.08%
Intern	152,682	3,970	2.60%
Customer Care Representative	140,674	5,403	3.84%
Business Advisor	117,961	5,215	4.42%
Operations Supervisor	115,594	6,381	5.52%
Retail Sales Associate	112,655	3,160	2.81%
Project Manager	95,736	4,566	4.77%
unknown	98,933	12,928	13.07%
General Analyst	92,393	4,116	4.45%

Table OA4 - Individual Violation Prevalence

This table reports the prevalence of each of the eleven individual violation categories among surviving and deleted reviews, along with the lift ratio (prevalence in deleted divided by prevalence in surviving). Categories are grouped by discretion level: low-discretion violations (verifiable) and high-discretion violations (judgment-based). Prevalence values are rounded to two decimal places; lift ratios are computed from unrounded prevalence. All variables are defined in Appendix A.

Violation	Surv (%)	Del (%)	Lift
Fraudulent/Inauthentic	0.00	0.04	8.00×
Privacy/Naming	0.58	3.02	5.17×
Trolling/Inflammatory	0.34	1.55	4.50×
Proprietary Information	0.02	0.10	4.40×
Abuse/Profanity	2.11	8.12	3.84×
Violence/Threats	0.19	0.72	3.83×
Irrelevant/Off-topic	0.03	0.12	3.53×
Discrimination/Hate	1.03	3.53	3.43×
Spam/Promotional	0.06	0.15	2.59×
Salary/Compensation	1.42	2.40	1.69×
Low Quality/Unintelligible	1.12	1.20	1.07×

Table OA5 - Composition and Violation Patterns in Deleted Reviews

This table examines how the characteristics of deleted reviews differ for EU-exposed firms relative to non-EU-exposed firms. Panel A compares the mean overall rating and sub-ratings of deleted reviews for treated firms in the pre-DSA and post-DSA periods. Panel B compares the prevalence of LLM-detected violations among deleted reviews for EU-exposed and non-EU-exposed firms. All variables are defined in Appendix A.

Panel A: Mean Characteristics of Deleted Reviews (Treated Only)

	Pre-DSA	Post-DSA	Diff
Overall Rating	3.3860	3.3111	−0.0749
Comp & Benefits	3.3421	3.2673	−0.0748
Culture & Values	3.3208	3.2245	−0.0963
Senior Leadership	3.0561	2.9826	−0.0736
Career Opportunities	3.2453	3.1574	−0.0878
Work-Life Balance	3.3740	3.3126	−0.0614

Panel B: Violation Patterns (Deleted Only)

	Treated / EU-Exposed (n=160,757)	Control / Non-EU-Exposed (n=208,453)	Diff
Any Violation	11.89%	15.35%	−3.46 pp
Low-Discretion	2.87%	4.63%	−1.76 pp
High-Discretion	10.63%	13.47%	−2.83 pp

Table OA6 - Determinants of Deletion – Review-Level Analysis

This table examines which review characteristics predict deletion at the individual review level, and how those predictors shift after the DSA. The specification is a linear probability model estimated on a matched sample containing all deleted reviews in our window and an approximately equal number of randomly selected surviving reviews. Each row reports the triple interaction (Determinant × *Post* × *Treat*) from a specification that includes all determinants simultaneously, along with country-by-industry fixed effects. All variables are defined in Appendix A. Standard errors clustered by country are in parentheses. ***, **, and * indicate significance at the 0.01, 0.05, and 0.10 levels, respectively.

	I(Deleted)
	(1)
Low-Disc Violation × Post × Treat	-0.027***
	(0.0083)
High-Disc Violation × Post × Treat	0.013**
	(0.0053)
Any Harsh × Post × Treat	0.000
	(0.0160)
Extreme Neg Tone × Post × Treat	0.020**
	(0.0081)
Extreme Pos Tone × Post × Treat	0.016
	(0.0146)
Main Effects	✓
Det × Post, Det × Treat interactions	✓
Observations	702,694
Sample	*Classified deleted + surviving reviews*

Table OA7 - Additional Cross-Sectional Heterogeneity

This table reports additional cross-sectional heterogeneity tests for the main deletion result. Each column interacts *Post* × *Treat* with a different firm-level moderator, Firm Size is split at the pre-DSA median. The dependent variable is N Deleted. Public Sector equals one for firms classified under two-digit NAICS code 92 (Public Administration). All specifications include country-by-industry and industry-by-year-month fixed effects and the full set of firm-level controls. Standard errors clustered by country are in parentheses. ***, **, and * indicate significance at the 0.01, 0.05, and 0.10 levels, respectively.

	N Deleted	
Moderator:	**Firm Size**	**Public Sector**
	(1)	**(2)**
Post × Treat	+0.0018	+0.0303***
	(0.0037)	(0.0072)
Post × Treat × Mod	**+0.0396*** **	**−0.0138**
	(0.0105)	(0.0099)
Country × Industry FE	✓	✓
Industry × Year-Month FE	✓	✓
Observations	2,325,550	2,326,875

Table OA8 - Main Results with Firm Fixed Effects

This table replicates the baseline difference-in-differences specification from Table 4 using firm-level fixed effects. Column (1) includes firm and year-month fixed effects; Column (2) includes firm and industry-by-year-month fixed effects. Firm fixed effects absorb the treatment indicator and all time-invariant firm characteristics, so controls are omitted. Standard errors clustered by country are in parentheses. ***, **, and * indicate significance at the 0.01, 0.05, and 0.10 levels, respectively.

	N Deleted	
	(1)	**(2)**
Post × Treat	+0.0267***	+0.0279***
	(0.0075)	(0.0073)
Firm FE	✓	✓
Year-Month FE	✓	
Industry × Year-Month FE		✓
Observations	2,325,550	2,325,550