Text-as-Data Final Project

Grace Flandreau

Table of contents

1 Introduction

In September 2024 the Federal Trade Commission launched Operation AI Comply, a coordinated enforcement sweep targeting companies that exaggerate or misrepresent artificial intelligence capabilities in their marketing, a practice commonly called "AI-Washing." Cases brought under or alongside this initiative, against companies like DoNotPay, Rytr, Ascend, and Air AI, have continued into 2026, and commentary from several law firms and trade groups suggests this is now a sustained enforcement priority rather than a one-off sweep.

This paper takes an exploratory text-as-data approach to the issue of AI-Washing in FTC cases. Rather than testing a pre-specified hypothesis about AI language concentration across predefined categories, the analysis asks what substantive themes emerge organically from the complaint text itself, and where AI-related language falls within that structure. Using a corpus of 104 deduplicated FTC consumer protection complaints filed between September 2024 and May 2026, the full active period of Operation AI Comply to date, application of Latent Dirichlet Allocation topic modeling aims to surface whatever enforcement themes are present in the text without imposing categories in advance. This approach is motivated by the methods developed in Egami et al. (2022), who provide a general framework for treating text as data in descriptive inference including applications to consumer complaint corpora, and Bastani, Namavari, and Shaffer (2019), who demonstrate that automated topic modeling can recover substantively meaningful enforcement categories from consumer complaint text without manual coding. Elsayed (2026) provides a typological framework for AI-Washing language that informs post-hoc interpretation of the topics that emerge.

The central finding of the analysis is that AI-Washing complaints do not form an isolated topic cluster. Instead, they load heavily onto a broader topic characterized by deceptive opportunity and capability claims, a topic also occupied by traditional business opportunity fraud cases with no AI angle. This suggests the FTC is applying an established doctrinal framework for deceptive opportunity claims to AI-Washing conduct rather than treating it as a categorically novel enforcement problem. The remaining five topics recover recognizable enforcement priorities including credit reporting and equipment financing, automotive and real estate marketing, gig economy fee disputes, mergers and franchise disputes, and negative option and subscription trap practices, providing useful context for situating the AI-Washing finding within the full scope of agency activity during this period. The analysis proceeds by first describing the data collection methodology and corpus construction, followed by an explanation of the text preprocessing and LDA modeling approach,

and concluding with a presentation and interpretation of the six topics that emerged and what they reveal about the FTC's treatment of AI-Washing conduct during this period.

2 Data Introduction

In order to analyze the contents of different FTC cases, I chose to create a corpus of complaint documents spanning the time from when Operation AI Comply was created to the most recent full month before I started analysis. Using the FTC's Case Document Search tool, I was not able to filter for the data range, but I could filter for documents that included the word "complaint", to ensure that I was only downloading complaint documents (Commision (n.d.)). I decided to go with complaint documents for my corpus for a few key reasons. The first reason was that every FTC matter has only one complaint per case, as opposed to motions or orders which may have multiples in some cases and none in others. This reason also ensures that stopwords can be largely normalized across different topics, as words that show up in every document are likely to be words that can be filtered out of the final analysis. Another reason I chose to work with complaint documents was because these documents are where the FTC describes the alleged conduct of the company which is being deliberated on. Understanding the specific deceptive practices or violations a company is being investigated on is essential in analyzing what topics have been the most relevant since the creation of Operation AI Comply.

The FTC's Case Document Search tool does not provide a date range filter, so the date restriction was applied post-hoc in R rather than at the point of collection. To collect the data, I filtered for "complaint" in the document title and manually downloaded the twelve pages that contained my desired date range. Since the interface is JavaScript-driven and does not support URL-based query parameters, automated scraping of search results was not feasible directly; instead, the search results

pages were saved manually as local HTML files and parsed in R. The twelve pages gathered from the FTC website were then parsed in R using the rvest package, which extracted the document title, case name, filing date, and PDF link for each result. The resulting table was filtered to retain only rows whose document title begins with "Complaint," "Amended Complaint," "Second Amended Complaint," "Third Amended Complaint," or "Final Complaint," removing false positives such as "Order Granting Complaint Counsel's Motion" that matched the initial keyword search but were not complaint documents. Next, I parsed the filing dates through lubridate and filtered to the September 2024 to May 2026 window, and relative PDF URLs were resolved to full addresses. The 183 qualifying complaint PDFs were downloaded using the curl package with browser-style request headers, and complaint text was extracted from each PDF using pdftools::pdf_text(), collapsing all pages into a single character string per document.

At this stage, the initial corpus of 183 complaints contained multiple filings for some matters, most significantly the Caremark Rx matter, which appeared more than 30 times across different defendant specific complaints filed within the same consolidated case. Retaining all filings would have allowed a single matter to dominate the corpus and distort the word co-occurrence patterns that LDA depends on. To address this, I deduplicated the corpus by retaining only the most recent complaint per unique case name, using dplyr::slice_max() on the parsed filing date. This reduced the corpus to 104 unique cases spanning the full collection window.

3 Analysis

3.1 Methods

Once I finalized my corpus of 104 complaint documents, it was important to create a stoplist of words relevant to the documents I was working with. While the stopword list included in tidytext was helpful, there were a large number of other legal words that would misrepresent the intended analysis if not removed. For my "legal stopword" list, I started by filtering for legal words. I built a custom stoplist as a tibble for use with anti_join(), supplementing the standard stop_words lexicon. Words in this first list included examples such as "defendant" and "judgement," words that would be common in a legal setting. To add on to this, I also chose to include a list of roman numerals into my legal stopword list, as many of the complaint documents include these in the formatting, and roman numerals do not count as numbers to R. In an upcoming section of my analysis, I ran into the issue of having random strings as well as general words that were showing up for nearly every topic. To remedy this, I first tried to use filter(str_detect(word, "[^1]+$")) to filter for only words in the English language. However, this didn't work, so I ended up printing out the top 10 words from every topic and adding odd character strings like "wkdw" and "zh" to my legal_stopwords lexicon. I also did the same thing with words that came up in multiple topics that I noticed were likely related to the formatting of FTC complaints as a whole. Examples of this include words like "document" and "business." One possible limitation that I noticed from this section was figuring out which words should be viewed as stop words, and which are important to different topics, but show up a lot. In adding words to the legal_stopwords vector, I did try to think through examples where a word may or may not be important to the topic, but I acknowledge that I may have made some over or undersimplifications in this process.

Similarly, an AI-term reference dictionary of 32 words was constructed drawing on Elsayed (2026)**'s**

typological framework. These terms covered direct technical references ("algorithm," "machine learning," "LLM") and broader capability-signaling vocabulary ("automated," "intelligent," "powered"). This dictionary was not used to filter or preprocess the corpus before modeling; the topic structure was allowed to emerge entirely from unsupervised word co-occurrence patterns. For each complaint, an AI-term score was computed by joining the cleaned token table against the dictionary and counting matching tokens per document. This score was used in two ways: to identify complaints with unusually high AI-term density, defined as 50 or more AI-related terms based on a natural break visible in the distribution of scores across the corpus, and to examine whether confirmed Operation AI Comply cases showed systematically higher AI-term counts than other complaints and whether their LDA topic assignments aligned with the AI-heavy topic surfaced by the model.

After creating these term lists, I moved on to starting my analysis using Tidytext and LDA. The raw complaint text was converted into an analyzable format using the tidytext package. Unnest_tokens() broke each complaint into one word per row, and stop words were removed in two passes using anti_join(): first against the standard stop_words lexicon, then against a custom legal_stopwords tibble targeting FTC-specific boilerplate. Three additional filters addressed PDF extraction noise, purely numeric tokens were removed to clear out page numbers, tokens shorter than three characters were dropped to eliminate extraction fragments, and tokens containing non-alphabetic characters were removed to address garbled sequences produced by font encoding issues in a subset of PDFs. The cleaned token table was aggregated into word counts per document using count() and cast into a DocumentTermMatrix using cast_dtm(), producing the sparse matrix format required by the LDA model.

A Latent Dirichlet Allocation model was fit using topicmodels::LDA(), an unsupervised method that surfaces latent topics purely from word co-occurrence patterns across documents without requiring

pre-assigned category labels. The number of topics k was selected through iterative testing. I found that k=4 produced topics too broad to be interpretable, while k=8 introduced redundancy, splitting what appeared to be one coherent theme into two nearly identical topics. k=6 was selected as the most interpretable solution, producing six cleanly distinguishable topics with minimal overlap. A fixed random seed ensures the results are fully reproducible across sessions. Two matrices were extracted from the fitted model using tidy(): the beta matrix of word-topic probabilities, used to characterize and label each topic by its most distinctive vocabulary, and the gamma matrix of document-topic probabilities, used to identify which complaints are most strongly associated with each topic and to examine how topic membership varies across complaint types.

3.2 Results

After running the LDA model for k=6, The six topics that emerged were: (1) credit reporting and equipment financing, (2) automotive, real estate, and consumer product marketing, (3) gig economy and unfair fee practices, (4) deceptive business opportunity and AI-Washing claims, (5) mergers, acquisitions, and franchise disputes, and (6) negative option and subscription trap practices. These labels were assigned by reading the top 10 highest-beta terms for each topic as a group and identifying what kind of FTC enforcement case would produce that vocabulary, cross-referenced against the known cases in the corpus. I came to the understanding that these were what the six topics represented by using Figure 1 to determine which words were the most highly associated with each topic. Something I noticed about the bar chart was that there is a combination of company names and topic terms that allude to the content of each topic. For example, Topic 4 includes the terms “ascend” and “income,” where the first word is likely in reference to the Ascend Ecom FTC case, and income likely relates to earnings claims central to business opportunity fraud. In conjunction with my own observations, I also shared this figure with ClaudeAI and asked the LLM

to identify what the six topics represent. Using this tool in my research allowed me to take advantage of the way LLM's use word prediction to categorize different terms. Together the six themes map reasonably well onto the known landscape of FTC enforcement during this period. Negative option practices, data broker regulation, merger review, and gig economy fee disputes are all well-documented agency priorities, and their emergence as distinct topics without any pre-labeling suggests the LDA is recovering real structural variation in the corpus rather than noise.

In an effort to compare the topics created by the LDA model and AI cases from the FTC, I chose to manually go through the FTC's page of reports and filter for "Artificial Intelligence." The goal of this was to create a list of confirmed AI related cases so that I could compare the true cases with the cases that had the highest AI-term counts. This list was manually created based on the last few months, so it is important to recognize the limitation that user error may be present (Commission (n.d.)). Based on these findings, I compiled a table with 13 AI cases within the time period of the study. Interestingly, there were three cases that showed up in the "Artificial Intelligence" tab that were not present in the case_name vector, these being NGL Labs, FBA Machine, and Workado LLC. There are a few reasons these may have been filtered out of the corpus, including falling right outside of the set date range or not having the right complaint document title. Despite the absences, the selection of 13 other cases is robust enough to check against other cases in the corpus.

For the first check, I created a stacked bar chart to represent the topics represented in AI-Washing cases compared to a random sample of other cases in the corpus. I decided to do a stacked bar chart so that it would be easy to visualize the differences found from topic modeling (Figure 2). The stacked bar chart compares topic composition for five confirmed Operation AI Comply cases against five comparison cases randomly selected from the rest of the sample, with gamma values indicating the proportion of each complaint attributable to each topic. The most striking find is that nearly all of the randomly selected AI-Washing cases load overwhelmingly onto Topic 4 (purple), with Cleo

AI, Growth Cave LLC, accessiBe Inc., and Empire Holdings Group LLC showing near-complete Topic 4 composition. In the random sample of comparison cases, none of the cases have a high representation of Topic 4. Based on this visualization, it appears as if AI language is incorporated into the LDA model's categorization of Topic 4. This conclusion is further supported in Figure 3, where I used a heatmap of the 13 confirmed AI cases to show the representation of each topic. Based on the heatmap, 10 of the 13 cases have at least a partial representation of Topic 4, with 8 of them having a representation of over 70%, making up the majority of the sample. In contrast, the other topics with high representations are topics 2, 3, and 6. Even accounting for the cases that do not load predominantly onto Topic 4, the proportion that do is large enough to support the conclusion that Topic 4 captures the enforcement language most closely associated with AI-washing conduct in this corpus.

In the final analysis test on my Topic 4 findings, I chose to create two tables (Tables 1 and 2). One of the tables simply selects the cases with the most AI-terms, based on the previously created ai_terms tibble. The other table uses the list of the 13 confirmed AI cases. For each case, I included the associated AI term count and dominant topic. The four complaints with the highest AI-term counts were all confirmed Operation AI Comply cases, though a number of non-AI cases in the corpus also recorded higher term counts than several of the confirmed AI-washing complaints, reflecting the limitations of dictionary-based scoring discussed in the methods section. Another notable finding in these tables is that Topic 4 only shows up in confirmed AI cases. One limitation here, and for Figure 2, is that this could be due to the random sample taken, amplifying the representation of Topic 4 in the AI cases and downplaying it for non-confirmed AI cases. Despite this limitation, the concentration of confirmed AI-washing cases in Topic 4 is substantial enough to support the conclusion that Topic 4's vocabulary reflects the enforcement language central to Operation AI Comply.

The most important implication is what Topic 4 reveals about where AI-Washing sits within the broader FTC enforcement landscape. Rather than forming an entirely isolated cluster, AI-Washing cases share Topic 4 with traditional business opportunity fraud cases involving fake earnings promises, inflated capability pitches, and misleading income claims that predate AI entirely. The top terms driving Topic 4, earnings, claims, income, growth, opportunity, cave, ascend, capture both the AI-Washing cases (Growth Cave, Ascend Ecom) and non-AI business opportunity fraud cases that share the same rhetorical structure of promising outsized returns through a supposedly superior system or technology. This suggests the FTC is applying an established deceptive opportunity framework to AI-Washing conduct rather than treating it as a categorically new enforcement problem, which is a substantively interesting finding about how the agency has absorbed AI-related complaints into its existing doctrinal toolkit. The fact that the LDA surfaced this pattern without any prior category assignment makes it a stronger finding than it would have been had we imposed it in advance.

4 Conclusion

This paper set out to examine where AI-Washing fits within the broader landscape of FTC consumer protection enforcement during the active period of Operation AI Comply, using an exploratory text-as-data approach rather than a pre-specified hypothesis. Applying Latent Dirichlet Allocation topic modeling to a corpus of 104 deduplicated FTC complaint documents filed between September 2024 and May 2026, the analysis surfaced six substantive enforcement themes without imposing any category labels in advance. The central finding is that AI-Washing complaints do not constitute a linguistically isolated enforcement category. Instead, they cluster predominantly within Topic 4, a topic characterized by the vocabulary of deceptive business opportunity and capability claims, a

theme that also encompasses traditional business opportunity fraud cases with no AI angle. The co-occurrence of confirmed Operation AI Comply cases alongside non-AI business opportunity fraud cases in the same topic cluster suggests the FTC is applying a familiar doctrinal framework to AI-Washing conduct rather than developing a novel enforcement theory from scratch.

Several limitations deserve acknowledgment. The corpus is restricted to complaint documents and does not capture how cases were ultimately resolved. The dictionary-based AI-term scoring approach under-counted cases where AI-Washing conduct is described in more technical or oblique legal language, as evidenced by the low AI-term counts for confirmed cases like Evolv Technologies and Intellivision. The legal stopword list required refinement, and some over or under removal of substantively meaningful terms is possible.

Future work could extend the corpus backward to the period before Operation AI Comply to test whether the Topic 4 cluster existed in FTC enforcement language prior to 2024, and could expand to include consent orders alongside complaints to compare how the agency describes alleged conduct at the outset versus how enforcement is ultimately resolved. I would also be interested in seeing how FTC topics created by LDA or other methods will evolve in the future. With AI becoming a constantly growing technology, there will likely be an expansion as to how it is approached by the FTC as AI becomes more accurate and less detectable for the general population. In the years ahead, as AI capabilities become more sophisticated and harder for consumers to evaluate, deceptive AI marketing claims are likely to remain a persistent enforcement concern for the FTC.

5 Figures and Tables

Figure 1: Top 10 terms per topic by beta-weighted word-topic probability. Terms are ordered within each facet by beta value. Topics are colored by Set2 palette and labeled numerically.

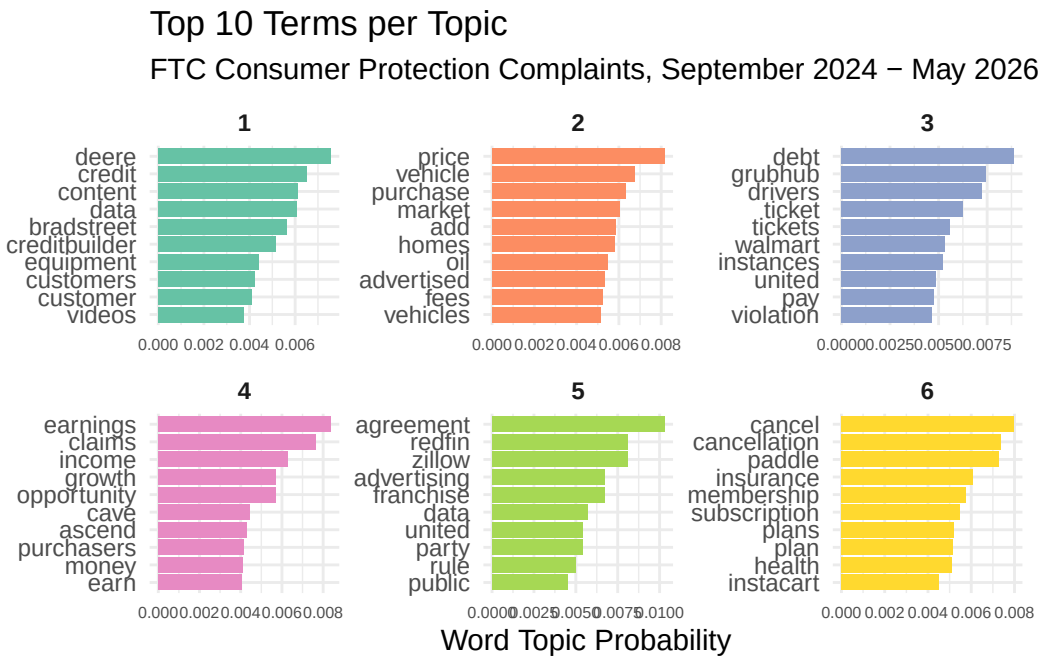


Figure 2: Topic composition for selected FTC complaints by document-topic probability (gamma). Cases are split into AI-washing and comparison groups. Each bar represents one complaint; colored segments show the proportion of that complaint attributed to each topic.

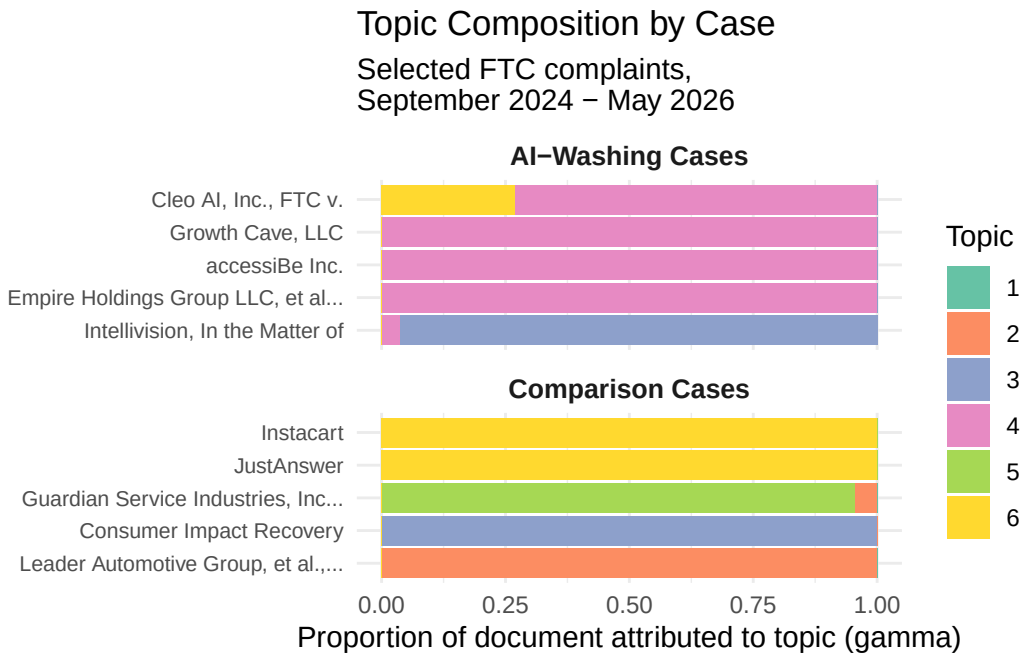
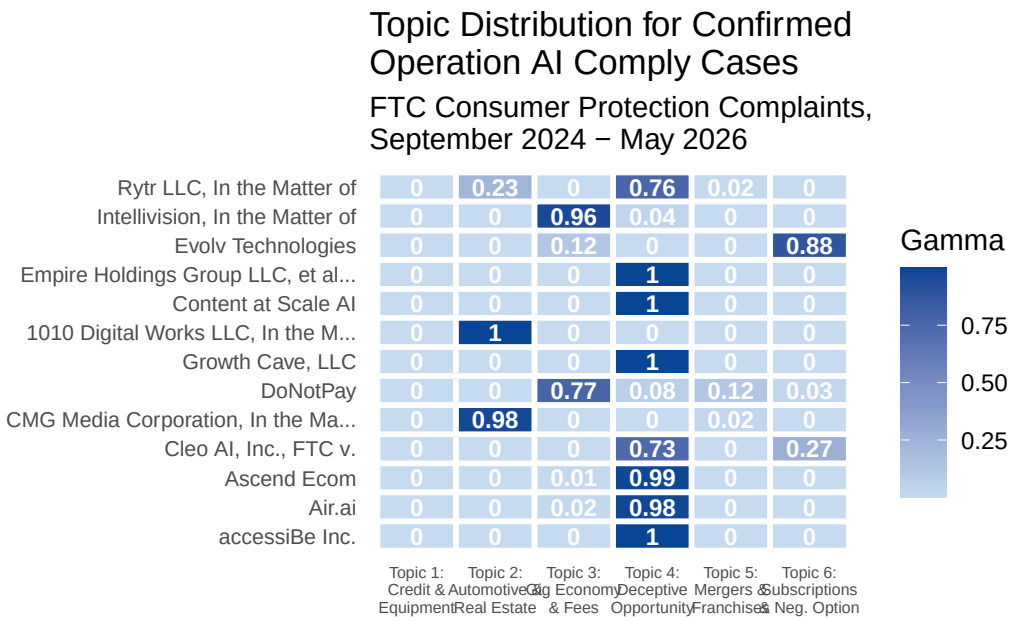


Figure 3: Document-topic probability (gamma) for confirmed Operation AI Comply cases across all six LDA topics. Darker blue indicates stronger topic association. Cases are ordered by AI-related term count.



```
# A tibble: 1 x 7
    min median  mean   max above_20 above_30 above_50
  <int>  <int> <dbl> <int>    <int>    <int>    <int>
1     1     10  16.3    67       33       16        4
```

Table 1: High AI-Term Complaints (30+ AI-related terms)

Case Name	AI Term Count	Dominant Topic
Growth Cave, LLC	67	4
Content at Scale AI	62	4
Xponential Fitness	52	5
Empire Holdings Group LLC, et al. FTC v.	51	4
Ascend Ecom	47	4
Qargo Coffee, Inc., et al., FTC v.	43	5
Disney	42	5
Cognosphere, LLC, U.S. v.	40	5
General Motors LLC., et al., In the Matter of	38	5
Pornhub/Mindgeek/Aylo	38	1
Leader Automotive Group, et al., FTC and State of Illinois v.	35	2
StubHub Holdings, FTC v.	35	3
Apitor	33	5
Iconic Hearts Holdings, Inc., U.S. v.	33	6
Zillow Group/Redfin Corp.	32	5
Deere & Company, FTC v.	31	1

Table 2: Confimed AI Cases, AI-Term Frequency and Dominant LDA Topic

Case Name	AI Term Count	Dominant Topic
Growth Cave, LLC	67	4
Content at Scale AI	62	4
Empire Holdings Group LLC, et al. FTC v.	51	4
Ascend Ecom	47	4
Air.ai	29	4
accessiBe Inc.	25	4
DoNotPay	23	3
Rytr LLC, In the Matter of	19	4
CMG Media Corporation, In the Matter of	13	2
Intellivision, In the Matter of	9	3
1010 Digital Works LLC, In the Matter of	7	2
Cleo AI, Inc., FTC v.	4	4
Evolv Technologies	1	6

6 LLM Statement

ClaudeAI was used as an assistive tool at three points in this analysis. First, during data collection, Claude was consulted to identify an approach for parsing JavaScript-rendered search results pages that could not be scraped directly via URL, resulting in the manual HTML saving and rvest parsing workflow described in the data section. Second, during preprocessing, Claude assisted in generating an initial set of candidate terms for the legal stopword list, which was subsequently

reviewed, modified, and expanded by me based on iterative inspection of topic model output. Third, during topic interpretation, Claude was consulted as a secondary check of my own reading of the top beta-weighted terms per topic; the final topic labels reflect the author's independent judgment and were not adopted directly from Claude's suggestions. All analytical decisions, including stopword selection, the choice of k=6, threshold selection for AI-term scoring, and the interpretation of results, were made by the author. Code was written in R using tidytext, topicmodels, ggplot2, and related packages; Claude assisted in debugging and code structure but did not write the final analytical pipeline independently.

References

Bastani, Kaveh, Hamed Namavari, and Jeffry Shaffer. 2019. "Latent Dirichlet Allocation (LDA) for Topic Modeling of the CFPB Consumer Complaints." *ScienceDirect* 127: 256–71. https://www.sciencedirect.com/science/article/pii/S095741741930154X.

Commision, Federal Trade. "Case Document Search." https://www.ftc.gov/legal-library/browse/cases-proceedings/case-document-search?search=&search_title=.

Commission, Federal Trade. "Technology/Artificial Intelligence." https://www.ftc.gov/industry/technology/artificial-intelligence?page=0.

Egami, Naoki et al. 2022. "How to Make Causal Inferences Using Texts." 8(42).

Elsayed, Nelly. 2026. "AI Washing and the Erosion of Digital Legitimacy: A Socio-Technical

Perspective on Responsible Artificial Intelligence in Business."