Disaggregating Narrative Disclosures into Accounting Topics

Jenna J. Burke
Associate Professor
University of Colorado Denver
jenna.j.burke@ucdenver.edu

Rani Hoitash
John E. Rhodes Professor of Accountancy
Bentley University
rhoitash@bentley.edu

Udi Hoitash
Lilian L. and Harry A. Cowan Professor of Accounting
Northeastern University
u.hoitash@northeastern.edu

Summer (Xia) Xiao
Assistant Professor
Northeastern University
x.xiao@northeastern.edu

August 2025

Acknowledgments: We thank Amazon AI finance group, participants at the 2024 Hawai'i Accounting Research Conference and workshops at Bentley University, Northeastern University, and the University of Colorado Denver for their helpful comments. We also thank our research assistants Sai Chandra, Akshay Nagireddy, and Mukul Rawat from Northeastern University.

Data Availability: All data are available from the public sources cited in the text.

JEL Classifications: C45; C81; M40; M41.

Keywords: Disaggregation; Small language model; Llama-2; accounting; topic classification; textual analysis; MD&A

Disaggregating Narrative Disclosures into Accounting Topics

Abstract

Researchers often examine narrative text in aggregate, but this approach can obscure economically important variation across specific accounting topics. Existing approaches to identifying topics covered in disclosures, such as reliance on commercial data aggregators, manual classification, or early forms of topic modeling, each have limitations that can hinder timely, resource-efficient, and accurate investigation. To address these limitations, we develop a domain-specific small language model that embeds accounting knowledge from authoritative reporting standards and classifies accounting topics efficiently in large volumes of unstructured narrative financial disclosures. We apply the model to MD&A text and show that topic-specific measures of disclosure content reveal insights missed by aggregate analyses. Topic-level variation predicts regulatory scrutiny (SEC comment letters), financial misstatements, and market reaction, relationships that are obscured when disclosures are studied only in the aggregate. Our study demonstrates the importance of disaggregating narrative disclosures into accounting topics and provides a practical tool for researchers seeking to better understand distinct areas of financial reporting.

Keywords: Disaggregation; Small language model; Llama-2; accounting; topic classification; textual analysis; MD&A

Disaggregating Narrative Disclosures into Accounting Topics

Introduction

Disclosure research often examines narrative text in aggregate, finding, for example, that investors react to the overall tone of financial disclosures. However, studying disclosures in this way can mask important and economically significant variation across specific topics, such as whether managers use a pessimistic tone in areas like revenue, tax, or contingencies. To move beyond these limitations and better understand narrative financial disclosures, we leverage innovations in generative artificial intelligence (AI) to build a domain-specific small language model (SLM) for classifying accounting topics within unstructured financial disclosures.[1] We then use this model to study the Management's Discussion and Analysis (MD&A) section of the 10-K, examining how topic-level coverage relates to regulatory scrutiny, financial reporting issues, and market reactions. These relationships are obscured when disclosures are analyzed only at the aggregate level.

Recognizing the richness of more granular information, there is growing interest in topic-specific research in accounting, where topics such as revenue, tax, and mergers and acquisitions are identified from various textual data (e.g., 10-Ks, 8-Ks, SEC comment letters, etc.). This line of work finds that specific accounting topics are important predictors of material outcomes, such as related corporate policies, financial reporting quality, and disclosures (e.g., Ahn et al. 2020; Bens et al. 2016; Christensen et al. 2014; Dechow et al. 2016; Drake et al. 2024; Dyer et al. 2017; Inger et al. 2018; Kubick et al. 2016). These studies typically use data classified by commercial data

[1]We focus on the classification of accounting topics, as defined by the FASB Accounting Standards Codification. While rooted in authoritative accounting standards, the relevance of these topics extends well beyond accounting. Discussions of accounting topics, such as revenue, fair value, and business combinations, reflect core economic activities that are studied in various research domains, including financial valuation, as evidenced by our market reaction tests.

aggregators, manual classification, or earlier versions of topic modeling to identify topics, which each have limitations that can hinder the timely, resource-efficient, and accurate investigation of economically important disclosures. For example, Audit Analytics is a popular data aggregator that identifies the topic of financial misstatements, material weaknesses, SEC comment letters, and critical audit matter disclosures. To utilize data aggregators, researchers must have access to the data and occasionally additional modules, must use the topics the aggregator provides even if additional granularity is desired, and must wait until the data is classified and released. Further, many disclosures (e.g., sections of 10-K filings) are not collected or are not classified into topics by data aggregators, so researchers must rely on manual classification. Manual classification presents its own challenges, where researchers must dedicate significant time to the task and will often need to limit the volume of disclosures examined (de Kok 2025). Lastly, early versions of topic modeling primarily focus on topic discovery, which is different than our objective to identify *specific* accounting topics.[2]

Analyzing accounting topics at scale requires a model that is efficient, accessible, and accurate, which is particularly challenging considering the abundance of unclassified textual data available in financial disclosures. We address this challenge by developing a domain-specific SLM, fine-tuned on Meta's Llama-2-7B, that classifies accounting topics in disclosures.[3] Meta's

[2] We therefore complement, rather than compete, with methods of topic modeling intended for topic discovery. For example, several studies use LDA (introduced by Blei et al. 2003) to discover groups of related words within disclosures (e.g., Bao and Datta 2014; Brown et al. 2020; Brown et al. 2024; Hoberg and Lewis 2017; Huang et al. 2018; Ryans 2021). LDA is useful for discovering topics based on word frequency and when labeled data is not available, but not ideal when the researcher is interested in specific topics such as revenue and inventory, as there is no guarantee these topics will be discovered or distinguishable from one another.

[3] Since we began this project, newer language models such as GPT-5 and Llama-3/4 have been released. Their availability does not diminish the contribution of our study. Our model is fine-tuned on XBRL-labeled data linked to specific accounting topics drawn from authoritative reporting standards, producing a domain-specific tool tailored to financial disclosures, whereas general purpose models (even more advanced ones) do not provide this specialization out of the box. In addition, large closed-source models are not free to use and can be costly at scale, making them less practical for academic research. Thus, while future work may adapt newer models, the approach and findings in this paper remain relevant and provide a foundation.

Llama-2-7B provides a strong foundation for our model because it is smaller, computationally efficient (can be used on a personal computer), and open source, meaning that it can be freely modified, replicated, and shared. This allows us to embed domain-specific financial reporting knowledge and make the resulting model available for future research.

To fine-tune the model, we leverage Extensible Business Reporting Language (XBRL). Financial statement notes filed in XBRL are "tagged" to specific concepts in the FASB Accounting Standards Codification, which ensures that the accounting topics we identify are important areas of financial reporting. XBRL, therefore, creates a unique opportunity to use a large sample of objectively labeled data to fine-tune Llama with domain-specific knowledge.[4] We identify 470,238 financial statement notes in 10-K filings for public companies from 2009 to 2021. We use the most frequent standard taxonomy XBRL tags, which map into 32 distinct topics in the FASB Codification.[5] We split this data into a training/evaluation sample (357,058 notes, of which 192,000 are ultimately used for fine-tuning) and an out-of-sample testing sample (113,180 notes).[6] During fine-tuning, the model learns to classify the 32 accounting topics, even where distinctions are subtle (e.g., separating impairment discussions across goodwill, PPE, and restructuring). In testing, the model achieves weighted average precision, recall, and F1 scores of 95.54, 95.48, and 95.47 percent, respectively. Importantly, because of its understanding of language and context, the

---

[4] LLMs' lack of domain-specific knowledge can be a limitation that, if not overcome by providing explicit examples, can cause them to hallucinate (de Kok 2025). Most research in supervised machine learning manually assigns labels to training data, which can be costly and subjective. For example, Huang et al. (2023) use a dataset of sentences manually labeled by researchers into positive, negative, and neutral sentiments. The use of XBRL to identify topics for supervised fine-tuning overcomes these challenges.

[5] Appendix A includes the mapping of the most frequent XBRL tags to accounting topics in the FASB Accounting Standards Codification. For example, the XBRL tag "FairValueDisclosuresTextBlock" maps to the Fair Value topic. 85 distinct XBRL tags map into these 32 topics, and together account for 92.5 percent of all notes with a taxonomy tag. To remain objective and follow the separate topics in the Codification, our topic classification is fairly granular. For example, there are separate topics for revenue and deferred revenue. Researchers can combine topics as needed for their research question.

[6] Each sample contains distinct companies. This is important because note disclosures are sticky and tend to repeat year over year. Therefore, using the same firm in the training, evaluation, and testing phases may provide an unfair advantage to our model.

model performs well even for topics that would be difficult to distinguish through other methods. For example, both PPE and goodwill and intangibles have high F1 scores at 98.14 and 98.91 percent, respectively. To ensure the conceptual advantages of fine-tuned Llama-2-7B for this task translate to actual performance advantages, we test different LLMs in additional analysis.

Having established the model's performance, we apply it to the MD&A, a critical narrative disclosure that contains meaningful topic-specific information that is difficult to parse at scale. Past research has typically examined the MD&A in aggregate (e.g., overall tone) or used machine learning methods to *discover* topics. Our model enables researchers to move beyond these approaches and quantify topic-specific content, allowing new questions to be asked about what is disclosed and how. We show our model the text of 3,937,612 MD&A paragraphs, and our model classifies 40.95 percent of these paragraphs as relating to one of the 32 accounting topics. The most common topics are tax, debt, equity, commitments and contingencies, and revenue. These topics represent economically important accounts to investors, other decision-makers, and researchers. Using these classifications, we study whether the extent and tone of topic-specific discussions predict SEC comment letters, financial misstatements, and market reactions around 10-K filings. This analysis combines our classification of accounting topics (what is being disclosed) with literature that examines textual measures of narrative disclosure (how it is being disclosed).

Our analyses deliver several insights. First, with one exception, both the extent and pessimistic tone of topic-specific discussions are significantly associated with the receipt of SEC comment letters addressing the same topic. Second, we find that a more extensive discussion of commitments and contingencies, along with the extent and pessimistic tone of revenue discussions, is associated with a higher likelihood of related misstatements. These findings demonstrate that topic-specific disclosures have important consequences for both regulatory oversight and financial reporting quality in related areas. Third, the pessimistic tone of commitments and contingencies discussions is associated with a negative market reaction, while other topics are not significant, reflecting investors' assessments of the economic implications of topic-specific disclosures. Taken together, these results demonstrate the importance of topic-specific analysis. To further demonstrate this, we contrast our topic-specific findings with aggregate measures. We find that the overall length and tone of the MD&A are not significantly associated with the likelihood of comment letter receipt or misstatement, which is inconsistent with the significant topic-level results and highlights the importance of examining this question at a more granular level. For market reaction, we find that the overall pessimistic tone of the MD&A is significantly associated with a negative market reaction, whereas our topic-specific findings reveal that specific accounting topics drive these overall findings. Overall, these findings provide evidence that our model captures disclosure content that is relevant to regulatory scrutiny, reporting quality, and investor response.[7]

Our study contributes to disclosure literature by showing that aggregate textual analysis can obscure meaningful relationships, whereas topic-level analysis provides sharper insights into regulatory scrutiny, financial reporting quality, and market response. We develop a scalable solution for identifying accounting topics in unstructured disclosures. Specifically, by fine-tuning an open-source SLM on XBRL-labeled data, we create a model that classifies accounting topics with high accuracy and enables researchers to study disclosure content at the topic level. This

[7] In additional analysis, we apply our model to custom notes to the financial statements (i.e., notes where a custom extended XBRL tag is used instead of a standard taxonomy tag). 54.1 percent of firm-year observations use at least one of these custom notes. Our classifications show that the most frequently customized topics are equity, debt, investments, commitments and contingencies, and receivables. Multivariate results indicate that using a custom tag for these topics is associated with a higher likelihood of related SEC comment letters, suggesting that topic-level classification can provide new insight into how regulators respond to customized disclosures.

approach offers a practical tool for extracting accounting topics from unstructured narrative text at scale.[8,9] Unlike prior work using topic discovery techniques, our model is explicitly grounded in authoritative accounting standards through XBRL-labeled training data. This tie to the FASB Codification ensures that topics correspond directly to important areas of financial reporting. Beyond methodological innovation, we show that topic-level disclosures and associated measures are economically important, as they shed new light on how managers communicate about distinct areas, how regulators respond, and how markets process this information.[10]

Background

Topic-specific disclosure and related research in accounting

Accounting researchers are increasingly focusing on specific topics, such as revenue, tax, or mergers and acquisitions, that are covered in broader disclosures. This literature shows that disaggregated or topic-level information enhances predictive power for material outcomes. For instance, studies find that specific topics of SEC comment letters lead to topic-specific regulatory and reporting consequences (Bens et al. 2016; Dechow et al. 2016; Kubick et al. 2016). Likewise, auditor expertise in specific accounting areas is associated with financial reporting quality and other outcomes in those areas (Ahn et al. 2020; Christensen et al. 2014; McGuire et al. 2012). Past research also finds that when auditors highlight certain topics in their disclosures, managers revise their disclosures in these areas (Burke et al. 2023; Drake et al. 2024), and the disclosures provide

[8] Our study differs from much of the academic research on generative AI to date, which often relies on powerful, computationally intensive, closed-source models such as GPT-4. Alongside a few concurrent working papers, we are one of the first in finance and accounting to fine-tune open-source generative LLMs/SLMs (Bernard et al. 2024; Konstantinidis et al. 2024; Luo et al. 2024; Pavlyshenko 2023).
[9] We will provide our code, data, and fine-tuned model on an open-source repository after publication.
[10] We also contribute to XBRL research. Specifically, by leveraging XBRL labeled data to fine-tune our model, we can identify specific topics across countless observations without any manual labeling. With this innovation, we contribute to the XBRL literature (e.g., Blankespoor 2019; Chen et al. 2022; Dong et al. 2016) by demonstrating its usefulness for machine learning.

forward-looking information (Nylen et al. 2025; Küster et al. 2025). Collectively, this evidence underscores the importance of disaggregating narrative text into accounting topics rather than studying disclosures solely in the aggregate.

To identify topics, these studies rely primarily on manual classification or commercial data aggregators such as Audit Analytics, which identifies the topic of misstatements, material weaknesses, comment letters, and critical audit matters. While these approaches have yielded useful insights, they also have important limitations. Manual classification requires significant time and judgment, often forcing researchers to restrict the volume of disclosures examined (de Kok 2025). Data provided by commercial data aggregators is expensive, limited in coverage, and subject to lags that can delay availability by months or years. Importantly, data providers do not cover many narrative sections of the 10-K, such as the MD&A, where economically important accounting discussions occur. Even when disclosures are collected, researchers must use the topics provided, even if additional granularity or a different classification is desired to distinguish economically distinct concepts (e.g., revenue and deferred revenue, debt and equity are combined in Audit Analytics coding).

Cleverly, a small number of studies have used non-LLM machine learning approaches to topic modelling. These methods can discover themes from text, and in some cases have identified some accounting topics in the process (e.g., Brown et al. 2020; Cong et al. 2024; Hoberg and Lewis 2017). However, these approaches are designed for topic discovery rather than the identification of specific accounting topics, leaving a gap for tools that enable systematic, topic-specific analysis of narrative financial disclosures.

Generative AI and Large and Small Language Models

Recent innovations in generative AI, and specifically large and small language models, are well suited for the task of accounting topic classification. This is particularly true as the increasing complexity of financial disclosures and the prevalence of unstructured and unlabeled disclosures require more sophisticated methods of text analysis.

LLMs are advanced machine learning models that exhibit a human-like understanding of text. Open AI's GPT, Meta's Llama, and Google's Gemini, amongst others, are prominent examples. These LLMs rely on a transformer neural network architecture that utilizes masked language modeling and next sentence prediction to learn contextual relationships within and across sentences. In masked language modeling, a word in a sentence is hidden and the model learns to predict it using surrounding context. In next sentence prediction, the model develops an understanding of order and context across different sentences by predicting whether two sentences are logically or randomly connected. These training mechanisms help the model develop a nuanced understanding of language, word order, and context.

While these models are most known for their ability to generate text, they also have lesser-known capabilities for natural language processing tasks. For our purposes, these models can perform classification tasks and, specifically, be modified (a process known as fine-tuning) to classify text into accounting topics, which overcomes the challenges of using commercial data aggregators or manual classification. The contextual awareness of these models is particularly important in accounting, where similar terminology may appear across different topics. For example, fair value discussions may appear in both business combinations and derivatives and hedging disclosures, and the model's understanding of the surrounding context helps distinguish between them.

While GPT-4 (the model behind ChatGPT) is the most well-known of these models, there are significant challenges when using it for research applications.[11] Meta's Llama-2-7B has several features that make it better suited for our task. Llama, an acronym for Large Language Model Meta AI, was first introduced by Meta in 2023.[12] An important advantage of Llama for research applications is that it is open source, meaning it is freely available and can be modified and distributed.[13] This enables us to download Llama-2-7B, customize it for our task of accounting topic classification, and make the resulting model available for future researchers. In contrast, closed-source models such as GPT-4 are costly to use at scale, restrict access, and cannot be freely shared or adapted. While these models can perform well out of the box (zero-shot), the ability to fine-tune can result in better performance for specific tasks such as ours. Specifically, in the fine-tuning process, models train on labeled data to "learn" to perform a specific natural language processing task, such as classifying data or predicting outcomes accurately in out-of-sample

[11] Nevertheless, several working papers have used ChatGPT to answer interesting research questions where these challenges are less restrictive (i.e., where costs are less prohibitive, where fine-tuning is not necessary, and the researchers' objective is not to share the model). For example, Kim et al. (2024) use ChatGPT to generate a measure of information bloat; Jha et al. (2024) and Jha et al. (2025) use ChatGPT on conference calls to measure managements' anticipated capital expenditures and economic outlook, respectively; Chen et al. (2024), Chen et al. (2025), and Lopez-Lira and Tang (2024) use ChatGPT to extract information to predict stock returns; Fan et al. (2024) use ChatGPT to measure firm-level misinformation; Bai et al. (2024) ask ChatGPT to answer earnings-related questions and compare its responses to those from managers during conference calls;

[12] Since it was introduced in 2023, Llama-2 is gaining interest to researchers. Working papers fine-tune Llama-2 to summarize and classify the sentiment of financial news (Konstantinidis et al. 2024; Luo et al. 2024; Pavlyshenko 2023) and bank earnings calls (Cook et al. 2024). Recently, Bernard et al. (2024) fine-tune Llama-3 to create a measure of business complexity.

[13] While Meta makes Llama-2 available at no cost for most, the Open Source Initiative notes that there are some commercial restrictions (licenses with over 700 million active daily users) that call into question the "open source" definition. These restrictions do not affect academic researchers and most commercial applications.

data.[14,15] Smaller fine-tuned models can perform comparably to or better than larger non-fine-tuned models (de Kok 2025).

Another benefit of Llama-2 is the availability of multiple model sizes. Specifically, in partnership with Microsoft, Meta released Llama-2 in three model sizes, 7B, 13B, and 70B.[16] We use the 7B model, a SLM that is especially accessible for academic researchers. It can be deployed locally on personal computers with graphics processing units (GPUs), substantially reducing the resource demands typically associated with LLMs. This addresses the downside of technological advancement, where there are concerns that LLMs require substantial energy and infrastructure resources and have a concerning impact on the carbon footprint (e.g., Erdenesanaa 2023; UNESCO 2024). While the raw performance of 7B is likely to be inferior to larger models, it can yield similar and even superior performance when fine-tuned for a specific task (de Kok 2025; UNESCO 2024).

Research Design

Our Model: Fine-tuning Llama-2-7B for Accounting Topic Classification

Our model results from fine-tuning Llama-2-7B for accounting topic classification. This section details the development of our model.

Labeled dataset: Accounting topics from XBRL and FASB Codification

[14] Fine-tuning is a specific type of supervised learning where a pre-trained model is further trained on a smaller labelled dataset. Supervised learning methods, including support vector machines, gradient boosting, supervised LDA, and random forest regression trees, have mostly been used for prediction tasks in finance and accounting research. Krupa and Minutti-Meza (2022) provide a review of these methods and related literature. For example, studies have used these methods to predict fraud (Cecchini et al. 2010) and returns (Frankel et al. 2022; Siano 2025). Bochkay et al. (2023) discuss how deep learning methods outperform these traditional supervised learning methods.

[15] A common example of supervised machine learning is the classification of spam emails into a separate folder, where the method learns from the user classification of emails to spam and then applies to new unseen emails.

[16] The "7B" (13B,70B) refers to the 7 billion parameters that the model learns during training. These parameters determine how the model processes input data and generates its output. Parameters in LLMs are crucial because they influence the model's ability to understand language. Generally, more parameters allow the model to capture more complex patterns and relationships in the data, which can lead to better performance. However, more parameters also increase the computational resources needed for training and inference.

One of the primary obstacles to fine-tuning models is the acquisition of a substantial amount of labeled data. In our case, we require textual data labeled with its accounting topic.[17] We overcome this challenge by leveraging Extensible Business Reporting Language (XBRL) to identify text referring to accounting topics drawn from authoritative reporting standards.

Public companies are required to produce financial reports in XBRL and "tag" each numerical value and financial statement note. The latter is particularly suitable for our task for several reasons. First, the financial statement notes cover a broad set of accounting topics, which is an essential requirement for our research objective. Second, preparers are required to tag each note with a standardized label (i.e., taxonomy tag) that maps the note to a specific accounting concept in the FASB Accounting Standards Codification, which is the most comprehensive source of such topics. This results in an extensive set of labeled data (more than 300,000 XBRL tags in our training set), which is typically unavailable in other machine-learning applications (de Kok 2025). Utilizing the mapping to the Codification also removes subjectivity and human error from the labeling process. Lastly, the notes comprise a substantial and diverse amount of topic-specific text, enabling Llama to associate specific text with corresponding accounting topics. We utilize this large, labeled dataset to develop our model and subsequently test its performance.

To collect financial statement notes, we use Python to download XBRL data for all public companies from 2009 to 2021. We limit the sample to firms with coverage in Compustat. We parse all the XBRL tags and keep only tags of financial statement notes.[18] We focus on the taxonomy tags with the highest frequency, resulting in 85 distinct tags that together account for 92.5 percent

[17] Prior studies that finetune machine learning models typically use smaller amounts of manually classified data (e.g., Huang et al. 2023).

[18] Notes (in their entirety) are tagged with "Textblock" tags. Textblocks tags have XBRL type: 'dtr-types:textBlockItemType'. To focus on topic-specific tags, we remove Textblocks that contain policies, tables, HTML links, or schedules. We also remove custom (extended) tags, as those are not mapped into the FASB codification. We use these custom tags in one of our research applications.

of all taxonomy tags disclosed in annual 10-K filings. For example, the most prevalent tag is "IncomeTaxDisclosureTextBlock" which is used by almost all companies and appears 43,848 times.[19] Each tag corresponds to an accounting concept in the FASB Codification, and multiple tags can map to a single concept, resulting in the same label being assigned to them. For example, the tags "BusinessCombinationDisclosureTextBlock", "MergersAcquisitionsAndDispositionsDisclosuresTextBlock" and "BusinessAcquisitionIntegrationRestructuringAndOtherRelatedCostsTextBlock", all map to the business combinations topic in the Codification. The complete list of the 85 XBRL tags and their corresponding mappings to accounting topics is presented in Appendix A.[20]

After processing the text data of all notes and eliminating outlier notes and firms,[21] the result is 470,238 notes covering 32 accounting topics. Overall and by topic, we randomly split this sample into approximately 75 percent of firms for training and evaluation (321,350 + 35,708 = 357,058 notes) and 25 percent held out for testing (113,180 notes). The actual split is 75.9 and 24.1 percent since we require each sample to belong to a distinct set of firms. This is important because the textual description of notes is sticky over time, and using the notes of the same firm across the training and test data would artificially inflate model performance.

Table 1 shows the frequency of each topic in each of these samples. Many topics are prevalent, with compensation, commitments and contingencies, tax, and debt leading in

[19] In contrast, the note "DisclosureOfIncomeTaxExplanatory" appeared only five times. We excluded this and other low-frequency tags from our final sample.
[20] In a few cases, we combine Codification topics where granularity does not seem necessary (e.g., compensation and stock compensation into compensation). We also removed certain notes that discuss broad issues such as quarterly or segment information and not a specific accounting topic.
[21] Specifically, we remove 621 firms with less than ten notes across our sample period and remove notes with less than 20 words, since Llama requires sufficient text to recognize topics. We also performed text normalization, including removing tables, converting all text into lowercase, and eliminating duplications, special characters, punctuation, and roman numerals.

frequency.[22] Notably, to remain objective, we keep topics separate when they are separate in the Codification (e.g., revenue and deferred revenue, liabilities and debt, etc.). This granularity allows flexibility to combine topics as needed.

Fine-tuning Llama-2-7B

We download Llama-2-7B from Huggingface and fine-tune it to perform our new task of accounting topic classification.[23] Since Llama is already trained to understand language and context, we provide the model with the entire text of the disclosures and do not perform further preprocessing steps, such as lemmatization or removing stop words.[24] For fine-tuning, we provide the model each note and its XBRL topic label, and the model "learns" by repeatedly adjusting its internal parameters to reduce the gap between its predicted topic and the correct label. In practical terms, this process strengthens the connections between certain words, phrases, and accounting topics so that the model becomes more accurate at predicting the 32 predefined accounting topics.

While 357,058 notes are available in the training sample for fine-tuning, we hold out approximately 10 percent of observations as an evaluation sample to periodically assess performance and ensure the model is not overfitting. Based on this evaluation, we stop the training process when the model's performance ceases to improve. Our resulting fine-tuning sample consists of 192,000 observations. Appendix B provides further technical details of our fine-tuning process.

Testing model performance using out-of-sample data

[22] Certain topics are industry-specific (e.g., financial services, oil and gas, real estate). We include them because they have specific disclosure requirements and point to a separate section of the Codification.
[23] Huggingface is a platform where machine learning models are freely shared. You can download the base Llama-2-7B model here: https://huggingface.co/meta-llama/Llama-2-7b.
[24] The model is equipped to handle 4,096 tokens (about 3,000 words). However, to make the fine-tuning process more efficient, we truncated the input to 1,024 tokens, which is the full note in 75 percent of cases.

Next, we evaluate performance by using our now fine-tuned model to predict topics in the testing sample (113,180 notes of unseen companies). During this testing stage, our model only sees the text of the notes and assigns one of the 32 topic labels. We can then assess the classification performance relative to the actual known topic.

Table 2 displays a matrix of our model's classification of each observation (columns) relative to each observation's actual topic (rows). For example, the testing sample includes 4,130 observations in the investments topic (the sum of values in Row 17). Of the 4,204 observations identified by the model as investments (the sum of values in Column 17), 3,951 are actually related to the investments topic (the intersection of Row 17 and Column 17) capturing 95.67 percent of the actual notes, and 253 are misclassified as the investment topic. Overall, the matrix follows a clear diagonal pattern representing accurate classification.

Three performance metrics are commonly used in machine learning: Precision, Recall, and the F1 score. Figure A illustrates a confusion matrix, which is a common technique to summarize the performance of a classification algorithm. We again use the investments topic as an example in this Figure. First, precision is a percentage calculated as the number of observations correctly classified into a topic (True Positive) scaled by the total number of observations classified into that topic (True and False Positives). Our model's precision for investments is 93.98 percent, calculated as 3,951 true positives divided by the sum of true positives and false positives (3,951+253). This illustrates that when classifying the Investments topic, our model has high precision and hence returns significantly more correct classifications relative to incorrect classifications. The second metric, recall, is a percentage calculated as the number of observations correctly classified into a topic (True Positive) scaled by the actual number of observations in that topic (True Positives and False Negatives). Our model's recall for investments is 95.67 percent, calculated as 3,951 true positives divided by the sum of true positives and false negatives (3,951+179). Lastly, the F1 score combines these two metrics by their harmonic mean.[25] The F1 score for investments for our model is 94.82.

Table 3 presents these performance metrics for each of the topics. Precision ranges from 78.2 to 100.0 percent and recall ranges from 70.1 to 100.0 percent. Our model performs particularly well at identifying common topics as well as topics that are important to researchers and investors (e.g., revenue and tax have F1 scores of 97.3 and 100.0 percent, respectively). Additionally, because of its understanding of context, our model is successful at distinguishing between topics that can contain similar language. For example, it learns which impairment discussions relate to goodwill and intangibles, PPE, and restructuring during fine-tuning and performs with an F1 score of 98.9, 98.1, and 96.8 percent for these topics in testing.

Overall, our model's weighted F1 score average is 95.5 percent. Our model's performance demonstrates that it is well suited for the task of classifying accounting topics.

Comparing to other generative language models

As discussed earlier in the paper, there are other generative language models suitable for our task, though each has its challenges. For example, Open AI's GPT-4 (the model behind ChatGPT) and Google's Gemini are closed-source, meaning they cannot be easily or freely fine-tuned, shared, and replicated by other researchers. Nevertheless, it is possible that using these models as they are (i.e., zero-shot without fine-tuning) can result in comparable performance. To assess this, we used the application processing interface (API) of GPT-4o, GPT-4o-mini, and Gemini-1.0-pro and instructed the models to perform our classification task on a subset of our test

[25] Specifically, the F1 score is calculated as 2 x (Precision x Recall) / (Precision + Recall).

data.[26] Table 4 summarizes the performance of each model, where the weighted average F1 scores are 86.0, 83.5, and 84.2 percent, respectively.[27] These results demonstrate that, in addition to the conceptual advantage of being open source, our model outperforms these closed-source models on this domain-specific task. This finding is consistent with the recent trend of training SLMs to perform specialized tasks.

We also compare our model to the base untrained version of Llama-2-7B (i.e., before our fine-tuning process), which results in a weighted average F1 score of 51.7 percent. This confirms the importance of fine-tuning, since smaller models like ours only outperform larger models when fine-tuned for a specific task.

Comparing to LDA

Latent Dirichlet allocation (LDA) is an earlier machine learning technology that discovers topics in textual data (Blei et al. 2003) and is popular in finance and accounting research (e.g., Bao and Datta 2014; Bellstam et al. 2021; Brown et al. 2020; Bybee et al. 2024; Hoberg and Lewis 2017; Huang et al. 2018; Lowry et al. 2020; Ryans 2021). While LDA is useful for discovering topics, there are several challenges to using it to identify specific topics. First, LDA is an unsupervised method, meaning it is not exposed to labeled data and instead discovers groups of related words. Thus, LDA is useful when looking for hidden themes but less ideal when looking for specific topics as it can miss topics and return topics unrelated to the research objective. Second, LDA uses a "bag of words" approach that ignores context and word order. An ambiguous word such as "impairment" could be discussing goodwill, PPE, or restructuring, and LDA would

[26] Because using the closed models is expensive, we select a stratified sample of 1,005 notes, which maintains the same proportion of note topics as the entire testing sample. For example, business combinations represent 4.6 percent of the testing sample, so the stratified sample also contains 4.6 percent business combinations notes.
[27] To ensure the smaller stratified testing sample is not driving the lower performance, we also apply our model to the 1,005 observation subset. The resulting F1 score is again 95.5 percent, which alleviates this concern and further confirms the superior performance of our model.

not be able to distinguish based on context, which can lead to the misclassification of topics. Lastly, researcher subjectivity is required at several points in the LDA process, which can affect its output and replicability across studies (Gentzkow et al. 2019).[28]

Nevertheless, we apply LDA to our data to ensure the conceptual advantages of our model translate to superior performance. LDA is shown the text of the notes and outputs groups of related words (based on the frequency of words that match) which we then label into topics.[29] 21 of the 32 topics appear in this output, meaning that 11 are not identified by LDA.[30] In testing, we then use the LDA identified word groupings to predict topics in the out-of-sample data. There is significant variation in its performance across the individual topics, with a lower F1 score than our model for every topic and a weighted average F1 score of 72.0 percent. In summary, this analysis confirms that our model is more comprehensive (i.e., identifies all 32 topics) and accurate (i.e., achieves higher performance metrics) than LDA for the task of accounting topic classification.

We recognize that researchers have also applied other methods of topic modeling that, like LDA, do not involve LLMs. In particular, Bae et al. (2025) use nonnegative matrix factorization to discover topics in the MD&A and risk factor sections of the 10-K. Cong et al. (2024) recently introduced a textual factors approach, which extends beyond LDA to allow topics (clusters of related words) to be both discovered and formed around researcher-selected seed words.[31] Our

---

[28] Specifically, researchers must decide on a number of topics (word groupings) for LDA to output a priori, which can be particularly subjective given that structure and topics are presumably unknown. Additionally, researchers must assign topic labels to the word groupings, typically by manually examining the most frequent words. This can be difficult when the word groupings appear irrelevant to the research objective or when a single theme does not clearly arise. Last, after labels are assigned to word groupings, researchers may decide to collapse similar topics together.

[29] Two of the authors independently labeled each group of words into one of the 32 topics. The two authors, along with a third author, reviewed and reconciled inconsistencies in classification. The process of labeling topics is difficult. Even though the topic labels are known, the researchers had a 19 percent disagreement in labeling because only specific words and their frequency, not their context, are observed.

[30] The unidentified are: asset retirement, cash, deferred revenue, insurance, inventory, liabilities, other assets, receivables, research and development, transfers and servicing, and warranties and guarantees.

[31] Rouen et al. (2024) similarly use seed words in semantic vector analysis to identify ESG content in corporate disclosures.

objective differs from LDA and these additional methods of topic classification in its intent to train a model to identify 32 *specific* topics drawn from authoritative reporting standards, rather than to discover topics.

Findings: Disaggregating the MD&A into Accounting Topics

Our study focuses on the value of disaggregating narrative disclosures into accounting topics rather than analyzing them only in aggregate. While prior research on the MD&A has typically examined overall tone or length, our model allows us to investigate economically important insights at the topic-level. The MD&A (Item 7 of the 10-K) is an important unstructured narrative disclosure in which management explains the financial statements and provides perspective on the performance and outlook of the company. Although the MD&A is a rich source of accounting-related information, it is unstructured and not collected by commercial data aggregators, making topic-level analysis difficult at scale. Manual classification is possible, but it is costly, subjective, and typically limited to small samples.

Research examining the MD&A is extensive (e.g., Bae et al. 2025; Brown and Tucker 2011; Brown et al. 2024; Cho and Muslu 2021; Davis and Tama-Sweet 2012; Feldman et al. 2010; Hoberg and Lewis 2017; Li 2010; Loughran and McDonald 2011; Mayew et al. 2015), but most studies focus on aggregate characteristics of the disclosure. Relatively few studies examine topics within the MD&A, and none isolate accounting topics specifically. Hoberg and Lewis (2017), for example, use LDA to discover discussion areas in the MD&A and find that firms involved in fraud have abnormally long discussions in certain areas such as revenue growth. Brown et al. (2020) use LDA and find that certain discussions in the 10-K (e.g., increase in income compared to prior periods) are associated with the likelihood of a misstatement (overall, not by topic). More recently, Bae et al. (2025) use nonnegative matrix factorization to similarly discover topics in the MD&A. These studies provide useful approaches for topic discovery, but as discussed in our earlier comparison to other models, there are several weaknesses when using them for our task of consistently identifying specific topics.

Classifying MD&A topics

We obtain the MD&A section of the 10-K from Calcbench and use our model to classify it into topics. Table 5, Panel A details the sample derivation, which results in 36,221 firm-year observations. We show our model the text of 3,937,612 MD&A paragraphs for these firm-year observations. The model provides a confidence score for each topic prediction, which ranges from 0 to 100 percent, and represents the likelihood that the topic prediction is correct. The model classifies 1,612,310 paragraphs into one of the 32 topics with over 90 percent confidence, and we use this cut-off in our analysis. Table 5, Panel B shows the 20 most common topics with at least one paragraph in the MD&A at the firm-year level, with the most frequent being tax, debt, equity, commitments and contingencies, and revenue.

Multivariate analysis

Using the topics identified by our model, we examine the information content of MD&A narratives by testing whether the extent and tone of specific accounting topic discussions predict economically important outcomes. This approach links what is disclosed (the accounting topic) with how it is disclosed (the extent and tone), extending prior research that has largely analyzed the MD&A in aggregate.[32] We evaluate whether these topic-specific measures are associated with regulatory scrutiny, financial reporting quality, and market response.

[32] For example, several studies leverage the narrative form of the MD&A to examine whether the tone used by management provides information (e.g., Cho and Muslu 2021; Davis and Tama-Sweet 2012; Feldman et al. 2010; Jiang et al. 2019; Kim et al. 2024; Li 2010; Loughran and McDonald 2011). Further, Hoberg and Lewis (2017) and Brown et al. (2020) examine areas of MD&A and general 10-K discussion that led to overall financial reporting issues.

Aggregate MD&A analysis

We begin by replicating the conventional aggregate approach. We examine whether the overall length and pessimistic tone of the MD&A are associated with regulatory scrutiny, financial reporting quality, and market response using the following model

$$DEPVAR_{i,t} = \beta_0 + \beta_1 MDA\text{-}NEG_{i,t} + \beta_2 MDA\text{-}LENGTH_{i,t} + \sum \beta_j Controls_{i,t} + Industry\ and\ year\ fixed\ effects + \varepsilon_{i,t} \quad (1)$$

We examine three dependent variables: comment letters (*CL*), financial misstatements (*MISSTATE*), and signed cumulative abnormal returns beginning with the 10-K date for two alternative windows (*CAR [0,2]* and *CAR [0,3]*).

To capture tone, we follow the majority of MD&A tone studies and focus on pessimistic tone (i.e., negative words) as defined by Loughran and McDonald (2011) and used in recent research (e.g., Cao et al. 2021; Huang et al. 2020). *MDA-NEG* is the percentage of words with a negative sentiment relative to the total words in the MD&A. *MDA-LENGTH* is the natural log of the total number of words in the MD&A. We control for a standard set of control variables that we expect to influence the dependent variables and include industry and year fixed effects.[33] Following prior literature, we include additional controls in the market reaction models (e.g., Arif et al. 2019; Frankel et al. 2022; Loughran and McDonald 2011). All variables used in the analyses are defined in Appendix C.

Table 6 displays the results of logit models for comment letters and misstatements (Columns 1 and 2, respectively) and OLS models for market reaction (Columns 3 and 4).[34] Results

---

With our model, it is possible to improve identification and reveal more nuanced relationships. Specifically, our model allows us to extend past research by quantifying textual measures within specific accounting topics and, where possible, matching topic-specific predictors to topic-specific dependent variables.

[33] We use Fama-French industry classification for industry fixed effects. To preserve as many observations as possible, we use Fama-French 12 in logit models. We use Fama-French 48 in OLS models.

[34] In certain logit models, observations drop due to perfect prediction of the dependent variable.

reveal that the aggregate textual measures do not significantly predict the likelihood of comment letters or misstatements. The coefficient on *MDA-NEG* is, however, significant and negatively associated with abnormal returns to the 10-K filing. This is consistent with prior literature, which largely finds that the pessimistic tone of the MD&A is associated with negative returns (e.g., Feldman et al. 2010; Frankel et al. 2022; Loughran and McDonald 2011).

These aggregate findings (and lack thereof) motivate our topic-specific analysis, as we propose that investigating textual measures within specific accounting topics will improve identification and potentially reveal more nuanced relationships.

Topic-specific MD&A analysis

Disaggregating the MD&A may reveal economically important relationships that are obscured at the aggregate level. To test this, we focus on the accounting topics our model reveals are most frequently discussed in the MD&A – tax, debt, equity, commitments and contingencies, and revenue.[35,36] These topic-specific discussions capture key elements of firm operations and risks. Tax discussions typically cover the company's current tax rate and related tax law updates. Debt and equity discussions provide insight into the company's capital structure and financing strategies. Commitments and contingencies discussions entail potential future obligations and uncertainties, such as legal proceedings or environmental liabilities. Lastly, revenue discussions provide insight into how the company generates income, concentration by major customers, recent trends, and more.

[35] For practical reasons, we cannot present an analysis for all 32 topics. We avoid subjectivity by focusing on the most prevalent topics from Table 5, Panel B. Future researchers can choose which topics are most relevant to address their research questions.
[36] For analysis, we combine equity and debt into one topic to match Audit Analytics coding of topics. We also combine revenue and deferred revenue into one topic since companies typically combine both discussions in the MD&A (unlike notes to the financial statements, where they are separate).

To measure the extent of discussion, we calculate the percentage of words devoted to each topic (*TAX-LENGTH*, *DEBT-EQUITY-LENGTH*, *CC-LENGTH*, *REVENUE-LENGTH*). To measure tone, we calculate the percentage of words with a pessimistic tone relative to total words for the given topic (*TAX-NEG*, *DEBT-EQUITY-NEG*, *CC-NEG*, and *REVENUE-NEG)*. Table 5, Panel C displays the descriptives of these variables, where we observe that the average firm-year dedicates 5.4, 8.3, 2.6, and 3.5 percent of the MD&A to tax, debt and equity, commitments and contingencies, and revenue, respectively. For each of our identified topics, 0.9, 0.6, 1.9, and 0.4 percent, respectively, of words have a pessimistic tone.

Regulatory scrutiny. We first examine whether the extent and tone of specific accounting topic discussions are associated with subsequent regulatory scrutiny. Prior literature finds that comment letters are more likely when accounts are riskier and more complex (e.g., Ahn et al. 2020; Cassell et al. 2013). Lengthier discussions or a more pessimistic tone may therefore draw regulatory attention, increasing the likelihood of receiving comment letters on the topic. Yet, it is also possible that common topics are less likely to have issues if their prominence suggests greater attention from management and other financial reporting stakeholders. To understand whether this is the case, it is important to match topic-specific predictors (e.g., revenue discussions in the MD&A) with topic-specific outcomes (e.g., revenue comment letters) rather than aggregate outcomes (e.g., aggregate comment letters). The majority of prior research has not attempted to link the disclosure of certain accounting topics to issues in related areas in this way.

Table 7, Panel A reports multivariate results using topic-specific comment letters as the dependent variables. In addition to the control variables used in aggregate analysis (Model (1) and Table 6), we include account-specific controls. The extent of discussion of all four topics (coefficients on *TAX-LENGTH, DEBT-EQUITY-LENGTH*, *CC-LENGTH*, and *REVENUE-*

LENGTH) are significantly positive, and the pessimistic tone of tax, commitments and contingencies, and revenue discussions (*TAX-NEG*, *CC-NEG,* and *REVENUE-NEG*) are also significantly positive. This indicates a higher likelihood of comment letters in these areas when there is more discussion and when a more pessimistic tone is used, suggesting that regulators respond to how accounting topics are disclosed. The average marginal effects of these associations (untabulated) indicate economic significance. Specifically, when holding other variables constant, the probability of receiving a tax-related comment letter increases by 9.95 percentage points relative to the sample mean when *TAX-LENGTH* moves from the 25^{th} to 75^{th} percentile. This increase is 18.31, 1.13, and 7.25 percentage points for *DEBT-EQUITY-LENGTH*, *CC-LENGTH*, and *REVENUE-LENGTH*, respectively. For tone, this increase is 6.36, 3.01, and 3.36 percentage points for tax, commitments and contingencies, and revenue, respectively.

Financial reporting quality. Our second analysis examines whether topic-specific disclosure predicts subsequently revealed misstatements. Misstatements represent a direct indicator of financial misreporting, and prior work shows that narrative disclosures can contain signals of aggregate misstatement risk (e.g., Hoberg and Lewis 2017; Brown et al. 2020). Most prior studies have similarly examined aggregate misstatements rather than linking specific disclosures, like the MD&A, to misstatements in the same area.

Table 7, Panel B reports results using topic-specific misstatements as the dependent variable. The extent of commitments and contingencies and the extent and pessimistic tone of revenue discussions (coefficients on *CC-LENGTH*, *REVENUE-NEG*, *REVENUE-LENGTH*) are significantly positive, showing that how certain important topics are discussed is associated with related misstatements. The probability of commitments and contingencies and revenue misstatements increase by 4.41 and 7.57 percentage points (relative to the sample mean) when the length of associated MD&A discussions increases from the 25th to 75th percentile. The probability of revenue misstatement increases by 7.24 percentage points when the pessimistic tone of revenue discussions increases from the 25th to 75th percentile.

Combined, the comment letter and misstatement analyses highlight the economic importance of disaggregating disclosures. While aggregate MD&A measures are not significant, topic-specific predictors are significantly associated with topic-specific regulatory scrutiny and reporting outcomes.

Market reaction. Third, we examine whether topic-specific disclosure is associated with market reactions. Prior studies find that the market reacts to the overall tone of the MD&A (e.g., Feldman et al. 2010; Frankel et al. 2022; Loughran and McDonald 2011), but it remains unclear which portions of the disclosure drive this relationship. Table 8, Panel A reports multivariate results of the topic-specific textual measures and market reaction. Results reveal that only the coefficient on *CC-NEG* is significantly negative, whereas none of the length measures or sentiment measures for other topics are significant. The *CC-NEG* finding is economically significant, where, for example, *CAR [0,3]* decreases by 0.22 percent, relative to the average sample CAR of -0.16 percent, when the pessimistic tone of commitments and contingencies discussions increases from the 25th to 75th percentile. The coefficient of -0.065 suggests CAR decreases 650 basis points for every unit increase in *CC-NEG*. This clarifies prior aggregate findings and demonstrates that the tone of commitments and contingencies discussions in the MD&A is informative to the market.

While it may at first seem unexpected that common and important accounts such as tax and revenue are not significant, the 10-K is generally redundant from earlier disclosures and our results confirm that the market does not view the discussion of these topics as new information. However, disclosures related to commitments and contingencies, which entail potential future obligations and uncertainties, are often viewed as inadequate (Standridge et al. 2023; Gleason and Mills 2002; Hennes 2014). Commitments and contingencies may be less likely to be covered in earlier releases (as opposed to an account like revenue or tax), whereas the MD&A is uniquely intended for management to provide a discussion of the company's outlook. Investors appear to view pessimistic language in this area as incremental and value-relevant information.

To validate this interpretation, we conduct two additional analyses. First, we apply our model to a sample of earnings announcements and find (untabulated) that the most frequent topics are equity, tax, debt, revenue, and business combinations. Commitments and contingencies rank much lower (eighth) than in the MD&A, consistent with the information having greater salience when the 10-K is released.[37] Second, we partition our sample by whether the 10-K is disclosed concurrently or non-concurrently with the earnings announcement. Table 8, Panel B displays this result, which indicates that the significant result is only within the non-concurrent sample when the new commitments and contingencies information would be most salient. The insignificant result in the concurrent earnings sample is likely due to the relatively lower attention and importance placed on new commitments and contingencies information amidst the volume of information disclosed in the 10-K and earnings announcement.

Summary of topic-specific MD&A analysis. Topic-specific measures of extent and tone predict the likelihood of receiving comment letters and of misstatements in related areas, and investors respond to pessimistic discussions of commitments and contingencies. These findings demonstrate the importance of moving beyond aggregate MD&A analysis and provide new evidence that topic-level variation in narrative disclosures has regulatory, reporting, and market consequences.

---

[37] The MD&A likely discusses accounting topics more extensively than the earnings announcement, so in reality, this differential in relative frequency is likely even greater.

Additional analysis: Classifying custom-tagged notes

While our main analyses demonstrate the importance of disaggregating narrative disclosures into accounting topics, another challenge arises when disclosures are already disaggregated but not easily interpretable. This is the case for financial statement notes with custom XBRL tags. Preparers use custom extended tags when they decide that no standard taxonomy tag is suitable, but this practice reduces comparability[38] because the tags are not linked to the Codification, and their accounting meaning is unclear without manually reading the note. The use of custom tags is common, with 54.1 percent of firm-year observations using at least one custom tag for their notes, and identifying the underlying accounting topic of these notes is difficult at scale.

We apply our model to classify the text of custom-tagged notes into accounting topics, enabling comparability across firms. Table 9, Panel A shows that the most common topics are equity, debt, investments, commitments and contingencies, and receivables. Notably, this distribution differs from the MD&A, illustrating that the model adapts to the disclosure context. In Table 9, Panel B, we examine whether classifying custom notes provides insight into regulatory scrutiny, given that comment letters often focus on how information is disclosed. For each of the four most common topics, we create indicator variables (*CUSTOMFN-DEBT-EQUITY*, *CUSTOMFN-INVESTMENTS*, *CUSTOMFN-CC*, and *CUSTOMFN-RECEIVABLES*) that equal to one when a firm-year uses a custom tag for a note and our model classifies the note to that topic with over 90 percent confidence. Results across all four topics show that using a custom tag in these topics is associated with a significantly higher likelihood of receiving a related SEC comment

[38] Prior research has noted that using custom tags undermines the objective of XBRL to enable machines to automatically extract and compare information across companies (Debreceny et al. 2011; Huang et al. 2019; Li and Nwaeze 2015; Li and Nwaeze 2018; SEC 2014).

letter. These findings suggest that classifying the text of custom-tagged notes into accounting topics can provide new insight into how regulators respond to these disclosures. More broadly, the analysis illustrates how applying our model to otherwise opaque disclosures makes it possible to study topic-specific outcomes in settings where comparability would otherwise be limited.[39]

Conclusion

As textual data analysis become increasingly important to accounting and finance research (Cong et al. 2019; Bochkay et al. 2023; de Kok 2025; Gentzkow et al. 2019), a persistent challenge has been the difficulty of identifying accounting topics in unstructured narrative disclosures. Aggregate approaches obscure economically meaningful variation, and reliance on data aggregators or manual classification to identify topics has its challenges. Our study addresses this challenge by showing that disaggregating narrative text into accounting topics provides new insights into disclosures that are otherwise hidden at the aggregate level.

We develop a domain-specific SLM, fine-tuned on XBRL-labeled financial statement notes, that accurately classifies accounting topics at scale. Applying this model to the MD&A, we show that topic-level variation in both extent and tone predicts the receipt of topic-specific SEC comment letters, the occurrence of related misstatements, and, in the case of commitments and contingencies, market reaction at the 10-K filings data. In contrast, aggregate measures of MD&A tone and length provide limited predictive power. These findings demonstrate the economic importance of examining narrative disclosures at the topic level.

[39] While we do not have reason to believe the use of custom tags to represent accounting topics will be associated with related misstatements, for completeness we examine this in untabulated additional analysis. All associations are insignificant.

Our study has several limitations. While our approach enables scalable topic-level analysis of narrative disclosures, it is designed to recognize predefined accounting topics and does not capture additional themes. Other versions of topic modeling, such as LDA, are better suited for this objective. Further, our analyses focus on U.S. public company filings, and the generalizability of findings to other disclosure regimes may be limited. Despite these constraints, we provide evidence that topic-level analysis uncovers economically important insights, and we offer a practical tool to facilitate future research.

References

Abbott, L., and W. L. Buslepp. 2023. Do critical audit matters provide decision-relevant information to investors? Evidence from merger and acquisition announcements. Working paper.

Ahn, J., R. Hoitash, and U. Hoitash. 2020. Auditor task-specific expertise: The case of fair value accounting. The Accounting Review 95(3): 1-32.

Arif, S., N. T. Marshall, J. H. Schroeder, and T. L. Yohn. 2019. A growing disparity in earnings disclosure mechanisms: The rise of concurrently released earnings announcements and 10-Ks. Journal of Accounting and Economics 68(1).

Bae, J. W., F. Belo, J. Li, X. Lin, and X. Zhao. 2025. The opposing effects of complexity and information content on uncertainty dynamics: Evidence from 10-K filings. Management Science, forthcoming.

Bai, J., N. Boyson, Y. Cao, M. Liu, and C. Wan. 2023. Executives vs. chatbots: Unmasking insights through human-AI differences in earnings conference Q&A. Working paper.

Bao, Y., and A. Datta. 2014. Simultaneously discovering and quantifying risk types from textual risk disclosures. Management Science 60(6): 1371-1391.

Bellstam, G., S. Bhagat, and J. A. Cookson. 2021. A text-based analysis of corporate innovation. Management Science 67(7): 4004-4031.

Bens, D. A., M. Cheng, and M. Neamtiu. 2016. The impact of SEC disclosure monitoring on the uncertainty of fair value estimates. The Accounting Review 91(2): 349-375.

Bernard, D. E. Blankespoor, T. de Kok, and S. Toynbee. 2024. Using GPT models to measure the complexity of business transactions. Working paper.

Blankespoor, E. 2019. The impact of information processing costs on firm disclosure choice: Evidence from the XBRL mandate. Journal of Accounting Research 57.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. Journal of Machine Learning Research.

Bochkay, K., S. V. Brown, A. J. Leone, and J. W. Tucker. 2023. Textual analysis in accounting: What's next? Contemporary Accounting Research 40(2): 765-805.

Brown, S. V., and J. W. Tucker. 2011. Large-sample evidence on firms' year-over-year MD&A modifications. Journal of Accounting Research 49(2): 309-346.

Brown, N. C., R. M. Crowley, and W. B. Elliott. 2020. What are you saying? Using topic to detect financial misreporting. Journal of Accounting Research 58(1): 237-291.

Brown, S. V., L. A. Hinson, and J. W. Tucker. 2024. Financial statement adequacy and firms' MD&A disclosures. Contemporary Accounting Research 41: 126-162..

Burke, J. J., R. Hoitash, U. Hoitash, and X. Xiao. 2023. The disclosure and consequences of U.S. critical audit matters. The Accounting Review 98(2): 59-95.

Bybee, L., B. Kelly, A. Manela, and D. Xiu. 2024. Business news and business cycles. Journal of Finance 79(5): 2905-3677.

Cao, S. S., V. W. Fang, and L. Lei. 2021. Negative peer disclosure. Journal of Financial Economics 140: 815-837.

Cassell, C. A., L. M. Dreher, and L. A. Myers. 2013. Reviewing the SEC's review process: 10-K comment letters and the cost of remediation. The Accounting Review 88(6): 1875-1908.

Cecchini, M.,H. Aytug, G. Koehler, and P.Pathak. 2010. Detecting management fraud in public companies. Management Science 56(7): 1146-1160.

Chen, X., Y. H. Cho, Y. Dou, and B. Lev. 2022. Predicting future earnings changes using machine learning and detailed financial data. Journal of Accounting Research 60(2).

Chen, Y., B. T. Kelly, and D. Xiu. 2024. Expected returns and large language models. Working paper.

Chen, J., G. Tang, G. Zhou, and W. Zhiu. 2025. ChatGPT and Deepseek: Can they predict the stock market and macroeconomy? Working paper.

Cho, H., and V. Muslu. 2021. How do firms change investments based on MD&A disclosures of peer firms? The Accounting Review 96(2): 177-204.

Christensen, B. E., A. J. Olson, and T. C. Omer. 2014. The role of audit firm expertise and knowledge spillover in earnings management through the tax accounts. The Journal of American Taxation Association 37(1).
Cong, L. W., T. Liang, B. Yang, and X. Zhang. 2019. Analyzing textual information at scale. *Information for efficient decision making: Big data, blockchain and relevance.*
Cong, L. W., T. Liang, X. Zhang, and W. Zhu. 2024. Textual factors: Scalable, interpretable, and data-driven approach to analyzing unstructured information. Working paper.
Cook. T. R., S. Kazinnik, A. Lundgaard Hansen, and P. McAdam. 2024. Evaluating local language models: An application to financial earnings calls. Working paper.
de Kok, T. 2025. ChatGPT for textual analysis? How to use generative LLMs in accounting research. Management Science, forthcoming.
Davis, A. K., and I. Tama-Sweet. 2012. Managers' use of language across alternative disclosure outlets: Earnings press releases versus MD&A. Contemporary Accounting Research 29(3): 804-837.
Debreceny, R. S., S. M. Farewell, M Piechocki, C. Felden, A. Graning, and A. d'Eri. 2011. Flex or break? Extensions in XBRL disclosures to the SEC? Accounting Horizons 25(4): 631-657.
Dechow, P. M., A. Lawrence, and J. P. Ryans. 2016. SEC comment letters and insider sales. The Accounting Review 91(2): 401-439.
Dong, Y., Li, O.Z., Lin, Y. and Ni, C., 2016. Does information-processing cost affect firm-specific information acquisition? Evidence from XBRL adoption. Journal of Financial and Quantitative Analysis, *51*(2), pp.435-462.
Drake, K. D., N. C. Goldman, S. J. Lusch, and J. J. Schmidt. 2024. Disclosure of tax-related critical audit matters and tax-related outcomes. Contemporary Accounting Research 41(2): 719-747.
Dyer, T., M. Lang, and L. Stice-Lawrence. 2017. The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. Journal of Accounting and Economics 64(2-3): 221-245.
Erdenesanaa, D. 2023. A.I. could soon need as much electricity as an entire country. The New York Times. October 10.
Fan, J., Q. Liu, Y. Song, and Z. Wang. 2024. Measuring misinformation in financial markets. Working paper.
Feldman, R., S. Govindaraj, J. Livnat, and B. Segal. 2010. Management's tone change, post earnings announcement drift and accruals. Review of Accounting Studies 15: 915-953.
Frankel, R., J. Jennings, and J. Lee. 2022. Disclosure sentiment: Machine learning vs. dictionary methods. Management Science 68(7): 5514-5532.
Gentzkow, M., B. Kelly, and M. Taddy. 2019. Text as data. Journal of Economic Literature 57(3): 535-574.
Gleason, C. A., and L. F. Mills. 2002. Materiality and contingent tax liability reporting. The Accounting Review 77(2): 317-342.
Hennes, K. M. 2014. Disclosure of contingent legal liabilities. Journal of Accounting and Public Policy 33(1): 32-50.
Hoberg, G., and C. M. Lewis. 2017. Do fraudulent firms produce abnormal disclosure? Journal of Corporate Finance 43: 58-85.
Hoitash, R., and U. Hoitash. 2018. Measuring accounting reporting complexity with XBRL. The Accounting Review 93(1): 259-287.
Huang, A. H., R. Lehavy, A. Y. Zang, and R. Zheng. 2018. Analyst information discovery and interpretation roles: A topic modeling approach. Management Science 64(6): 2473-2972.
Huang, F., W. G. No, and M. A. Vasarhelyi. 2019. Do managers use extension elements strategically in the SEC's tagged data for financial statements? Evidence from XBRL complexity. Journal of Information Systems 33(3): 61-74.
Huang, A. G., H. Tan, and R. Wermers. 2020. Institutional trading around corporate news: Evidence from textual analysis. The Review of Financial Studies 33(10): 4627-4675.
Huang, A. H., H. Wang, and Y. Yang. 2023. FinBERT: A large language model for extracting information from financial text. Contemporary Accounting Research 40(2): 806-841.

Inger, K. K., M. D. Meckfessel, M. Zhou, and W. Fan. 2018. An examination of the impact of tax avoidance on the readability of tax footnotes. Journal of the American Taxation Association 40(1): 1-29.
Jha, M., J. Qian, M. Weber, and B. Yang. 2024. ChatGPT and corporate policies. Working paper.
Jha, M., J. Qian, M. Weber, and B. Yang. 2025. Harnessing generative AI for economic insights. Working paper.
Jiang, F., J. Lee, X. Martin, and G. Zhou. 2019. Manager sentiment and stock returns. Journal of Financial Economics 132(1): 126-149.
Kim, A. G., M. Muhn, and V. V. Nikolaev. 2024. Bloated disclosures: Can ChatGPT help investors process information? Working paper.
Konstantinidis, T., G. Iacovides, M. Xu, T. G. Constantinides, and D. Mandic. 2024. FinLlama: Financial sentiment classification for algorithmic trading applications. Working paper.
Krupa, J., and M. Minutti-Meza. 2022. Regression and machine learning methods to predict discrete outcomes in accounting research. Journal of Financial Reporting 7(2): 131-178.
Kubick, T. R., D. P. Lynch, M. A. Mayberry, and T. C. Omer. 2016. The effects of regulatory scrutiny on tax avoidance: An examination of SEC comment letters. The Accounting Review 91(6): 1751-1780.
Küster, S., T. Steindl, and M. Göttsche. 2025. The informational content of key audit matters: Evidence from using artificial intelligence in textual analysis. *Contemporary Accounting Research*, forthcoming.
Li, F. 2010. The information content of forward-looking statements in corporate filings—A naive Bayesian machine learning approach. Journal of Accounting Research 48(5).
Li, S., and E. T. Nwaeze. 2015. The association between extensions in XBRL disclosures and financial information environment. Journal of Information Systems 29(3): 73-99.
Li, S., and E. T. Nwaeze. 2018. Impact of extensions in XBRL disclosures on analysts' forecast behavior. Accounting Horizons 32(2): 57-79.
Lopez-Lira A, and Y. Tang. 2024. Can ChatGPT forecast stock price movements? Return predictability and large language models. Working paper.
Loughran, T., and B. McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance 66(1): 35-65.
Loughran, T., and B. McDonald. 2016. Textual analysis in accounting and finance: A survey. Journal of Accounting Research 54(4): 1187-1230.
Lowry, M., R.Michaely, and E. Volkova. 2020. Information revealed through the regulatory process: Interactions between the SEC and companies ahead of their IPO. Review of Financial Studies 33(12); 5510-5554.
Luo, W., D. Gong, and X. Jiaotong. 2024. Pre-trained large language models for financial sentiment analysis. Working paper.
Mayew, W. J., M. Sethuraman, and M. Venkatachalam. 2015. MD&A disclosure and the firm's ability to continue as a going concern. The Accounting Review 90(4): 1621-1651.
McGuire, S. T., T. C. Omer, D. Wang. 2012. Tax avoidance: Does tax-specific industry expertise make a difference? The Accounting Review 87(3).
Nylen, J. A., D. D. Wangerin, and K. M. Zehms. 2025. Do key audit matter disclosures about M&A transactions predict future performance? *Review of Accounting Studies*, forthcoming.
Pavlynshenko, B. M. 2023. Financial news analytics using fine-tuned Llama 2 GPT model. Working paper.
Rouen, E., K. Sachdeva, and A. Yoon. 2024. Sustainability meets substance: Evaluating ESG reports in the context of 10-Ks and firm performance. Working paper.
Ryans, J. P. 2021. Textual classification of comment letters. Review of Accounting Studies 26(1): 37-80.
Securities and Exchange Commission (SEC). 2014. Staff observations of custom tag rates. Available at: www.sec.gov/structureddata/announcement/osd-announcement-070714-staff-observations-custom-axis-tag#_ftnref19.

Siano, F. 2025. The news in earnings announcement disclosures: Capturing word context using LLM methods. Management Science, forthcoming.
Standridge, K, A. M. Allen, and J. Thornock. 2023. The (in)adequacy of legal contingency reporting. Working paper.
United Nations Educational, Scientific and Cultural Organization (UNESCO). 2024. Small language models (SLMs): A cheaper greener route into AI. March 18. Available at: https://www.unesco.org/en/articles/small-language-models-slms-cheaper-greener-route-ai#:~:text=SLMs%20are%20well%20recognized%20for,reinforce%20their%20commitment%20to%20sustainability.

Appendix A – Accounting topics

Accounting Topic from FASB Codification	XBRL TextBlock Taxonomy Tag
Asset Retirement	AssetRetirementObligationDisclosureTextBlock
Business Combinations	BusinessCombinationDisclosureTextBlock
	MergersAcquisitionsAndDispositionsDisclosuresTextBlock
	BusinessAcquisitionIntegrationRestructuringAndOtherRelatedCostsTextBlock
Cash	CashAndCashEquivalentsDisclosureTextBlock
	CashCashEquivalentsAndMarketableSecuritiesTextBlock
	CashCashEquivalentsAndShortTermInvestmentsTextBlock
Collaborative Arrangements	CollaborativeArrangementDisclosureTextBlock
Commitments and Contingencies	CommitmentsDisclosureTextBlock
	CommitmentsAndContingenciesDisclosureTextBlock
	CommitmentsContingenciesAndGuaranteesTextBlock
	LegalMattersAndContingenciesTextBlock
	LossContingencyDisclosures
Compensation	CompensationAndEmployeeBenefitPlansTextBlock
	CompensationRelatedCostsGeneralTextBlock
	CompensationAndEmployeeBenefitPlansOtherThanShareBasedCompensationTextBlock
	DisclosureOfCompensationRelatedCostsShareBasedPaymentsTextBlock
	ShareholdersEquityAndShareBasedPaymentsTextBlock
	DisclosureOfShareBasedCompensationArrangementsByShareBasedPaymentAwardTextBlock
Consolidation	MinorityInterestDisclosureTextBlock
	VariableInterestEntityDisclosureTextBlock
Debt	DebtDisclosureTextBlock
	LongTermDebtTextBlock
	ShortTermDebtTextBlock
	FederalHomeLoanBankAdvancesDisclosureTextBlock
	SubordinatedBorrowingsDisclosureTextBlock
Deferred Revenue	DeferredRevenueDisclosureTextBlock
Derivatives and Hedging	DerivativeInstrumentsAndHedgingActivitiesDisclosureTextBlock
	FinancialInstrumentsDisclosureTextBlock

	DerivativesAndFairValueTextBlock
Equity	StockholdersEquityNoteDisclosureTextBlock PreferredStockTextBlock TreasuryStockTextBlock PartnersCapitalNotesDisclosureTextBlock
Fair Value	FairValueDisclosuresTextBlock FairValueMeasurementInputsDisclosureTextBlock
Financial Services	RegulatoryCapitalRequirementsUnderBankingRegulationsTextBlock
Goodwill and Intangibles	GoodwillAndIntangibleAssetsDisclosureTextBlock GoodwillDisclosureTextBlock IntangibleAssetsDisclosureTextBlock
Insurance	ReinsuranceTextBlock
Inventory	InventoryDisclosureTextBlock
Investments	InvestmentsInDebtAndMarketableEquitySecuritiesAndCertainTradingAssetsDisclosureTextBlock InvestmentTextBlock InvestmentHoldingsTextBlock InvestmentsAndOtherNoncurrentAssetsTextBlock MarketableSecuritiesTextBlock EquityMethodInvestmentsDisclosureTextBlock
Leases	LeasesOfLesseeDisclosureTextBlock DebtAndCapitalLeasesDisclosuresTextBlock LeasesOfLessorDisclosureTextBlock OperatingLeasesOfLesseeDisclosureTextBlock CapitalLeasesInFinancialStatementsOfLesseeDisclosureTextBlock LesseeOperatingLeasesTextBlock OperatingLeasesOfLessorDisclosureTextBlock LesseeFinanceLeasesTextBlock
Liabilities	AccountsPayableAndAccruedLiabilitiesDisclosureTextBlock DepositLiabilitiesDisclosuresTextBlock OtherLiabilitiesDisclosureTextBlock AccountsPayableAccruedLiabilitiesAndOtherLiabilitiesDisclosureCurrentTextBlock
Oil and Gas	OilAndGasExplorationAndProductionIndustriesDisclosuresTextBlock

Other Assets	OtherAssetsDisclosureTextBlock OtherCurrentAssetsTextBlock
Pension and Post-employment	PensionAndOtherPostretirementBenefitsDisclosureTextBlock PostemploymentBenefitsDisclosureTextBlock
PPE	PropertyPlantAndEquipmentDisclosureTextBlock DisposalGroupsIncludingDiscontinuedOperationsDisclosureTextBlock
Real Estate	RealEstateDisclosureTextBlock RealEstateOwnedTextBlock RealEstateAndAccumulatedDepreciationDisclosureTextBlock
Receivables	LoansNotesTradeAndOtherReceivablesDisclosureTextBlock FinancingReceivablesTextBlock AllowanceForCreditLossesTextBlock
Related Party	RelatedPartyTransactionsDisclosureTextBlock
Research and Development	ResearchDevelopmentAndComputerSoftwareDisclosureTextBlock
Restructuring	AssetImpairmentChargesTextBlock RestructuringImpairmentAndOtherActivitiesDisclosureTextBlock RestructuringAndRelatedActivitiesDisclosureTextBlock
Revenue	LongTermContractsOrProgramsDisclosureTextBlock RevenueFromContractWithCustomerTextBlock
Tax	IncomeTaxDisclosureTextBlock
Transfers and Servicing	TransfersAndServicingOfFinancialAssetsTextBlock RepurchaseAgreementsResaleAgreementsSecuritiesBorrowedAndSecuritiesLoanedDisclosureTextBlock
Warranties and Guarantees	ProductWarrantyDisclosureTextBlock GuaranteesTextBlock

Appendix B – Technical details for our model

In this Appendix, we provide further detail on our process for fine-tuning Llama-2-7B for the new downstream prediction task of topic classification. Our training and evaluation dataset includes textual data (notes to the financial statements) and the corresponding topic label (XBRL taxonomy tag). The objective at this stage is to fine-tune the model to predict the probability that a note belongs to a given topic.

The model is trained on the text and labels, adjusting its parameters for optimal prediction. We used a single A100 80GB GPU that is part of a computer cluster to create the model, with a learning rate of 2e-5. Learning rate is a hyperparameter that controls how much to change the model in response to the estimated error each time the model weights are updated. Each step has a batch size of 8 observations. We use a cosine learning rate scheduler and a weight decay of 0.1. A cosine learning rate scheduler gradually decreases the learning rate following a cosine curve, promoting smoother convergence during training. Specifically, the model makes larger changes to its weights at the beginning of the finetuning process when it has more to learn and gradually decreases the magnitude of these changes. Weight decay regularization technique penalizes large weights by adding a small penalty term to the loss function, which helps prevent overfitting. Model evaluation was conducted every 2,000 steps with a batch size of 8 notes. We utilized the "Weights & Biases" machine learning platform to monitor the model's performance. This platform enables developers to build more efficient models and visually assess their results.

A common concern in machine learning is overfitting, meaning that the model performs well within the training/fine-tuning data, but will not perform well when seeing new data. Specifically, as training progresses, a model can start to memorize the training data, leading to poor generalization on unseen data. To address this, we used Early Stopping as a regularization technique. This technique is often used during the training of machine learning models and helps prevent overfitting by stopping the training process when the performance ceases to improve (Yao et al. 2007). Early Stopping also optimizes training time by allowing for more efficient use of computational resources (Chen et al. 2023). We set maximum epochs to three. An epoch is one complete pass through the entire training dataset. In our case, one epoch = 40,169 steps (train_samples/batch_size = 321,350/8). Our training stopped at 24,000 steps (24,000*8 = 192,000 samples), or 0.6 epoch. We ran another experiment forcing the model to run two complete epochs and discovered that the best results were around 0.6 epochs; beyond that, the model experienced significant overfitting.

Overall, the model revealed the best results after 24,000 steps utilizing 192,000 observations (notes), which is 60 percent of the overall training sample. The overall accuracy when applied to the evaluation sample was around 95 percent.

Appendix C – Variable definitions

Variable Name	Definition
Test variables	
TAX-NEGTONE	Percentage of negative sentiment words relative to total words on income taxes in the MD&A
TAX-LENGTH	Percentage of words on income taxes in the MD&A
DEBT-EQUITY-NEGTONE	Percentage of negative sentiment words relative to total words on debt or equity in the MD&A
DEBT-EQUITY-LENGTH	Percentage of words on debt or equity in the MD&A
CC-NEGTONE	Percentage of negative sentiment words relative to total words on commitments and contingencies in the MD&A
CC-LENGTH	Percentage of words on commitments and contingencies in the MD&A
REVENUE-NEGTONE	Percentage of negative sentiment words relative to total words on revenues in the MD&A
REVENUE-LENGTH	Percentage of words on revenues in the MD&A
Dependent variables	
CL	1 if the firm subsequently receives an SEC comment letter that comments on a financial reporting issue in the 10-K filing, and 0 otherwise
MISSTATE	1 if the 10-K filing contains a misstatement, and 0 otherwise
CAR [0,2]	Three-day cumulative market-adjusted abnormal returns (CAR) beginning with the 10-K filing date [0, 2]
CAR [0,3]	Four-day cumulative market-adjusted abnormal returns (CAR) beginning with the 10-K filing date [0, 3]
CL-TAX	1 if the firm subsequently receives an SEC comment letter that comments on tax in the 10-K filing, and 0 otherwise.
CL-DEBT_EQUITY	1 if the firm subsequently receives an SEC comment letter that comments on debt and equity in the 10-K filing, and 0 otherwise.
CL-CC	1 if the firm subsequently receives an SEC comment letter that comments on commitments and contingencies in the 10-K filing, and 0 otherwise.
CL-REVENUE	1 if the firm subsequently receives a SEC comment letter that comments on revenue in the 10-K filing, and 0 otherwise.
MISSTATE-TAX	1 if the firm subsequently announces a tax-related restatement for the 10-K filings, and 0 otherwise
MISSTATE-DEBT-EQUITY	1 if the firm subsequently announces a debt or equity-related restatement for the 10-K filings, and 0 otherwise
MISSTATE-CC	1 if the firm subsequently announces a commitments and contingencies-related restatement for the 10-K filings, and 0 otherwise
MISSTATE-REVENUE	1 if the firm subsequently announces a revenue-related restatement for the 10-K filings, and 0 otherwise
Control variables	
MDA-NEG	Percentage of negative sentiment words relative to total words in the MD&A section
MDA-LENGTH	Natural log of the total number of words in the MD&A section
ARC	Natural log of the total number of distinct monetary taxonomy XBRL tags in financial statements of the 10-K filings (Hoitash and Hoitash 2018 – www.xbrlresearch.com)
MV	Natural log of market value of equity
AGE	Natural log of the number of years for which total assets are reported in Compustat.
BIG4	1 if Big 4 auditor, and 0 otherwise
SALES_GROWTH	Current sales minus lagged sales divided by lagged sales
SEGMENTS	Natural log of 1 plus the total number of segments from the Compustat WRDS_SEGMERGED dataset
LEV	Total debt divided by total assets

LOSS	1 if income before extraordinary items is negative, and 0 otherwise
Z_SCORE	Altman's Z-score is measured following DeFond and Hung (2003) and Altman (1968)
MA	1 if issuer reports non-zero pre-tax earnings impact of mergers or acquisitions, and 0 otherwise
RESTRUCTURE	1 if issuer reports non-zero pre-tax restructuring costs, and 0 otherwise
UTB	Total uncertain tax benefits (TXTUBEND) multiplied by 100, divided by total assets
EXT-FIN	Sale of common and preferred stock (SSTK) plus issuance of long-term debt (DLTIS) divided by total assets
CONTINGENCIES	Contingent liabilities – guarantees (CLG) multiplied by 100, divided by total assets
DEF-REVENUES	Total deferred revenue (DRC + DRLT) divided by total assets
BTM	Book to market ratio, calculated as the book value of equity divided by market value of equity.
TURNOVER	Prefile date trading activities, calculated as the natural log of the cumulative daily turnover in trading days [−252, −6] relative to the 10-K filing date. At least 60 observations of daily volume must be available to be included in the sample.
PRE-FFALPHA	Prefile date Fama–French alpha, estimated from a regression of their three-factor model using trading days [−252, −6] relative to the 10-K filing date. At least 60 observations of daily returns must be available to be included in the sample.
INST-OWN	Percentage of shares owned by institutional investors at the fiscal year end.
NASDAQ	1 if the firm is traded on the NASDAQ, and 0 otherwise.
UE	Unexpected earnings, calculated as the actual EPS for the year minus the last mean consensus analyst forecasted EPS, divided by fiscal year-end stock price.
CONCUR	1 if the earnings announcement date is the same as or one day before the 10-K filing date, and 0 otherwise.
FILE-LAG	Natural log of one plus the number of days between fiscal year end and the filing date of the 10-K.
Additional analysis	
CUSTOMFN-DEBT-EQUITY	1 if the firm has a note with extended XBRL tag that our model classifies as Debt or Equity with at least 90 percent confidence, and 0 otherwise.
FNLENGTH-DEBT-EQUITY	Natural log of the number of words in the footnote on debt or equity.
CUSTOMFN-INVESTMENTS	1 if the firm has a note with extended XBRL tag that our model classifies as Investments with at least 90 percent confidence, and 0 otherwise.
FNLENGTH-INVESTMENTS	Natural log of the number of words in the footnote on investments.
CUSTOMFN-CC	1 if the firm has a note with extended XBRL tag that our model classifies as Commitments and Contingencies with at least 90 percent confidence, and 0 otherwise.
FNLENGTH-CC	Natural log of the number of words in the footnote on commitments and contingencies.
CUSTOMFN-RECEIVABLES	1 if the firm has a note with extended XBRL tag that our model classifies as Receivables with at least 90 percent confidence, and 0 otherwise.
FNLENGTH-RECEIVABLES	Natural log of the number of words in the footnote on receivables.
CL-INVESTMENTS	1 if the firm subsequently receives an SEC comment letter that comments on investments in the 10-K filing, and 0 otherwise.
CL-RECEIVABLES	1 if the firm subsequently receives an SEC comment letter that comments on receivables in the 10-K filing, and 0 otherwise.
INVESTMENTS	Total investment (IVST + IVAEQ + IVAO) divided by total assets.
RECEIVABLES	Total receivables divided by total assets.

Figure A – Confusion Matrix for Investment Topic Prediction

		Predicted Topic Negative (108,976)	**Predicted Topic** Positive (4,204)
Actual Topic	Negative (109,050)	True Negative (TN) 108,797	False Positive (FP) 253
Actual Topic	Positive (4,130)	False Negative (FN) 179	True Positive (TP) 3,951

The above figure shows an example confusion matrix for Table 2 using the Investments topic as an example.

Precision is a percentage calculated as the number of observations correctly classified into a topic scaled by the total number of observations classified into that topic. **Precision is 93.98 percent for the Investment topic.** The calculation is as follows: TP/(TP+FP)=3,951/(3,951+253)=0.9398.

Recall is a percentage calculated as the number of observations correctly classified into a topic scaled by the actual number of observations of that topic. **Recall is 95.67 percent for the Investment topic.** The calculation is as follows: TP/(TP+FN)=3,951/(3,951+179)=0.9567.

The F1 score is the harmonic mean of the two metrics. **F1 score is 94.82 percent for the Investment topic.** The calculation is as follows: 2*(Precision*Recall)/(Precision + Recall)=2*(0.9398*0.9567)/(0.9398+0.9567)=0.9482.

TABLE 1 Frequency of Accounting Topics

Accounting Topic	Full available sample	Training sample	Evaluation sample	Fine-tuning sample	Testing sample
Asset Retirement	1,966	1,341	184	781	441
Business Combinations	21,879	15,225	1,420	9,093	5,234
Cash	2,454	1,648	227	962	579
Collaborative Arrangements	1,639	981	256	594	402
Commitments and Contingencies	44,076	29,540	3,905	17,684	10,631
Compensation	44,538	30,411	3,412	18,215	10,715
Consolidation	4,153	3,012	232	1,837	909
Debt	42,038	28,606	3,308	17,011	10,124
Deferred Revenue	529	358	47	212	124
Derivatives and Hedging	18,618	13,007	903	7,712	4,708
Equity	30,607	20,390	2,932	12,038	7,285
Fair Value	29,563	20,408	1,999	12,248	7,156
Financial Services	3,775	2,632	199	1,567	944
Goodwill and Intangibles	26,971	18,886	1,757	11,334	6,328
Insurance	1,099	787	51	468	261
Inventory	5,262	3,638	400	2,162	1,224
Investments	17,172	12,089	953	7,201	4,130
Leases	14,943	10,075	1,155	6,134	3,713
Liabilities	6,856	4,636	549	2,776	1,671
Oil and Gas	802	493	91	282	218
Other Assets	2,268	1,579	160	948	529
Pension and Post-employment	20,306	14,233	1,106	8,534	4,967
PPE	27,146	18,437	2,243	10,995	6,466
Real Estate	3,632	2,537	171	1,492	924
Receivables	11,186	7,838	676	4,696	2,672
Related Party	20,625	13,332	2,384	7,964	4,909
Research and Development	558	328	56	192	174
Restructuring	10,631	7,673	455	4,574	2,503
Revenue	7,471	5,246	435	3,121	1,790
Tax	43,848	29,372	3,879	17,627	10,597
Transfers and Servicing	1,946	1,409	69	829	468
Warranties and Guarantees	1,681	1,203	94	717	384
Total	470,238	321,350	35,708	192,000	113,180

TABLE 2 Classification Matrix

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**	**(13)**	**(14)**	**(15)**	**(16)**	**(17)**
(1) Asset Retirement	432	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(2) Business Combinations	0	5,039	0	1	6	0	24	0	0	0	4	0	0	16	0	0	37
(3) Cash	0	0	519	0	0	0	0	1	0	5	0	2	0	0	0	0	49
(4) Collaborative Arrangements	0	7	0	348	2	0	1	1	0	0	0	0	0	6	0	0	12
(5) Commitments and Contingencies	3	11	0	11	10,383	10	0	8	0	1	2	0	9	2	0	0	5
(6) Compensation	0	4	0	0	3	9,173	4	0	2	0	590	0	0	0	0	0	0
(7) Consolidation	0	10	0	0	0	0	839	1	0	0	16	0	0	0	0	0	39
(8) Debt	0	1	4	6	6	0	0	9,899	1	2	22	0	0	0	0	0	1
(9) Deferred Revenue	0	0	0	4	0	0	0	0	97	0	0	0	0	1	0	0	0
(10) Derivatives and Hedging	0	0	7	0	30	4	1	2	0	4,267	12	356	0	0	0	0	18
(11) Equity	0	5	0	0	3	177	7	1	0	8	7,042	1	18	0	0	0	6
(12) Fair Value	0	1	18	0	2	1	0	0	0	104	0	7,012	0	0	0	0	10
(13) Financial Services	0	0	0	0	0	0	0	3	0	0	39	0	902	0	0	0	0
(14) Goodwill and Intangibles	0	13	0	21	0	0	0	0	0	0	2	1	0	6,251	0	0	0
(15) Insurance	0	0	0	0	0	0	0	0	0	0	0	0	0	0	261	0	0
(16) Inventory	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,221	1
(17) Investments	0	7	44	0	2	0	14	1	0	11	5	12	0	1	0	7	3,951
(18) Leases	0	0	0	1	69	1	0	177	0	0	0	0	0	6	0	0	0
(19) Liabilities	0	0	5	0	5	0	0	38	1	2	1	0	0	1	0	0	0
(20) Oil and Gas	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(21) Other Assets	0	3	0	0	2	1	0	0	0	0	0	1	0	1	0	5	21
(22) Pension and Post-employment	0	0	0	0	1	521	0	0	0	0	5	0	0	0	0	0	0
(23) PPE	0	39	0	0	8	0	0	0	1	0	1	1	0	1	0	0	4
(24) Real Estate	0	19	0	0	0	0	0	0	0	0	0	1	0	6	0	5	15
(25) Receivables	0	2	0	0	5	0	1	2	0	0	2	0	0	2	0	0	8
(26) Related Party	0	5	0	2	12	4	2	21	0	0	18	0	0	0	1	0	19
(27) Research and Development	0	0	0	32	3	0	0	0	0	0	0	0	0	7	0	0	0
(28) Restructuring	0	16	0	0	1	3	0	0	0	0	0	0	0	9	0	1	3
(29) Revenue	0	0	0	19	3	2	0	0	8	0	0	0	0	0	0	0	0
(30) Tax	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
(31) Transfers and Servicing	0	0	0	0	2	0	13	5	0	0	6	0	0	2	0	0	5
(32) Warranties and Guarantees	0	0	0	0	30	1	0	0	0	4	11	0	0	0	0	0	0
Total	**435**	**5,183**	**597**	**445**	**10,578**	**9,898**	**906**	**10,160**	**111**	**4,404**	**7,778**	**7,387**	**929**	**6,312**	**262**	**1,239**	**4,204**

TABLE 2 columns continued:

	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)	(31)	(32)	Total
(1) Asset Retirement	0	6	0	0	2	1	0	0	0	0	0	0	0	0	0	**441**
(2) Business Combinations	0	0	0	1	0	62	14	0	3	0	27	0	0	0	0	**5,234**
(3) Cash	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	**579**
(4) Collaborative Arrangements	0	1	0	0	0	0	0	0	4	2	1	17	0	0	0	**402**
(5) Commitments and Contingencies	145	1	0	0	2	0	2	2	14	0	1	1	0	0	18	**10,631**
(6) Compensation	0	3	0	2	908	0	0	0	23	0	3	0	0	0	0	**10,715**
(7) Consolidation	0	0	0	0	0	1	0	1	2	0	0	0	0	0	0	**909**
(8) Debt	58	29	0	1	0	0	1	23	40	0	1	2	0	27	0	**10,124**
(9) Deferred Revenue	1	1	0	1	0	1	0	0	0	0	0	18	0	0	0	**124**
(10) Derivatives and Hedging	0	1	0	0	0	0	0	0	0	0	0	0	0	0	10	**4,708**
(11) Equity	0	0	0	0	0	1	0	0	16	0	0	0	0	0	0	**7,285**
(12) Fair Value	0	3	0	0	0	0	0	4	1	0	0	0	0	0	0	**7,156**
(13) Financial Services	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**944**
(14) Goodwill and Intangibles	0	0	0	3	0	3	0	0	4	7	21	0	0	2	0	**6,328**
(15) Insurance	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**261**
(16) Inventory	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	**1,224**
(17) Investments	0	5	0	33	0	5	7	14	10	0	0	0	0	1	0	**4,130**
(18) Leases	3,434	0	0	0	2	7	4	2	5	0	3	2	0	0	0	**3,713**
(19) Liabilities	5	1,580	0	11	0	0	0	7	9	0	3	1	1	1	0	**1,671**
(20) Oil and Gas	0	0	214	3	0	0	0	0	0	0	0	0	0	0	0	**218**
(21) Other Assets	0	3	1	469	1	1	0	14	4	0	1	0	0	1	0	**529**
(22) Pension and Post-employment	0	1	0	0	4,430	0	0	0	1	0	8	0	0	0	0	**4,967**
(23) PPE	6	0	1	5	0	6,335	28	0	3	0	33	0	0	0	0	**6,466**
(24) Real Estate	2	0	0	0	0	10	862	0	1	0	1	2	0	0	0	**924**
(25) Receivables	1	0	0	7	0	0	0	2,628	4	0	0	5	0	5	0	**2,672**
(26) Related Party	2	0	0	0	0	1	1	3	4,810	0	1	1	0	2	4	**4,909**
(27) Research and Development	0	0	0	0	0	0	0	0	1	122	0	9	0	0	0	**174**
(28) Restructuring	1	3	0	1	0	13	1	2	1	0	2,448	0	0	0	0	**2,503**
(29) Revenue	0	1	0	0	0	0	0	6	0	0	0	1,751	0	0	0	**1,790**
(30) Tax	0	1	0	0	0	0	0	0	0	0	0	0	10,595	0	0	**10,597**
(31) Transfers and Servicing	0	0	0	1	0	3	0	12	0	0	0	0	0	419	0	**468**
(32) Warranties and Guarantees	1	1	0	0	0	0	0	1	0	0	0	0	0	4	331	**384**
Total	**3,656**	**1,641**	**216**	**538**	**5,345**	**6,444**	**920**	**2,719**	**4,956**	**131**	**2,553**	**1,809**	**10,596**	**465**	**363**	**113,180**

TABLE 3 Model performance

Accounting Topic	Obs.	Precision	Recall	F1 Score
Asset Retirement	441	99.31%	97.96%	98.63%
Business Combinations	5,234	97.22%	96.27%	96.75%
Cash	579	86.93%	89.64%	88.27%
Collaborative Arrangements	402	78.20%	86.57%	82.17%
Commitments and Contingencies	10,631	98.16%	97.67%	97.91%
Compensation	10,715	92.68%	85.61%	89.00%
Consolidation	909	92.60%	92.30%	92.45%
Debt	10,124	97.43%	97.78%	97.60%
Deferred Revenue	124	87.39%	78.23%	82.55%
Derivatives and Hedging	4,708	96.89%	90.63%	93.66%
Equity	7,285	90.54%	96.66%	93.50%
Fair Value	7,156	94.92%	97.99%	96.43%
Financial Services	944	97.09%	95.55%	96.32%
Goodwill and Intangibles	6,328	99.03%	98.78%	98.91%
Insurance	261	99.62%	100.00%	99.81%
Inventory	1,224	98.55%	99.75%	99.15%
Investments	4,130	93.98%	95.67%	94.82%
Leases	3,713	93.93%	92.49%	93.20%
Liabilities	1,671	96.28%	94.55%	95.41%
Oil and Gas	218	99.07%	98.17%	98.62%
Other Assets	529	87.17%	88.66%	87.91%
Pension and Post-employment	4,967	82.88%	89.19%	85.92%
PPE	6,466	98.31%	97.97%	98.14%
Real Estate	924	93.70%	93.29%	93.49%
Receivables	2,672	96.65%	98.35%	97.50%
Related Party	4,909	97.05%	97.98%	97.52%
Research and Development	174	93.13%	70.11%	80.00%
Restructuring	2,503	95.89%	97.80%	96.84%
Revenue	1,790	96.79%	97.82%	97.30%
Tax	10,597	99.99%	99.98%	99.99%
Transfers and Servicing	468	90.11%	89.53%	89.82%
Warranties and Guarantees	384	91.18%	86.20%	88.62%
Weighted average	**113,180**	**95.54%**	**95.48%**	**95.47%**

TABLE 4 Comparison to other models

Model	Provider	Availability	Weighted average F1 score
Our model: Llama-2-7B (fine-tuned)	Meta	Open source	95.5%
GPT-4o	Open AI	Closed source	86.0%
GPT-4o-mini	Open AI	Closed source	83.5%
Gemini-1.0-pro	Google	Closed source	84.2%
Llama-2-7B (untrained)	Meta	Open source	51.7%
LDA	N/A	N/A	72.0%

TABLE 5 Research Applications - Sample construction and descriptives

Panel A: Sample construction

Firm-year observations with note data from Table 1	47,152
Less: Firm-years without Audit Analytics data	(16)
Less: Firm-years without Compustat data	(149)
Less: Firm-years before 2012 or after 2020	(8,924)
Less: Missing data for control variables in Model (1)	(1,560)
Less: Firm-years without MD&A data	(282)
	36,221

Panel B: Firm-years with at least one MD&A paragraph on listed topic

	Predicted Labels	**Number of firm-years**
1	Tax	32,052
2	Debt	30,890
3	Equity	26,679
4	Commitments and Contingencies	25,745
5	Revenue	24,821
6	Fair Value	22,726
7	Investments	22,657
8	Goodwill and Intangibles	21,429
9	Restructuring	20,080
10	Compensation	19,689
11	Receivables	18,602
12	PPE	15,708
13	Business Combinations	15,577
14	Cash	15,547
15	Derivatives and Hedging	15,260
16	Inventory	12,908
17	Leases	12,214
18	Consolidation	9,573
19	Research and Development	9,196
20	Warranties and Guarantees	8,266

TABLE 5 (continued)

Panel C: Descriptive statistics

	N	Mean	Median	Std Dev	25th Pctl	75th Pctl
Test variables						
TAX-NEGTONE	36,221	0.94%	0.82%	0.89%	0.20%	1.40%
TAX-LENGTH	36,221	5.36%	4.55%	4.52%	1.95%	7.72%
DEBT-EQUITY-NEGTONE	36,221	0.62%	0.47%	0.70%	0.10%	0.88%
DEBT-EQUITY-LENGTH	36,221	8.25%	6.91%	6.44%	3.62%	11.48%
CC-NEGTONE	36,221	1.87%	1.10%	2.25%	0.00%	3.22%
CC-LENGTH	36,221	2.64%	1.33%	4.09%	0.00%	3.44%
REVENUE-NEGTONE	36,221	0.39%	0.00%	0.67%	0.00%	0.59%
REVENUE-LENGTH	36,221	3.49%	1.73%	4.73%	0.00%	5.18%

Note – For ease of interpretation, we display test variables in their percent form in this Panel.

	N	Mean	Median	Std Dev	25th Pctl	75th Pctl
Dependent variables						
CL-TAX	36,221	0.024	0.000	0.154	0.000	0.000
CL-DEBT_EQUITY	36,221	0.013	0.000	0.112	0.000	0.000
CL-CC	36,221	0.015	0.000	0.122	0.000	0.000
CL-REVENUE	36,221	0.033	0.000	0.178	0.000	0.000
MISSTATE-TAX	36,221	0.015	0.000	0.120	0.000	0.000
MISSTATE-DEBT-EQUITY	36,221	0.011	0.000	0.103	0.000	0.000
MISSTATE-CC	36,221	0.001	0.000	0.031	0.000	0.000
MISSTATE-REVENUE	36,221	0.015	0.000	0.123	0.000	0.000
CAR [0,2]	25,224	-0.002	-0.001	0.067	-0.023	0.022
CAR [0,3]	25,224	-0.002	-0.001	0.073	-0.026	0.025
Control variables						
MDA-NEG	36,221	0.013	0.012	0.005	0.009	0.015
MDA-LENGTH	36,221	9.072	9.115	0.652	8.753	9.451
ARC	36,221	5.642	5.670	0.414	5.366	5.943
MV	36,221	6.393	6.534	2.383	4.746	8.043
AGE	36,221	2.770	2.944	0.860	2.079	3.367
BIG4	36,221	0.630	1.000	0.483	0.000	1.000
SALES-GROWTH	36,221	0.181	0.051	0.797	-0.042	0.173
SEGMENTS	36,221	1.596	1.609	0.585	1.099	2.079
LEV	36,221	0.297	0.211	0.482	0.052	0.416
LOSS	36,221	0.355	0.000	0.478	0.000	1.000
ZSCORE	36,221	1.980	1.780	14.469	0.293	3.983
MA	36,221	0.430	0.000	0.495	0.000	1.000
RESTRUCTURE	36,221	0.284	0.000	0.451	0.000	1.000
UTB	36,221	0.611	0.000	1.435	0.000	0.544
EXT-FIN	36,221	0.208	0.060	0.381	0.005	0.260
CONTINGENCIES	36,221	0.074	0.000	0.357	0.000	0.000
DEF-REVENUES	36,221	0.036	0.000	0.087	0.000	0.024
BTM	25,224	0.508	0.429	0.652	0.213	0.735
TURNOVER	25,224	0.470	0.482	0.821	0.017	0.936
PRE-FFALPHA	25,224	0.000	0.000	0.002	-0.001	0.001
INST-OWN	25,224	0.662	0.746	0.277	0.490	0.886
NASDAQ	25,224	0.512	1.000	0.500	0.000	1.000
UE	25,224	-0.004	0.000	0.066	-0.002	0.003
CONCUR	25,224	0.345	0.000	0.475	0.000	1.000
FILE-LAG	25,224	4.106	4.094	0.210	3.989	4.249

TABLE 6 Aggregate MD&A Analysis

	(1) *CL*	(2) *MISSTATE*	(3) *CAR [0,2]*	(4) *CAR [0,3]*
MDA-NEG	-5.839	-2.397	-0.196*	-0.192*
	(-1.541)	(-0.358)	(-1.901)	(-1.729)
MDA-LENGTH	0.041	0.049	0.000	0.000
	(1.246)	(1.074)	(0.302)	(0.143)
ARC	0.252***	0.721***	0.001	0.002
	(3.567)	(6.056)	(0.652)	(1.074)
MV	0.255***	-0.060***	0.000	0.000
	(20.122)	(-3.157)	(0.916)	(0.338)
AGE	-0.125***	-0.184***	-0.000	-0.000
	(-5.213)	(-4.725)	(-0.610)	(-0.468)
BIG4	0.098*	-0.001	0.003**	0.004**
	(1.866)	(-0.013)	(2.287)	(2.525)
SALES-GROWTH	0.076***	0.029	-0.000	-0.000
	(3.386)	(1.123)	(-0.274)	(-0.347)
SEGMENTS	0.164***	0.229***	-0.000	-0.000
	(4.205)	(3.461)	(-0.475)	(-0.304)
LEV	0.156***	0.123**	-0.002	-0.003
	(3.724)	(2.259)	(-0.748)	(-0.905)
LOSS	0.261***	0.180***	-0.002	-0.003**
	(5.748)	(2.856)	(-1.462)	(-2.085)
ZSCORE	0.004***	0.006**	0.000*	0.000**
	(2.667)	(2.302)	(1.906)	(1.965)
MA	0.147***	0.125**	-0.001	-0.001
	(3.869)	(2.129)	(-1.323)	(-1.447)
RESTRUCTURE	0.106***	0.169***	-0.000	-0.000
	(2.604)	(2.713)	(-0.386)	(-0.200)
BTM			-0.000	-0.000
			(-0.158)	(-0.242)
TURNOVER			-0.002***	-0.002***
			(-2.628)	(-1.966)
PRE-FFALPHA			0.067	0.067
			(0.180)	(-0.935)
INST-OWN			0.007***	0.007***
			(3.306)	(3.074)
NASDAQ			0.001	0.001
			(0.922)	(1.129)
UE			0.032**	0.028*
			(2.251)	(1.849)
CONCUR			-0.002**	-0.003**
			(-2.094)	(-2.276)
FILE-LAG			-0.009***	-0.009***
			(-2.579)	(-2.672)
Constant	-4.549***	-5.926***	0.022	0.021
	(-11.321)	(-9.655)	(1.211)	(1.073)
Industry and year fixed effects	Included	Included	Included	Included
Observations	36,221	36,221	25,224	25,224
Adjusted R2	N/A	N/A	0.019	0.022
ROC	0.748	0.650	N/A	N/A

TABLE 7 Topic-specific MD&A analysis – Accounting topic length and tone, regulatory scrutiny, and financial reporting quality

Panel A: Regulatory scrutiny

	(1) *CL-TAX*	(2) *CL-DEBT-EQUITY*	(3) *CL-CC*	(4) *CL-REVENUE*
TAX-NEG	10.848**			
	(2.328)			
TAX-LENGTH	4.101***			
	(5.240)			
DEBT-EQUITY-NEG		-10.496		
		(-1.223)		
DEBT-EQUITY-LENGTH		4.152***		
		(6.183)		
CC-NEG			9.137***	
			(4.960)	
CC-LENGTH			3.242***	
			(4.582)	
REVENUE-NEG				8.675*
				(1.872)
REVENUE-LENGTH				2.163***
				(3.364)
MDA-NEG	0.062	-26.114*	3.198	-42.091***
	(0.007)	(-1.959)	(0.331)	(-5.084)
MDA-LENGTH	-0.007	0.232**	-0.050	0.161***
	(-0.095)	(2.321)	(-0.709)	(2.749)
ARC	0.658***	0.287	0.288	0.152
	(3.927)	(1.321)	(1.540)	(1.093)
MV	0.174***	0.145***	0.220***	0.227***
	(6.288)	(3.894)	(7.139)	(9.954)
AGE	-0.155***	-0.216***	-0.034	-0.111***
	(-2.993)	(-3.365)	(-0.539)	(-2.651)
BIG4	-0.008	-0.058	-0.034	-0.036
	(-0.063)	(-0.418)	(-0.239)	(-0.372)
SALES-GROWTH	0.021	0.062	0.089	0.095***
	(0.406)	(1.219)	(1.547)	(2.823)
SEGMENTS	0.408***	0.051	0.189**	0.068
	(4.650)	(0.463)	(1.985)	(0.937)
LEV	0.146	0.208***	0.141*	-0.037
	(1.602)	(2.642)	(1.885)	(-0.407)
LOSS	-0.091	0.491***	0.424***	0.350***
	(-0.861)	(3.803)	(3.582)	(4.146)
ZSCORE	0.009**	0.005	0.002	0.004
	(2.284)	(1.531)	(0.539)	(1.400)
MA	0.097	0.212**	0.152	0.119*
	(1.124)	(2.025)	(1.545)	(1.693)
RESTRUCTURE	0.286***	-0.101	0.113	0.076
	(3.227)	(-0.808)	(1.080)	(1.029)
UTB	0.043*	-0.043	-0.026	0.005
	(1.760)	(-1.018)	(-0.763)	(0.279)
EXT-FIN	-0.070	0.207*	0.161	-0.193*
	(-0.539)	(1.857)	(0.972)	(-1.771)
CONTINGENCIES	-0.012	0.055	0.030	-0.026
	(-0.151)	(0.531)	(0.311)	(-0.297)
DEF-REVENUES	-0.747	-0.772	-0.030	0.894***
	(-1.307)	(-1.158)	(-0.046)	(2.822)

Constant	-8.631***	-8.159***	-6.888***	-6.471***
	(-9.521)	(-7.073)	(-6.669)	(-8.605)
Industry and year fixed effects	Included	Included	Included	Included
Observations	36,221	36,221	36,221	36,221
ROC	0.819	0.777	0.793	0.742

Panel B: Financial reporting quality

	(1) *MISSTATE-TAX*	(2) *MISSTATE-DEBT-EQUITY*	(3) *MISSTATE-CC*	(4) *MISSTATE-REVENUE*
TAX-NEG	-0.370			
	(-0.050)			
TAX-LENGTH	2.041			
	(1.578)			
DEBT-EQUITY-NEG		5.861		
		(0.769)		
DEBT-EQUITY-LENGTH		0.802		
		(0.910)		
CC-NEG			13.717	
			(1.381)	
CC-LENGTH			6.545**	
			(2.062)	
REVENUE-NEG				16.303**
				(2.419)
REVENUE-LENGTH				1.955*
				(1.754)
MDA-NEG	19.077	-58.647***	26.326	-6.528
	(1.333)	(-3.306)	(0.569)	(-0.421)
MDA-LENGTH	0.250**	-0.135	-0.305	0.020
	(2.089)	(-1.484)	(-1.070)	(0.216)
ARC	0.789***	1.294***	1.247	0.587**
	(2.678)	(4.757)	(1.047)	(2.250)
MV	-0.115***	-0.044	0.199	-0.023
	(-2.966)	(-0.964)	(1.110)	(-0.545)
AGE	-0.152*	-0.476***	-0.455*	-0.052
	(-1.733)	(-5.367)	(-1.682)	(-0.565)
BIG4	-0.108	-0.013	-1.506***	-0.195
	(-0.618)	(-0.080)	(-2.651)	(-1.045)
SALES-GROWTH	0.047	-0.052	0.240	0.042
	(0.621)	(-0.804)	(1.417)	(0.770)
SEGMENTS	0.344**	-0.226	0.845***	0.193
	(2.247)	(-1.543)	(3.034)	(1.358)
LEV	-0.139	0.165*	-1.230	-0.012
	(-0.581)	(1.876)	(-1.053)	(-0.088)
LOSS	-0.171	0.488***	0.134	-0.064
	(-1.179)	(3.334)	(0.221)	(-0.431)
ZSCORE	0.018**	-0.003	-0.005	0.008**
	(2.061)	(-0.982)	(-0.398)	(1.961)
MA	0.348***	0.042	1.896***	0.472***
	(2.853)	(0.306)	(3.253)	(3.402)
RESTRUCTURE	0.308**	0.167	0.639*	0.231*
	(2.290)	(1.071)	(1.908)	(1.838)
UTB	-0.091*	-0.023	-0.202*	-0.109*
	(-1.724)	(-0.377)	(-1.781)	(-1.780)
EXT-FIN	0.015	0.120	-0.172	-0.143
	(0.086)	(1.224)	(-0.234)	(-0.664)
CONTINGENCIES	-0.133	-0.725***	-4.289	-0.120

	(-0.792)	(-2.728)	(-1.276)	(-0.821)
DEF-REVENUES	1.348*	-1.126	1.156	2.659***
	(1.659)	(-1.151)	(0.289)	(4.452)
Constant	-9.830***	-7.364***	-14.254*	-7.019***
	(-6.518)	(-4.895)	(-1.914)	(-4.811)
Industry and year fixed effects	Included	Included	Included	Included
Observations	36,221	36,221	25,911	36,221
ROC	0.733	0.743	0.904	0.698

TABLE 8 Topic-specific MD&A analysis – accounting topic length and tone and market reaction

Panel A: Multivariate analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	CAR [0,2]	*CAR [0,3]*	*CAR [0,2]*	*CAR [0,3]*	*CAR [0,2]*	*CAR [0,3]*	*CAR [0,2]*	*CAR [0,3]*
TAX-NEG	0.052	0.060						
	(0.833)	(0.864)						
TAX-LENGTH	-0.001	-0.004						
	(-0.050)	(-0.359)						
DEBT-EQUITY-NEG			-0.043	-0.070				
			(-0.530)	(-0.801)				
DEBT-EQUITY-LENGTH			-0.004	-0.006				
			(-0.468)	(-0.599)				
CC-NEG					-0.042**	-0.065***		
					(-1.977)	(-2.762)		
CC-LENGTH					-0.004	-0.003		
					(-0.382)	(-0.241)		
REVENUE-NEG							-0.071	-0.016
							(-1.111)	(-0.230)
REVENUE-LENGTH							0.010	0.001
							(0.778)	(0.098)
MDA-NEG	-0.204**	-0.201*	-0.186*	-0.178	-0.114	-0.077	-0.167	-0.181
	(-1.962)	(-1.802)	(-1.774)	(-1.592)	(-1.053)	(-0.651)	(-1.612)	(-1.610)
MDA-LENGTH	0.000	-0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.091)	(-0.061)	(0.321)	(0.223)	(0.453)	(0.390)	(0.215)	(0.068)
ARC	0.001	0.002	0.001	0.002	0.001	0.002	0.001	0.002
	(0.584)	(1.015)	(0.638)	(1.064)	(0.552)	(0.960)	(0.670)	(1.060)
MV	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.950)	(0.379)	(0.802)	(0.212)	(0.991)	(0.457)	(0.914)	(0.306)
AGE	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
	(-0.701)	(-0.531)	(-0.723)	(-0.570)	(-0.648)	(-0.479)	(-0.678)	(-0.512)
BIG4	0.003**	0.004**	0.003**	0.004**	0.003**	0.004**	0.003**	0.004**
	(2.148)	(2.396)	(2.173)	(2.435)	(2.193)	(2.461)	(2.173)	(2.419)
SALES-GROWTH	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
	(-0.273)	(-0.359)	(-0.284)	(-0.367)	(-0.284)	(-0.372)	(-0.289)	(-0.362)
SEGMENTS	-0.000	-0.000	-0.001	-0.000	-0.000	-0.000	-0.000	-0.000
	(-0.491)	(-0.292)	(-0.527)	(-0.351)	(-0.440)	(-0.225)	(-0.478)	(-0.309)
LEV	-0.001	-0.001	-0.000	-0.001	-0.001	-0.002	-0.000	-0.001
	(-0.188)	(-0.435)	(-0.088)	(-0.284)	(-0.231)	(-0.478)	(-0.164)	(-0.437)
LOSS	-0.002*	-0.003**	-0.002*	-0.003**	-0.002*	-0.003**	-0.002*	-0.003**

	(-1.751)	(-2.359)	(-1.660)	(-2.248)	(-1.722)	(-2.334)	(-1.688)	(-2.254)
ZSCORE	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**
	(2.140)	(2.183)	(2.106)	(2.111)	(2.203)	(2.243)	(2.145)	(2.182)
MA	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	(-1.152)	(-1.293)	(-1.181)	(-1.325)	(-1.160)	(-1.291)	(-1.159)	(-1.319)
RESTRUCTURE	-0.000	-0.000	-0.000	0.000	-0.000	-0.000	-0.000	0.000
	(-0.181)	(-0.000)	(-0.154)	(0.010)	(-0.173)	(-0.010)	(-0.115)	(0.012)
UTB	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
	(-0.969)	(-0.906)	(-0.976)	(-0.937)	(-0.901)	(-0.835)	(-0.972)	(-0.912)
EXT-FIN	-0.002	-0.001	-0.002	-0.001	-0.002	-0.001	-0.002	-0.001
	(-1.249)	(-0.943)	(-1.191)	(-0.854)	(-1.235)	(-0.907)	(-1.242)	(-0.922)
CONTINGENCIES	0.002*	0.002	0.002*	0.002	0.002**	0.002*	0.002**	0.002
	(1.944)	(1.595)	(1.944)	(1.598)	(1.990)	(1.652)	(1.961)	(1.618)
DEF-REVENUES	0.018***	0.018**	0.018***	0.017**	0.019***	0.018**	0.018**	0.018**
	(2.688)	(2.309)	(2.677)	(2.295)	(2.734)	(2.366)	(2.545)	(2.288)
BTM	0.000	-0.000	0.000	-0.000	0.000	-0.000	0.000	-0.000
	(0.056)	(-0.058)	(0.049)	(-0.064)	(0.025)	(-0.095)	(0.053)	(-0.065)
TURNOVER	-0.002**	-0.002*	-0.002**	-0.002*	-0.002**	-0.002*	-0.002**	-0.002*
	(-2.574)	(-1.928)	(-2.534)	(-1.880)	(-2.516)	(-1.852)	(-2.555)	(-1.903)
PRE-FFALPHA	0.061	-0.381	0.068	-0.372	0.056	-0.389	0.060	-0.377
	(0.166)	(-0.953)	(0.183)	(-0.932)	(0.152)	(-0.975)	(0.162)	(-0.947)
INST-OWN	0.007***	0.007***	0.007***	0.007***	0.007***	0.007***	0.007***	0.007***
	(3.131)	(2.947)	(3.177)	(2.960)	(3.207)	(3.008)	(3.169)	(2.970)
NASDAQ	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	(0.867)	(1.094)	(0.867)	(1.077)	(0.874)	(1.104)	(0.862)	(1.091)
UE	0.031**	0.028*	0.032**	0.028*	0.032**	0.028*	0.032**	0.028*
	(2.209)	(1.812)	(2.211)	(1.812)	(2.218)	(1.821)	(2.212)	(1.820)
CONCUR	-0.002**	-0.003**	-0.002**	-0.003**	-0.002**	-0.003**	-0.002**	-0.003**
	(-2.045)	(-2.239)	(-2.041)	(-2.222)	(-2.033)	(-2.213)	(-2.062)	(-2.249)
FILE-LAG	-0.008**	-0.009***	-0.008**	-0.009***	-0.009**	-0.009***	-0.008**	-0.009***
	(-2.514)	(-2.620)	(-2.517)	(-2.605)	(-2.550)	(-2.660)	(-2.535)	(-2.631)
Constant	0.023	0.021	0.022	0.020	0.021	0.019	0.021	0.020
	(1.238)	(1.107)	(1.185)	(1.029)	(1.142)	(0.972)	(1.148)	(1.046)
Industry and year fixed effects	Included	Included	Included	Included	Included	Included	Included	Included
Observations	25,224	25,224	25,224	25,224	25,224	25,224	25,224	25,224
Adjusted R2	0.020	0.022	0.020	0.022	0.020	0.022	0.020	0.022

TABLE 8 (continued)
Panel B: Within concurrent and non-concurrent samples

	Concurrent		Non-concurrent	
	(1)	(2)	(3)	(4)
	CAR [0,2]	*CAR [0,3]*	*CAR [0,2]*	*CAR [0,3]*
CC-NEG	-0.020	-0.097	-0.046***	-0.046**
	(-0.377)	(-1.637)	(-2.722)	(-2.331)
CC-LENGTH	-0.015	-0.020	0.004	0.010
	(-0.555)	(-0.756)	(0.454)	(0.979)
Controls from Panel A	Included	Included	Included	Included
Observations	8,704	8,704	16,520	16,520
Adjusted R2	0.022	0.022	0.023	0.029

TABLE 9 Additional analysis: Classifying custom-tagged notes

Panel A: Frequency of custom-tagged note topics at firm-year level

	Predicted Labels	Frequency	Percent
1	Equity	6,176	17.53
2	Debt	3,472	9.85
3	Investments	2,161	6.13
4	Commitments and Contingencies	2,054	5.83
5	Receivables	2,034	5.77
6	Compensation	1,680	4.77
7	Business Combination	1,602	4.55
8	Leases	1,394	3.96
9	PPE	1,255	3.56
10	Revenue	1,241	3.52
11	Liabilities	1,186	3.37
12	Consolidation	1,123	3.19
13	Transfers and Servicing	1,040	2.95
14	Restructuring	898	2.55
15	Goodwill and Intangibles	827	2.35
16	Related Party	798	2.26
17	Pension and Post-employment	792	2.25
18	Warranties and Guarantees	782	2.22
19	Other Assets	755	2.14
20	Cash	689	1.96

Panel B: Multivariate results of custom notes and regulatory scrutiny

	(1) *CL_DEBT_EQUITY*	(2) *CL_INVESTMENTS*	(3) *CL_CC*	(4) *CL_RECEIVABLES*
CUSTOMFN-DEBT-EQUITY	0.348***			
	(2.804)			
FNLENGTH-DEBT-EQUITY	0.324***			
	(5.384)			
CUSTOMFN-INVESTMENTS		0.354**		
		(2.153)		
FNLENGTH-INVESTMENTS		0.148***		
		(5.683)		
CUSTOMFN-CC			0.440***	
			(2.672)	
FNLENGTH-CC			0.262***	
			(4.903)	
CUSTOMFN-RECEIVABLES				0.250*
				(1.806)
FNLENGTH-RECEIVABLES				0.063***
				(2.603)
ARC	0.277	0.201	0.037	0.317
	(1.363)	(0.762)	(0.193)	(1.573)
MV	0.116***	0.158***	0.200***	0.171***
	(3.191)	(3.567)	(6.305)	(5.412)
AGE	-0.210***	-0.165**	0.018	-0.135**
	(-3.317)	(-2.075)	(0.283)	(-2.187)
BIG4	-0.076	0.107	-0.055	0.109
	(-0.525)	(0.567)	(-0.390)	(0.866)

SALES-GROWTH	0.039	0.141**	0.076	0.148**
	(0.768)	(1.976)	(1.308)	(2.419)
SEGMENTS	0.037	0.172	0.142	0.283**
	(0.346)	(1.378)	(1.531)	(2.530)
LEV	0.246***	0.347***	0.115	0.126
	(3.290)	(4.014)	(1.532)	(1.158)
LOSS	0.296**	-0.007	0.299**	0.338***
	(2.251)	(-0.043)	(2.498)	(2.758)
ZSCORE	0.008***	0.011**	0.004	0.007
	(2.915)	(2.073)	(0.822)	(1.509)
MA	0.254**	0.184	0.150	0.104
	(2.466)	(1.481)	(1.538)	(1.133)
RESTRUCTURE	-0.162	-0.088	0.097	-0.084
	(-1.343)	(-0.561)	(0.952)	(-0.714)
EXT-FIN	0.191*	-0.063	0.179	0.080
	(1.736)	(-0.449)	(1.176)	(0.802)
INVESTMENTS	-0.201	0.893***	-0.674**	0.803***
	(-0.684)	(3.098)	(-2.156)	(3.058)
CONTINGENCIES	0.084	-0.144	0.030	-0.030
	(0.824)	(-1.157)	(0.317)	(-0.332)
RECEIVABLES	-1.741***	-0.675*	-0.431	1.723***
	(-4.010)	(-1.720)	(-1.208)	(5.452)
Constant	-7.686***	-7.344***	-7.014***	-7.569***
	(-7.600)	(-5.409)	(-7.260)	(-6.907)
Industry and year fixed effects	Included	Included	Included	Included
Observations	36,503	36,503	36,503	36,503
ROC	0.783	0.850	0.795	0.808