An NLP Framework for Analyzing Corporate Strategic Behavior in the Opioid Industry Documents Archive

Duy Dang Phu and Dang Van Thin
University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam
Vietnam National University Ho Chi Minh City, Vietnam
24520010@gm.uit.edu.vn, thindv@uit.edu.vn

Abstract

The Opioid Industry Documents Archive (OIDA) provides extensive internal corporate records that offer valuable insight into the drivers of the opioid crisis, yet its use in systematic analysis of corporate strategy remains limited. In this study, we propose an NLP-based framework to analyze strategic behavior in large-scale litigation archives, combining relevance filtering and topic modeling with large language model (LLM)-assisted interpretation. Applied to documents from Insys Therapeutics and Mallinckrodt Pharmaceuticals, our approach uncovers systematic differences in corporate strategies and organizational priorities. These results highlight the potential of integrating representation learning and LLMs for large-scale analysis in public health and corporate accountability research.

1 Introduction

The opioid crisis constitutes a severe and ongoing public health emergency, characterized by widespread opioid dependence, high rates of overdose mortality, and profound social and economic consequences. In the wake of extensive litigation against pharmaceutical manufacturers, distributors, pharmacy chains, and consulting firms, millions of internal corporate documents have been made publicly available. These disclosures provide an unprecedented window into corporate decision-making processes and strategic conduct within the opioid industry.

The Opioid Industry Documents Archive (University of California, San Francisco and Johns Hopkins University, n.d.) is a digital repository developed by the University of California, San Francisco and Johns Hopkins University. It contains internal corporate documents, emails, and presentations produced by opioid manufacturers and related organizations. Given its scope and depth, OIDA represents a uniquely valuable data source for research in social science, public health, policy, and law—particularly for studies aimed at analyzing corporate strategic behavior. However, despite its substantial research potential, systematic large-scale computational analyses of corporate strategic behavior within OIDA remain limited. The unstructured nature of the textual data, combined with relatively sparse metadata, complicates comprehensive large-scale analysis.

To address these challenges, this study develops a multi-stage computational framework for analyzing corporate strategic behavior in OIDA. The framework integrates keyword-based retrieval, transformer-based text embeddings, and a K-Nearest Neighbors classifier to refine the selection of strategy-relevant texts. We then apply topic modeling, augmented by large language model –assisted interpretation, to identify and synthesize recurring forms of corporate strategic behavior. We focus on two organizations—Insys Therapeutics and Mallinckrodt Pharmaceuticals—by systematically characterizing and comparing their corporate strategic behavior during the same historical period. As a result, we present a computational framework that integrates large-scale text analysis with LLM-assisted interpretation to enable scalable and reproducible analysis of extensive litigation corpora. Substantively, our findings provide systematic empirical evidence of recurring corporate strategic behavior patterns in the opioid industry.

2 Related Work

Computational text analysis has been widely applied to large-scale corpus exploration, including topic modeling, semantic clustering, sentiment analysis, and named entity recognition. These approaches enable structured analysis of unstructured textual data across scientific, legal, and corporate domains.

Prior work has demonstrated the utility of multi-

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 113–122

stage NLP pipelines for knowledge discovery. For example, (Polpinij et al., 2026) integrates transformer-based embeddings, deep clustering, and relation extraction to construct domain-specific knowledge representations from agricultural research literature. Similarly, (Azher et al., 2025) combines BERTopic and large language models to structure and summarize limitation sections in scientific articles. Beyond scientific corpora, NLP methods have been applied to corporate disclosures and financial documents. Studies such as (Kang and Kim, 2022) and (Faccia et al., 2024) employ sentiment analysis and semantic similarity measures to examine thematic emphasis, disclosure patterns, and linguistic risk signals in corporate reports. These works illustrate how computational methods can surface structured patterns within corporate communication.

Such approaches are particularly relevant for archives like the *Opioid Industry Documents Archive*, which contains internal corporate communications documenting strategic planning, marketing activities, and regulatory positioning. While OIDA has been introduced as a valuable research resource, computational engagement with the archive remains limited. Existing work, such as OIDA-QA (Shen et al., 2025), primarily focuses on benchmark construction. However, an integrated computational framework explicitly designed to identify and characterize corporate strategic behavior in large-scale litigation archives remains absent. This gap motivates the development of the structured NLP framework proposed in this study.

3 Data Description

The dataset is collected from the Opioid Industry Documents Archive, a public repository of internal documents related to the opioid industry. The archive contains various types of documents, including emails, reports, memoranda, and presentation slides. Each document is accompanied by metadata such as document ID, document type, date, author, and other descriptive attributes.

The primary source for textual analysis in this study is the *ocr_text* field, which contains text extracted from scanned original documents using Optical Character Recognition (OCR). This field serves as the main input for NLP analysis. However, due to the nature of OCR-based extraction, the text may contain structural inconsistencies, formatting irregularities, and noise (e.g., broken words, misrecognized characters, or misplaced line breaks). These issues introduce additional preprocessing challenges before conducting downstream NLP tasks.

4 Methodology

All detailed prompt templates, model configurations, and hyperparameters are provided in the Appendix B.

4.1 Data Selection and Preprocessing

This study draws on documents from the Opioid Industry Documents Archive. We selected presentation materials associated with two major opioid manufacturers, Insys Therapeutics and Mallinckrodt Pharmaceuticals, both of which have been centrally implicated in litigation and public investigations related to the U.S. opioid crisis. Focusing on these firms enables a concentrated examination of corporate strategic practices during periods of heightened commercial activity and regulatory scrutiny.

To reduce topical noise and focus on high-level strategic content, we restricted our analysis to presentation documents. Compared to emails or general reports, presentation slides are more likely to summarize strategic planning, business positioning, and key corporate initiatives, making them more suitable for downstream semantic analysis. After collecting the presentation files, we performed text cleaning on the *ocr_text* field to remove OCR-related artifacts and malformed characters. To facilitate embedding generation, each presentation was segmented into text chunks of 300 words with a 50-word overlap between consecutive segments. This chunking strategy was determined through preliminary experiments to balance contextual completeness and embedding quality.

4.2 Data Filtering

To enable topic modeling to focus on strategic-level content, we implemented a two-stage filtering procedure.

Stage 1: Keyword-based Pre-filtering. We first constructed a domain-informed keyword list capturing common corporate strategies. Only text segments containing at least one of these keywords were retained. This step serves two purposes: (1) reducing topical noise and (2) lowering computational costs for subsequent modeling stages.

Stage 2: LLM-assisted Labeling and Similarity-based Propagation. From each company, we randomly sampled 150 text segments and used a large language model with chain-of-thought (Wei et al., 2022) prompting to infer whether each segment was strategically relevant. For text representation, we employed the embedding model *avsolatorio/GIST-Embedding-v0* (Solatorio, 2024), selected to balance computational efficiency and semantic performance.

Using the labeled subset as supervision, we trained a k-nearest neighbors classifier ($k = 5$). The value of k was determined through empirical evaluation, with F1-score used as the primary selection metric. The trained KNN model was subsequently applied to the full dataset, where strategic relevance for each segment was determined via majority voting among its nearest neighbors in the embedding space.

4.3 Topic Modeling

We employed BERTopic (Grootendorst, 2022) to cluster semantically similar text segments into coherent and interpretable topics. To ensure methodological consistency across stages, the same embedding model used during the filtering phase (*avsolatorio/GIST-Embedding-v0*) was retained for topic modeling.

For topic representation, we applied the *KeyBERTInspired* approach to extract representative keywords for each topic. Prior to generating topic representations, additional preprocessing steps were conducted, including stopword removal. This procedure was intended to reduce lexical noise, improve semantic coherence, and enhance the interpretability of the resulting topics. Recent work by Yang et al. (2025) demonstrates that large language models (LLMs) can generate high-quality topical descriptors that align closely with human judgment in topic model evaluation. This finding suggests that LLMs possess substantial potential for extracting structured and meaningful information from document clusters. Motivated by this insight, we operationalized a two-stage LLM-assisted analytical framework by designing the following prompts:

Stage 1: Prompt for Topic-Level Strategy Classification This prompt was developed to determine whether a given topic explicitly reflects one of the predefined corporate strategic categories: *Sales Expansion & Promotion*, *Influence & Narrative Management*, *Regulatory Risk Management & Evasion*, or neither. The model was instructed to rely strictly on explicit textual evidence extracted from representative document excerpts and to provide a justification for its classification decision. This step serves as a filtering mechanism to exclude topics that do not substantively reflect corporate strategies or strategies, thereby improving the validity of subsequent interpretation.

Stage 2: Prompt for Topic Interpretation This prompt was applied exclusively to topics that had been classified as strategy-related in the preceding stage. Its primary objective was to generate a structured and analytically coherent interpretation of each topic, thereby facilitating clearer substantive analysis and supporting subsequent qualitative examination.

During the prompt design process, we observed that the inclusion of certain explicitly strategy-laden terms tended to induce speculative reasoning, leading the model to produce overextended or inferential claims not firmly grounded in the source material. To mitigate this tendency, the prompt was carefully calibrated to constrain interpretive latitude. Specifically, the model was instructed to frame the described actions as observable industry strategies without inferring hidden intentions unless such intentions were directly evidenced in the text. Furthermore, the prompt required that every substantive claim in the generated explanation be traceable to the provided textual excerpts. This evidentiary constraint was intended to reinforce analytical rigor, minimize unwarranted extrapolation, and ensure that interpretations remained firmly anchored in the documentary record.

5 Analysis

After applying the LLM to strategy-classified topics, we obtained the results summarized in Table 1. To better understand the substantive characteristics of the identified topics, we now turn to a company-level qualitative analysis. This allows us to examine how strategic patterns manifest differently across organizational contexts rather than relying solely on aggregate topic counts.

5.1 Insys

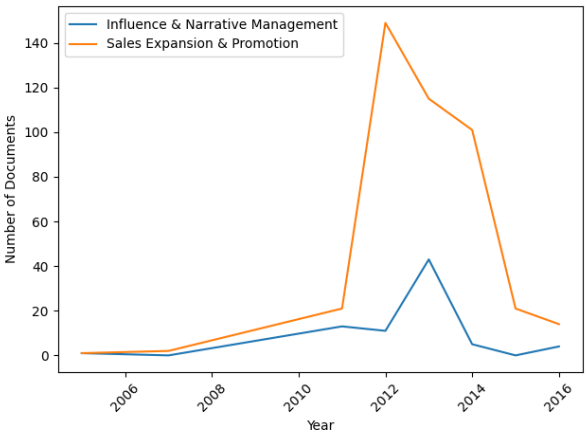
Figure 1 presents the yearly distribution of strategy-related text chunks identified for Insys. The results show that the majority of documents categorized under *Sales Expansion & Promotion* and *Influence & Narrative Management* are concentrated in the

Table 1: Strategy-related topics identified by the LLM. Sales Exp. = Sales Expansion & Promotion; Influence = Influence & Narrative Management; Reg. Risk = Regulatory Risk Management & Evasion; Total = total number of extracted topic groups.

Strategy	Insys	Mallinckrodt
Sales Exp.	24	9
Influence	5	1
Reg. Risk	0	5
Total	51	25

Note: A single topic may be assigned to more than one strategy category.

Figure 1: Number of Document per Strategy Type Over Years



2011–2015 period. Because most strategy-related materials fall within this time frame, our subsequent analysis focuses primarily on these years.

This temporal concentration overlaps with key milestones in the commercial trajectory of *Subsys*, which received approval from the U.S. Food and Drug Administration (FDA) in January 2012 for the treatment of breakthrough cancer pain in opioid-tolerant patients. Publicly available records report substantial revenue growth between 2012 and 2015, with annual revenues reaching approximately $330 million by 2015 (Opioid Industry Documents Archive, n.d.). However, it is important to note that the prominence of strategy-related content during 2011–2015 may also partly reflect the larger overall volume of available documents from this period. In other words, the higher frequency of identified strategies is not necessarily indicative of a proportional increase in strategic activity, but may be influenced by the greater density of archived materials.

To further examine the internal composition of these activities, we identified the five strategy-related topics with the highest number of text documents within the 2011–2015 period and conducted qualitative summaries using LLM-assisted abstraction of the corresponding documents. Although automated summarization may introduce minor semantic imprecision, manual inspection of a subset of documents suggests that the extracted summaries preserve their primary thematic content. Figure 2 displays the quarterly distribution of these five topics.

Substantively, the five topics capture complementary dimensions of commercialization strategy. Topic 3 reflects an integrated, multi-channel marketing framework combining awareness campaigns, targeted outreach, speaker engagement, and medical-education initiatives supported by internal staff and IT infrastructure, oriented toward expanding prescriber adoption and utilization intensity. Topic 10 centers on a structured speaker-program system in which promotional events are financially supported and monitored through prescription-linked performance metrics. Topic 11 describes organizational efforts to enhance sales-force efficiency by reallocating logistical responsibilities to specialized liaisons, thereby increasing representatives' focus on prescriber engagement and market expansion. Topics 15 and 30 both concern performance-based compensation structures, including region-specific sales targets, national adjustment factors, and layered bonus mechanisms designed to align managerial incentives with corporate revenue objectives.

The quarterly patterns suggest differentiated temporal dynamics across topics. Topic 3 appears consistently from early 2011 through late 2014, indicating that the activities captured under this topic were sustained over multiple years rather than confined to a short-term initiative. In contrast, the remaining topics emerge primarily after January 2012. Topics 10 and 11 become visible relatively early in the post-approval phase, particularly between 2012 and early 2013, while Topics 15 and 30 appear more prominently from late 2013 to mid-2014. These patterns may reflect shifts in strategic emphasis during Insys's peak commercialization period.

Taken together, the visualization suggests that most identifiable strategy-related documentation

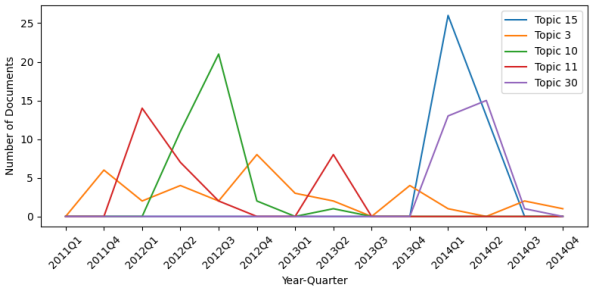


Figure 2: Number of Document per Topic Over Quarter-Years

clusters within 2011–2015, a period that also contains a relatively high volume of overall archival materials. While these observations remain descriptive and do not establish causality, they provide distributional evidence that may reflect evolving organizational priorities during a document-dense phase of product commercialization.

5.2 Mallinckrodt

Similar to the pattern observed for Insys, the majority of strategy-related documents identified in the Mallinckrodt corpus are concentrated in the 2010–2014 period. This temporal clustering spans the years immediately before and after a significant regulatory event: in 2011, Mallinckrodt came under investigation by the Drug Enforcement Administration (DEA) for failing to meet its obligations to monitor and report suspicious orders of controlled substances (Opioid Industry Documents Archive, n.d.). As with Insys, the concentration of identified strategies during this period partly reflects the relatively high volume of available documents from these years. Accordingly, our analysis focuses on 2010–2014, while acknowledging that distributional density may be influenced by archival coverage.

In contrast to the case of Insys, our analysis of Mallinckrodt's corpus reveals the presence of text chunks categorized under *Regulatory Risk Management & Evasion*. Although these documents do not constitute the largest share of strategy-related content, their emergence during a period of heightened regulatory scrutiny is analytically noteworthy. However, a closer qualitative inspection of these topics reveals that they primarily pertain to standard corporate legal and regulatory procedures, rather than explicit strategies for navigating or countering the concurrent DEA investigation. This pattern suggests that while the organization was actively managing its baseline regulatory compliance alongside commercial expansion, the automated extraction did not capture direct strategic responses to the DEA probe within this topic subset. Within the 2010–2014 window, we identified the five strategy-related topics with the highest document counts and conducted qualitative abstraction of their representative texts.

Substantively, the five topics capture complementary dimensions of Mallinckrodt's commercialization strategy during the focal period. Topic 1 reflects sustained efforts to expand prescription volume through sales-force growth, payer segmentation and rebate arrangements to secure formulary access, and the mobilization of key opinion leaders, collectively oriented toward increasing market penetration and revenue attainment. Topic 2 centers on coordinated launch campaigns that integrate sales training, defined performance metrics, cross-product promotion, expanded coverage, and patient-support mechanisms within a structured commercialization framework. Topic 11 describes a comprehensive pre-launch strategy that aligns marketing, medical affairs, and managed-markets functions around pricing, regulatory positioning, payer analytics, and unbranded market-development initiatives. Topic 10 emphasizes managed-care engagement and contract optimization, supported by systematic payer segmentation and financial modeling to balance coverage, profitability, and market share. Topic 15 concerns organizational and channel design decisions, including territory structuring, sales-force deployment, and evaluation of internal versus outsourced commercial models, thereby shaping the operational infrastructure of market entry and expansion.

The temporal distribution of these topics suggests differentiated strategic horizons. Topics 1 and 2 appear over relatively extended intervals, indicating sustained commercial initiatives. Topic 1 spans nearly the entire 2010–2014 period, while Topic 2 is prominent from early 2011 through approximately mid-period, suggesting an extended launch and expansion phase. In contrast, Topics 10, 11, and 15 exhibit more temporally concentrated patterns, consistent with time-bound planning or optimization initiatives.

Taken together, the quarterly distributions indicate a combination of long-term commercialization strategies and shorter-term strategic adjustments during a period characterized by both market expansion and increased regulatory scrutiny. While

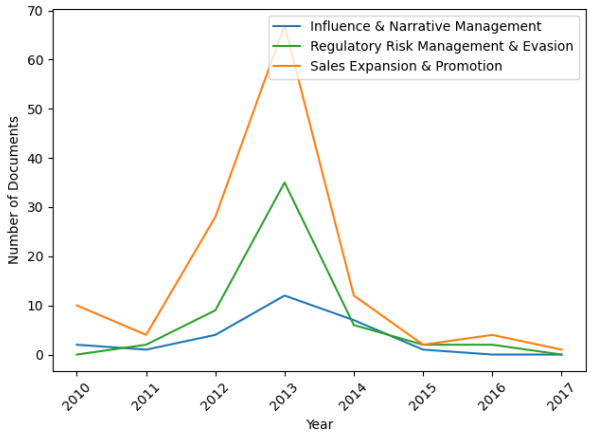


Figure 3: Number of Document per Strategy Type Over Years

these observations remain descriptive and do not establish causal relationships, they provide distributional evidence of evolving strategic emphases within Mallinckrodt's documented activities during 2010–2014.

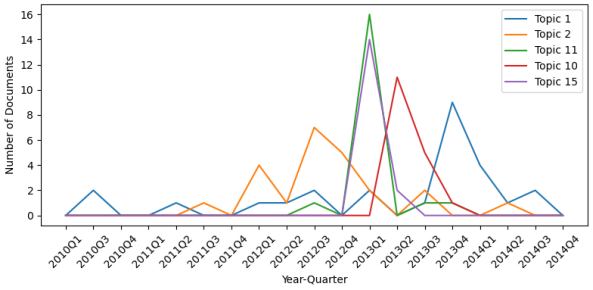


Figure 4: Number of Document per Topic Over Quarter-Years

5.3 Comparative Strategic Emphases

Synthesizing the topic distributions reveals a distinct divergence in the documented strategic priorities of the two organizations during their respective periods of peak commercialization. While both companies focused on market expansion, their operational emphases differed significantly along the commercialization value chain.

Insys's corpus reflects a localized, micro-level strategy heavily oriented toward direct prescriber engagement. Their strategic documentation indicates a highly proactive direct-to-prescriber outreach model, predominantly focusing on prescriber-level interventions. This is evidenced by the prominence of topics detailing speaker programs linked to sales volume and performance-based compensation structures designed to monitor prescription metrics. In essence, Insys's documented strategy during this period prioritized prescriber acquisition, engagement, and utilization intensity. Notably, these computationally derived patterns align closely with the historical record; subsequent federal investigations documented that Insys heavily relied on "speaker programs" to drive and reward prescription volumes (U.S. Department of Justice, 2019).

Conversely, Mallinckrodt demonstrated a macro-level strategy focused on institutional market access and commercial infrastructure. Rather than concentrating primarily on individual physicians, Mallinckrodt's internal discourse prioritized payer segmentation, pricing strategies, and contract optimization. Their emphasis on pre-launch coordination among marketing, medical affairs, and managed markets, alongside careful considerations of organizational channel design, indicates a structural approach. The documents suggest that Mallinckrodt aimed to navigate the broader distribution environment and secure formulary access to facilitate subsequent sales efforts. Notably, external historical records from (Opioid Industry Documents Archive, n.d.) indicate that Mallinckrodt became one of the largest suppliers of generic oxycodone in the United States during the 2008–2016 period. This expansion temporally overlaps with the internally documented emphasis on institutional market access and commercial infrastructure development, suggesting compatibility between strategic orientation and observed distribution scale rather than implying direct causality.

Ultimately, the computational analysis moves beyond merely quantifying strategy occurrences; it illustrates how commercialization strategies can manifest either as targeted behavioral interventions at the prescriber level (Insys) or as the structural optimization of market access and payer dynamics (Mallinckrodt).

6 Conclusion

In this paper, we proposed a structured NLP framework for analyzing corporate strategic behavior from large-scale corporate presentation documents. The framework integrates (i) keyword-based prefiltering, (ii) supervised refinement using a KNN classifier trained on embedding representations of LLM-labeled text chunks, and (iii) topic mod-

eling with BERTopic to uncover latent thematic structures. LLMs were further employed to identify strategy-relevant topics and generate human-readable explanations, thereby enhancing interpretability.

Applying this framework to documents from Insys Therapeutics and Mallinckrodt Pharmaceuticals, we systematically characterized both shared and divergent forms of corporate strategic behavior. The results demonstrate that combining embedding-based classification, topic modeling, and LLM-assisted interpretation enables a structured, scalable, and reproducible analysis of corporate strategic behavior in large archival corpora. Overall, the proposed framework reduces manual coding effort, improves analytical consistency, and accelerates the extraction of meaningful insights into corporate strategic behavior from unstructured litigation documents.

Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

7 Limitations and Future Work

Several stages of the analysis rely on Large Language Models (LLMs) for strategy extraction and interpretation. While this approach substantially reduces manual effort and enables scalable processing, the accuracy and interpretive validity of LLM-generated outputs have not been systematically validated by domain experts. As a result, potential issues such as misclassification, oversimplification, or latent bias may persist. Future research should incorporate structured expert evaluation and inter-rater validation frameworks to assess the reliability, consistency, and robustness of extracted strategic themes.

This study focuses exclusively on presentation documents within the OIDA dataset. Although presentations provide structured and strategy-oriented insights, OIDA contains additional document types—such as internal communications, reports, and correspondence—that may offer complementary or contrasting perspectives. Restricting the analysis to a single document category may therefore limit the comprehensiveness of the findings. Extending the scope to include multiple document types would enable a more holistic reconstruction of organizational behavior.

The approach relies mainly on qualitative interpretation rather than a shared quantitative framework, which limits the rigor of cross-corpus comparisons. Future work may explore the design of standardized quantitative evaluation protocols for more reliable cross-corpus benchmarking.

Another limitation concerns the uneven temporal distribution and inherent gaps within the dataset, which together constrain the longitudinal analysis. Documentation from certain periods—particularly prior to 2010—is relatively limited, introducing potential blind spots in the reconstruction of strategic evolution. Because we use raw document counts to reflect the intensity of strategic activity, fluctuations in data availability may partially confound our interpretations. Consequently, observed shifts in topic prevalence or apparent strategic inflection points might reflect archival density and data discontinuities rather than genuine, substantive changes in corporate strategy. To address this, future studies could employ normalization techniques or weighting schemes to better disentangle true strategic signals from artifacts of data availability.

In addition, this study attempted to identify strategies related to mitigating legal and regulatory risks within the opioid market. However, such strategies may not be explicitly articulated in corporate language and are often embedded in indirect or coded expressions. Consequently, the absence of clearly identified legal-avoidance strategies should not be interpreted as definitive evidence of their nonexistence. Future work may benefit from incorporating legal-domain expertise or alternative analytical frameworks designed to detect implicit regulatory positioning strategies.

Finally, the prompting framework employed in this study is designed to extract passages that explicitly describe strategies. While this improves precision, it may reduce recall by overlooking segments that imply strategies indirectly or require deeper contextual inference. With expert supervision or hybrid analytical pipelines, future research could allow LLM systems to flag potentially implicit or strategically sensitive content for subsequent human review, thereby enabling a more comprehensive and nuanced analysis.

8 Ethical Statement

We emphasize that the strategic categorizations and thematic summaries presented in this study are computational artifacts generated by Large Lan-

guage Models (LLMs) processing an archived text corpus. The outputs reflect the model's algorithmic abstraction of the text based on our defined prompts, rather than established legal, economic, or historical facts. Our framework is designed as an exploratory computational tool to process large-scale text data, not as an adjudicative mechanism. Consequently, the findings should not be interpreted as definitive representations of corporate intent or objective truth. Translating these computational signals into substantive conclusions about legal compliance, market manipulation, or economic strategy requires rigorous, independent evaluation by domain experts, including legal scholars, economists, and regulatory analysts.

References

Ibrahim Al Azher, Venkata Devesh Reddy Seethi, Akhil Pandey Akella, and Hamed Alhoori. 2025. Limtopic: Llm-based topic modeling and text summarization for analyzing scientific articles limitations. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, New York, NY, USA. Association for Computing Machinery.

Alessio Faccia, Julie McDonald, and Babu George. 2024. Nlp sentiment analysis and accounting transparency: A new era of financial record keeping. *Computers*, 13(1).

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Hyewon Kang and Jinho Kim. 2022. Analyzing and visualizing text information in corporate sustainability reports using natural language processing methods. *Applied Sciences*, 12(11).

Opioid Industry Documents Archive. n.d. Timeline of the opioid crisis. `https://timeline.oida-resources.jhu.edu/`. Accessed: 2 March 2026.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jantima Polpinij, Manasawee Kaenampornpan, Christopher S. G. Khoo, Wei-Ning Cheng, and Bancha Luaphol. 2026. A multi-stage nlp framework for knowledge discovery from crop disease research literature. *Mathematics*, 14(2).

Xuan Shen, Brian Wingenroth, Zichao Wang, Jason Kuen, Wanrong Zhu, Ruiyi Zhang, Yiwei Wang, Lichun Ma, Anqi Liu, Hongfu Liu, and 1 others. 2025. Oida-qa: A multimodal benchmark for analyzing the opioid industry documents archive. *arXiv preprint arXiv:2511.09914*.

Aivin V. Solatorio. 2024. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning. *arXiv preprint arXiv:2402.16829*.

University of California, San Francisco and Johns Hopkins University. n.d. Opioid industry documents archive (oida). `https://www.industrydocuments.ucsf.edu/opioids/`. Accessed: 6 February 2026.

U.S. Department of Justice. 2019. Founder and four executives of insys therapeutics convicted of racketeering conspiracy. Accessed: 4 March 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xiaohao Yang, He Zhao, Dinh Phung, Wray Buntine, and Lan Du. 2025. LLM reading tea leaves: Automatically evaluating topic models with large language models. *Transactions of the Association for Computational Linguistics*, 13:357–375.

A Selection of *K* in KNN

To determine the optimal value of K in the *K-Nearest Neighbors* (KNN) classifier, we employed *Repeated Stratified K-Fold Cross-Validation* (Pedregosa et al., 2011). Specifically, the evaluation procedure was configured with 5 folds and repeated 5 times (resulting in 25 total evaluations), in order to reduce variance induced by random data partitioning and to ensure the robustness of the performance estimates. The dataset consists of 300 text chunks, with each company contributing 150 randomly sampled chunks from a subset prefiltered using domain-specific keywords. These chunks were subsequently labeled in a binary manner ("yes"/"no") by a LLM, indicating whether the chunk is related to the company's commercialization strategy.

We evaluated the classifier across values of K ranging from 1 to 100 and used the F1-score as the primary performance metric, given the potential class imbalance in the dataset. The results are illustrated in Figure 5.

The model achieved its highest F1-score at $K = 5$. Compared to the baseline approach—where all keyword-filtered chunks were directly classified as "yes" (yielding an F1-score of 0.7277)—the KNN classifier with $K = 5$ improved the F1-score by

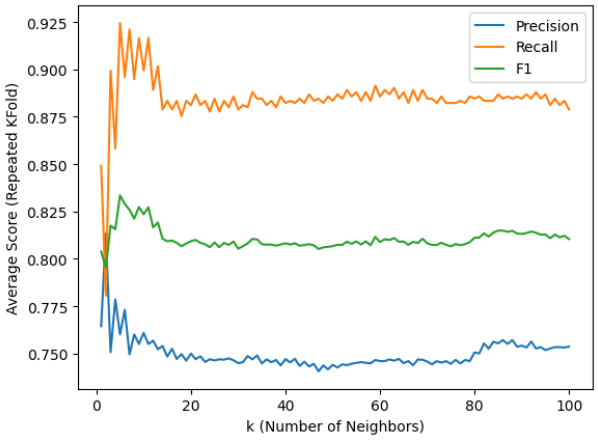


Figure 5: Model Performance over Different Values of K

0.1059. This improvement indicates that incorporating a supervised classification step via KNN effectively reduces noise introduced by keyword-based filtering and enhances the overall predictive performance of the system.

B Prompt Design and Model Configuration

We employed the `openai/gpt-oss-120b` large language model to classify document chunks according to whether they describe a corporate strategy. To ensure deterministic and reproducible outputs, the temperature parameter was set to 0.

B.1 Prompt for Strategy Classification

The following prompt was used to determine whether a given text chunk describes a coordinated corporate strategy or strategic action:

```
{
    "role": "system",
    "content": "You are a research analyst
        studying corporate strategic behavior
        in the pharmaceutical and opioid
        industry.

    Your task is to determine whether a given
        text chunk describes a COMPANY STRATEGY.

    A strategy is defined as a coordinated,
        intentional corporate action designed
        to achieve commercial, regulatory,
        legal, reputational, or market
        objectives.

    A TRUE strategy must:
    - Describe an intentional corporate action
        or coordinated effort
    - Aim to influence revenue, market share,
        regulation, litigation exposure, or
        public perception
    - Go beyond neutral reporting or operational
        background

    Non-strategies include:
    - Lists of advisors or experts
    - Neutral scientific discussion
    - Study design without corporate intent
    - Meeting logistics

    INSTRUCTIONS:
    1. Provide a brief explanation (2-5
        sentences).
    2. On a new line write exactly one of:
       LABEL: YES
       LABEL: NO
    3. Do not write anything after the LABEL
        line."
},
{
    "role": "user",
    "content": "TEXT CHUNK:
    {chunk_text}"
}
```

B.2 Prompt for Topic-Level Strategy Classification

The following prompt was used to classify whether a topic reflects explicit corporate strategic behavior in the opioid industry. The model was instructed to rely strictly on explicit textual evidence and to avoid inference or interpretation beyond the provided excerpts.

```
{
    "role": "system",
    "content": "You are an independent evaluator
        assessing whether a topic explicitly
        reflects corporate strategic behavior
        related to the opioid industry.
    Important:
    - The Topic Representation is contextual
        only.
    - The final decision MUST be based strictly
        on explicit statements in the Document
        Excerpts.
    - Do NOT infer intent.
    - Do NOT rely on background knowledge.
    - Do NOT interpret implications.
    - If it is not explicitly stated, it does
        NOT count.

    Only output the final structured answer.
    "
},
{
    "role": "user",
    "content": "Your task is to assign ALL
        applicable labels based ONLY on
        explicit textual evidence in the
        Document Excerpts.

    LABEL OPTIONS:
    - Sales Expansion & Promotion
    - Influence & Narrative Management
    - Regulatory Risk Management & Evasion
```

```
If none apply, return: None

DEFINITIONS:

Sales Expansion & Promotion requires
    explicit mention of:
- increasing prescription volume
- extending treatment duration
- revenue maximization
- aggressive sales targeting
- formulary positioning through rebates or
    contracting
- expanding into new or higher-risk patient
    populations
- undermining competitors through pricing,
    contracting, or messaging

Influence & Narrative Management requires
    explicit mention of:
- downplaying addiction risks
- shaping scientific evidence or publications
- funding or leveraging key opinion leaders
- sponsoring medical education aligned with
    promotion
- supporting advocacy groups to influence
    public opinion or healthcare policy

Regulatory Risk Management & Evasion
    requires explicit mention of:
- influencing FDA review or regulatory
    processes
- lobbying to shape opioid-related laws or
    policy
- structuring compliance to reduce legal
    exposure
- coordinating litigation defense or
    liability reduction
- managing adverse event reporting or safety
    data to limit regulatory consequences
STRICT RULES:
- Topic Representation provides context ONLY.
- The decision MUST rely on explicit wording
    in the excerpts.
- No inference allowed.
- If wording is vague, it does NOT count.
- Each assigned label MUST have its own
    supporting quote.
- If no label is explicitly supported,
    return None.
OUTPUT FORMAT:
Evidence:
Sales Expansion & Promotion:
- "<exact quoted sentence>"
Influence & Narrative Management:
- "<exact quoted sentence>"
Regulatory Risk Management & Evasion:
- "<exact quoted sentence>"
Justification:
2-4 sentences grounded ONLY in the quoted
    evidence.
Labels:
[comma-separated list of applicable labels
    OR "None"]
---
Topic Representation:
{topic_representation}

Document Excerpts:
{document_excerpts}
"
}
```

B.3 Prompt for Topic Interpretation

The following prompt was used to generate an analytical explanation of each topic identified by the topic modeling procedure. The model was instructed to describe only those promotional or strategic activities that are explicitly supported by the provided topic representation and representative document excerpt.

```
{
    "role": "system",
    "content": "You are analyzing the output of
        a topic modeling system
    for a research project on industry
        promotional strategies.

    Your task is to identify and explain any
        promotional or strategic
    activities that are explicitly described or
        strongly evidenced
    in the provided keywords (Representation)
        and representative document excerpt.

    Focus strictly on strategies, actions, or
        coordinated efforts that are
    clearly supported by the text. Do NOT infer
        hidden motives or intentions.
    Only describe strategic elements that can be
        directly traced to
    specific wording or content in the document.

    Requirements:
    - Write in clear, analytical language.
    - Limit the response to 3-5 sentences.
    - Do not list the keywords.
    - Do not copy phrases verbatim from the
        document.
    - Identify relevant actors (e.g., companies,
        clinicians, regulators, payers).
    - Describe concrete strategic actions only
        if explicitly mentioned.
    - Frame these actions as observable industry
        strategies without speculation.
    - If no clear strategic element is present,
        state that the material is primarily
        descriptive.
    - Ensure that every claim is directly
        traceable to the provided text."
},
{
    "role": "user",
    "content": "
    Representation:
    {representation}

    Representative_Doc:
    {representative_docs}
    "
}
```