VILNIUS UNIVERSITY

FACULTY OF MATHEMATICS AND INFORMATICS

DATA SCIENCE STUDY PROGRAMME

Master's Thesis

Predicting Patent Litigation Outcomes Using Large Language Model-Based Data Extraction and Machine Learning

Patentų ginčų baigčių prognozavimas taikant didžiaisiais kalbos modeliais pagrįstą duomenų išgavimą ir mašininį mokymąsi

Tomas Stoškus

Supervisor : Associate Professor, Ph.D. Viktor Medvedev

Vilnius
2026

Santrauka

Esamuose patentų ginčų duomenų rinkiniuose dažnai trūksta patikimos informacijos apie bylų baigtis, todėl ribojami tyrimai, nagrinėjantys, kaip bylos baigiasi po jų iškėlimo. Šiame tyrime išplečiame USPTO patentų ginčų bylų (Patent Litigation Docket Reports) duomenis, pasitelkdami didelį kalbos modelį bylų sąrašo (docket) įrašų tekste užfiksuotoms baigtims išgauti. Vėliau šias pažymėtas bylas susiejame su patentų charakteristikomis iš „PatentsView“ ir sukonstruojame advokatų patirties rodiklius pagal ankstesnių ginčų istoriją. Išgavimo procesas suklasifikavo 29 084 vieno patento bylas, iškeltas 2003-2020 m., o tikslumas, palyginti su rankiniu būdu užkoduotomis bylomis, 300 imtyje, siekė 81.1%.

Baigčių prognozavimą suskaidėme į tris dvejetainės klasifikavimo užduotis, atitinkančias bylinėjimosi etapus: išlikimo (bylos atmetimas prieš bylos tąsą), susitarimo (taikus susitarimas prieš sprendimą teisme) ir adjudikacijos (ieškovo laimėjimas prieš atsakovo laimėjimą). XGBoost pasiekė geriausią diskriminaciją prognozuojant išlikimą (AUC 0,771) ir susitarimą (AUC 0,716), o Random Forest buvo geriausias adjudikacijos užduotyje (AUC 0,778; 95% PI: 0,718–0,834). Savybių svarbos analizė parodė, kad ankstyvuoju etapu prognozavimą daugiausia lėmė advokatų charakteristikos ir bylos nagrinėjimo vietos (venue) indikatoriai; susitarimo etape didesnę reikšmę įgijo patentų sudėtingumo rodikliai; o adjudikacijos etape stipriausias veiksnys buvo technologijų klasifikacija. Patentų teksto įterpiniai (embeddings) nė viename etape prognozavimo nepagerino. Laikinis patvirtinimas (angl. Temporal validation), taikant 2016 m. ribą, atskleidė susitarimo prognozavimo sumažėjimą, kuris atspindi sistematinius pokyčius tyrinėjamame laikotarpyje.

Raktažodžiai: patentų ginčai, rezultatų prognozavimas, dideli kalbos modeliai, informacijos išgavimas, mašininis mokymasis, teisės analizė

Summary

Existing patent litigation datasets lack reliable outcome information, limiting research on how cases resolve after filing. This study extends the USPTO Patent Litigation Docket Reports Data by using a large language model to extract litigation outcomes from docket entry text, then links these labeled cases to patent characteristics from PatentsView and constructs attorney experience metrics from the litigation history. The extraction pipeline classified 29,084 single-patent cases filed between 2003 and 2020, achieving 81.1% accuracy when validated against hand-coded labels on a sample of 300 cases.

We decompose outcome prediction into three binary classification tasks corresponding to litigation stages: survival (dismissal versus continuation), settlement (settlement versus adjudication), and adjudication (plaintiff versus defendant win). XGBoost achieved the best discrimination for survival (AUC 0.771) and settlement (AUC 0.716), while Random Forest performed best for adjudication (AUC 0.778, 95% CI: 0.718-0.834). Feature importance analysis revealed that attorney characteristics and venue indicators dominated early-stage prediction, patent complexity measures gained prominence at the settlement stage, and technology classification was the strongest predictor at adjudication. Patent text embeddings did not improve prediction at any stage. Temporal validation using a 2016 cutoff showed performance degradation for settlement prediction, reflecting shifts in the litigation landscape during the study period.

Keywords: patent litigation, outcome prediction, large language models, information extraction, machine learning, legal analytics

Contents

Santrauka **2**

Summary **3**

Introduction **6**

1 Literature Review **7**
1.1 Patent Litigation: Background and Process 7
1.2 Empirical Research on Litigation Outcomes 9
1.3 Machine Learning for Litigation Prediction 10
1.4 Attorney Effects in Litigation 11
1.5 Patent Text Representation 12
1.6 LLM-Based Legal Documents Analysis 12

2 Methodology **13**
2.1 Data Source 14
2.1.1 USPTO Patent Litigation Docket Reports Data 14
2.1.2 PatentsView Data 14
2.2 Feature Engineering 15
2.2.1 Attorney Experience Features 15
2.2.2 Patent Characteristics 16
2.2.3 Venue and Procedural Features 16
2.2.4 Patent Text Embeddings 16
2.3 Outcome Extraction 17
2.3.1 Deterministic Extraction Attempts 17
2.3.2 LLM-Based Extraction Pipeline 19
2.3.3 Validation 20
2.4 Final Dataset Construction 21
2.5 Prediction Modeling 22
2.6 Implementation 23

3 Results and Analysis **24**
3.1 Large Language Model Extracted Outcomes 24
3.1.1 Outcome Distribution 24
3.1.2 Classification Confidence by Outcome 25
3.1.3 Procedural Complexity by Outcome 26
3.1.4 Temporal Trends 27
3.1.5 Variation by Technology Sector 28
3.1.6 Venue Patterns 30
3.2 Model Results 32
3.2.1 Experimental Setup and Data Summary 32
3.2.2 Model Performance Comparison 33
3.2.3 Feature Importance Analysis 34
3.2.4 Embedding Contribution 35
3.2.5 Temporal Validation 35

4 Discussion **37**

5 Limitations **39**

6 Future Work **40**

7 Conclusions **41**

References **43**

Appendix A Patent Document Metadata Plots **47**

Appendix B Prompt for Docket Outcome Extraction **49**

Appendix C GitHub Repository **53**

Introduction

Existing patent litigation datasets lack reliable outcome information, and prior prediction work relies on hand-crafted or proprietary data with limited access. We extend the USPTO litigation dataset using LLM extraction on docket entries, enabling prediction and explanation of litigation outcomes at scale.

Researchers have successfully predicted which patents will be litigated, but predicting how litigation resolves remains largely unexplored. Existing models can help patent offices and firms identify litigation risk, yet a more consequential question remains unanswered: once a case is filed, what determines whether the plaintiff wins, the defendant wins, or the parties settle?

This gap stems from the lack of labeled outcome data. The USPTO Patent Litigation Docket Reports, the most comprehensive public dataset, does not reliably code case outcomes. Where outcome information exists, it must be inferred from docket entry text. Prior studies either rely on proprietary data sources or hand-label cases, limiting scale and reproducibility.

The goal of this thesis is to make patent litigation outcomes usable at scale by extracting them from public docket text, and then use those labels to predict and explain how cases tend to resolve after filing.

To achieve this goal, the thesis addresses four tasks:

1. Analyze related work on patent litigation, outcome prediction, and information extraction from legal text.

2. Develop a methodology for extracting litigation outcomes from USPTO docket records and construct a labeled dataset from publicly available sources,

3. Engineer features from patent and litigation data and evaluate prediction models across litigation stages.

4. Analyze which factors predict outcomes at each stage and discuss limitations and future directions.

The remainder of this thesis is organized as follows. Section 1 reviews the literature on patent litigation, outcome prediction, and related methods. Section 2 describes the data sources, outcome extraction pipeline, and feature engineering. Section 3 presents the predictive models and results. And then we further discuss the findings, limitations, and directions for future work.

1 Literature Review

1.1 Patent Litigation: Background and Process

When a patent holder believes their intellectual property has been infringed, they may initiate a civil lawsuit in a U.S. district court seeking injunctive relief and monetary damages. The United States patent system grants inventors exclusive rights to their inventions for a limited period, typically twenty years from the application filing date, in exchange for public disclosure of the invention. Patents are classified according to the Cooperative Patent Classification (CPC) system, a hierarchical taxonomy jointly developed by the USPTO and the European Patent Office. The CPC organizes inventions from broad sections, such as ”Electricity”, ”Physics” or ”Chemistry; Metallurgy” down through classes, subclasses, and detailed subgroups. This classification structure reflects the technical domain of the invention and, as this study will demonstrate, may influence litigation dynamics and outcomes differently across technology areas.

Patent infringement disputes center on whether an accused product or process falls within the scope of the patent's claims. Claims are the numbered statements at the end of a patent document that define the legal boundaries of the patent holder's exclusive rights; they function as the ”metes and bounds” of intellectual property in much the same way that a deed describes the boundaries of real property. Defendants in infringement suits typically pursue one of two strategies: arguing that their product does not meet the claim limitations (non-infringement) or arguing that the patent itself should not have been granted (invalidity). Although U.S. patents carry a presumption of validity, invalidity defenses based on anticipation by prior art, obviousness, or inadequate written description remain common litigation strategies [6].

Patent litigation in U.S. district courts proceeds through several distinct stages, each representing a potential resolution point. The process begins when the patent holder files a complaint alleging infringement. The defendant responds with an answer and typically counterclaims seeking a declaratory judgment of non-infringement or invalidity. Following initial pleadings, the case enters the discovery phase, during which parties exchange documents, respond to interrogatories, take depositions, and gather evidence regarding both infringement and validity. Discovery is typically the most resource-intensive phase of patent litigation, with costs often reaching several million dollars for complex cases involving large corporate defendants [30]. The expense reflects the technical complexity of patent disputes: parties routinely retain expert witnesses in both the relevant technology and damages calculation, and document production can involve millions of pages of technical specifications, source code, and business records.

After discovery, courts conduct a claim construction hearing to determine the meaning of disputed claim terms. This proceeding, known as a Markman hearing after the Supreme Court's decision in *Markman v. Westview Instruments* [39], has become a important event in patent litigation. In *Markman*, the Court held that claim construction is a question of law for the judge rather than a question of fact for the jury. The practical consequence is that the judge's interpretation of claim language often proves decisive: once the court defines the scope of the patent claims, the likely infringement and validity outcomes become substantially clearer. Empirical research suggests that

claim construction rulings strongly influence subsequent case trajectories, as parties can better assess their litigation prospects and adjust settlement positions accordingly [33]. Cases that survive claim construction may proceed to summary judgment motions and, for the small fraction that remain contested, to trial.

The majority of patent cases never reach trial. Empirical research consistently finds that approximately 95% of filed patent suits terminate before going to trial, with most resolutions occurring early in the litigation process [23]. The high settlement rate reflects both the substantial costs of sustained litigation and the inherent uncertainty in predicting trial outcomes. For patent holders, particularly individual inventors and small firms, the expense of litigation through trial can become prohibitive, creating pressure to accept settlement terms that may undervalue their claims. Lanjouw and Schankerman [24] documented this asymmetry, finding that cases involving individual patent holders against corporate defendants settled at higher rates and on terms less favorable to plaintiffs than cases between comparably resourced parties. Conversely, defendants facing the risk of injunction may prefer settlement to the possibility of being excluded from a product market. These economic pressures mean that litigation outcomes are shaped not only by the legal merits of a case but also by the relative financial resources and risk tolerances of the parties involved.

The structure of the litigation process creates significant challenges for empirical research on case outcomes. Because cases can terminate at any stage, whether through voluntary dismissal, negotiated settlement, summary judgment, or trial verdict, and because settlement agreements are typically confidential, comprehensive outcome data is difficult to obtain from public records. Court dockets record that a case was ”dismissed,” but this entry does not distinguish between dismissals resulting from settlement agreements, dismissals for procedural deficiencies such as improper venue, and dismissals on substantive grounds such as failure to state a claim. The technical documentation for the USPTO Patent Litigation Docket Reports Data explicitly acknowledges this limitation, noting that ”the settlement field in PACER records is not determinative and analysts should examine individual docket entries to determine case outcomes” [2, 4]. This ambiguity in publicly available records has constrained prior empirical work and motivates the present study’s methodological approach: using large language models to extract outcome information from the full text of docket entries.

The characteristics of patents that end up in litigation also warrant consideration, as they differ systematically from the general patent population. Lanjouw and Schankerman [22] established several stylized facts about litigated patents: they tend to have more claims, receive more forward citations, and belong to owners with smaller patent portfolios than non-litigated patents. These patterns suggest that litigation selects for patents that are both more valuable, as proxied by citations and claim scope, and held by owners with fewer alternative assets to protect. Multi-patent litigation, in which plaintiffs assert multiple patents against a defendant in a single suit, is common and associated with longer case duration and greater procedural complexity. The present study restricts analysis to single-patent cases to enable clean attribution of patent-level characteristics to case outcomes, though this restriction limits generalizability to multi-patent disputes.

Finally, the rise of non-practicing entities has reshaped the patent litigation landscape over the past two decades. Non-practicing entities, sometimes called patent assertion entities or pejoratively

”patent trolls,” are organizations that acquire patents primarily for licensing revenue and litigation rather than for commercializing products. Because NPEs face no risk of counter-assertion, since they have no products that could infringe a defendant’s patents, they operate under different strategic constraints than practicing companies. Empirical research suggests that NPE litigation exhibits distinct patterns: Love [25] found that NPEs begin asserting patents later in the patent term and continue litigation closer to expiration than product-producing companies. Whether these timing differences translate to different outcome distributions remains an open question, though NPE status itself is not directly observable in the data and must be inferred from party characteristics.

1.2 Empirical Research on Litigation Outcomes

Research on patent litigation characteristics has established several stylized facts about which patents end up in court. [3] examined ”valuable patents” that survive to full term and are maintained through all maintenance fee payments, finding that litigated patents have significantly more claims, prior art citations, and forward citations than non-litigated patents. Forward citations emerged as the strongest predictor of litigation across multiple studies, suggesting that citation-based features should be important in outcome prediction models.

Given that the majority of patent cases resolve through settlement rather than adjudication, understanding settlement determinants is critical for outcome prediction. Somaya [35] modeled settlement decisions as strategic choices, identifying two primary influences: the use of patents as ”isolating mechanisms” to protect valuable strategic stakes, and their ”defensive” role in obtaining technology access through mutual hold-up. His analysis of computer and pharmaceutical patents found that strategic stakes significantly predicted non-settlement, while evidence for mutual hold-up was stronger in computer patents. Cremers [12], studying German patent litigation, found that prior opposition proceedings negatively impacted settlement probability. Patents that had survived validity challenges were less likely to settle, perhaps because their validity had been partially validated through the opposition process. Bar and Kalinowski [**BarKalinowski2019**] developed a theoretical model highlighting the role of evidence discovery in settlement timing: patentees with weak patents prefer early settlement to avoid the risk of invalidation, while those with strong patents may prefer to litigate to establish precedent.

The duration of patent litigation varies substantially by venue, case complexity, and party type. Love [25] conducted an empirical analysis of litigation timing that revealed stark differences between practicing entities and non-practicing entities. Product-producing companies predominantly enforce their patents soon after issuance and complete enforcement well before expiration, while NPEs begin asserting patents relatively late in the patent term and frequently continue litigation until expiration. This variance is so dramatic that all claims asserting the average product-company patent are resolved before the average NPE patent is asserted for the first time. NPEs, despite enforcing only 20% of studied patents, were responsible for over two-thirds of all suits in the final three years of the patent term. These timing patterns suggest that patent age at litigation may interact with party type in predicting outcomes, a relationship that prediction models should capture.

Prior empirical work on litigation outcomes has largely been confined to specific technology do-

mains [1]. Studies of pharmaceutical patent disputes have found that remaining market exclusivity and the number of patents asserted are strong predictors of settlement, consistent with the theory that settlement reflects the discounted expected value of continued exclusivity. This domain-specific focus reflects both data availability constraints and the unique regulatory structures governing certain patent categories. Whether these findings generalize across technology areas remains an open question.

Despite this substantial body of research on litigation determinants, a significant gap remains: prior work has focused on predicting which patents will be litigated or examining settlement behavior within specific technology domains, but no study has systematically predicted how patent litigation resolves at scale using publicly available data. The prediction studies by Juranek and Otneim[37], and Chen and Lai [11] address litigation occurrence rather than outcomes. The settlement analyses by Somaya [35] and Cremers [12] either rely on proprietary data sources or restrict attention to single industries. The most comprehensive public dataset, the USPTO Patent Litigation Docket Reports, explicitly lacks reliable outcome coding, forcing researchers to either hand-label small samples or aggregate proprietary data.

1.3 Machine Learning for Litigation Prediction

Machine learning approaches to patent litigation have primarily focused on predicting which patents will be litigated rather than how litigation resolves. Juranek and Otenim [37] used gradient boosting and random forest models to predict litigation occurrence from patent characteristics, achieving substantial predictive power from patent value indicators and owner characteristics. Their analysis found that extending the feature set had the largest positive impact on prediction performance, with model choice providing additional but smaller gains. Tree-based models proved superior to logistic regression, a result the authors attribute to non-linearities and interactions across predictors. Chen and Lai [11] achieved 79% accuracy in predicting litigation using USPTO examination and assignment data. However, these studies predict the binary outcome of whether litigation occurs, not how cases resolve once filed.

The broader field of Legal Judgment Prediction provides methodological guidance for patent outcome prediction. A systematic review by Dina et al.[13] identified 21 NLP methods and numerous machine learning approaches applied to legal prediction tasks, with Support Vector Machines and TF-IDF representations being most common in earlier work. Medvedeva et al. [28] achieved 75% accuracy predicting European Court of Human Rights decisions using SVMs on case facts, while Katz et al. [21] achieved 70% accuracy predicting U.S. Supreme Court outcomes using random forests trained on case metadata. Critically, Medvedeva et al.[27] distinguished between ”outcome identification” (extracting verdicts from judgment text), ”outcome categorization” (classifying completed cases), and ”outcome forecasting” (predicting future cases). Their analysis revealed that performance drops substantially when models trained on historical data are applied to future cases: forecasting achieved a maximum F1-score of 66% compared to 92% for categorization on the same ECHR cases.

Following the methodological precedent established in prior patent and legal prediction work, this study employs logistic regression, random forest, and XGBoost as prediction models. Logistic

regression serves as an interpretable baseline that has been widely used in legal prediction research. Random forest and gradient boosting methods have consistently outperformed simpler approaches in litigation prediction tasks, particularly when non-linearities and feature interactions are present Juraneko and Otneim[37]. The comparison across model classes allows us to assess whether the additional complexity of ensemble methods yields meaningful predictive gains for litigation outcome prediction.

Several methodological considerations from the legal prediction literature inform our approach. The distinction between categorization and forecasting matters: our task is fundamentally a forecasting problem since we aim to predict outcomes for cases using only information available at filing. Temporal validation is essential to avoid inflated performance estimates from training on future information. Class imbalance is endemic to litigation data, where settlements dominate and adjudicated outcomes are rare. Prior work in patent litigation prediction has addressed imbalance through stratified sampling and appropriate evaluation metrics, practices we adopt here.

1.4 Attorney Effects in Litigation

The foundational framework for understanding attorney and party effects in litigation is Galanter's [15] theory of repeat players versus one-shotters. Galanter argued that repeat players, parties who engage in frequent litigation, enjoy systematic advantages: they can develop expertise, amortize start-up costs across cases, establish relationships with legal institutions, and strategically "play for rules" by settling cases likely to produce unfavorable precedent while litigating those likely to establish favorable law. In patent litigation, this framework suggests that experienced patent litigators and firms with large patent portfolios should outperform occasional litigants and their counsel. Empirical tests of party capability theory have generally supported these predictions. Hamzehzadeh [19], studying repeat players in litigation, found that financially and organizationally stronger parties tend to prevail against weaker parties. Songer et al. [36] confirmed that organizational litigants and government parties outperform individuals, though they also found that amicus support could partially offset the advantages enjoyed by repeat players.

Despite the theoretical importance of attorney effects, empirical measurement has been challenging. Most litigation studies use crude proxies for party capability, classifying litigants as individuals versus organizations or identifying government parties, rather than directly measuring attorney experience. McGuire [26] measured Supreme Court attorney experience by counting prior appearances, finding that experience predicted success even after controlling for party type. Chang et al. [10] developed the most comprehensive attorney experience dataset to date for tort litigation in Taiwan, measuring multiple experience dimensions and finding that experienced attorneys achieved better damages outcomes for their clients. However, these approaches have not been applied to patent litigation, where attorney specialization by technology area, district court, or party side may be particularly relevant given the technical complexity of patent disputes.

Beyond absolute experience levels, the relative experience of opposing counsel may affect litigation outcomes. Cases where one side enjoys substantial experience advantages may resolve differently than evenly matched contests. The disadvantaged party may settle earlier or on less favorable

terms, while cases that proceed to judgment may exhibit selection effects based on the nature of the experience imbalance. This study addresses the gap in the patent litigation literature by constructing attorney experience metrics from the complete USPTO litigation dataset, enabling analysis of how attorney characteristics interact with patent features to predict case outcomes.

1.5 Patent Text Representation

Patent documents present unique challenges for natural language processing. Patent claims are drafted in highly technical, domain-specific language that serves both legal and scientific functions, describing the invention with sufficient precision to define the scope of protection while satisfying statutory disclosure requirements. This results in text characterized by lengthy sentences, nested claim structures, specialized terminology (including neologisms coined for specific inventions), and deliberate ambiguity employed strategically by patent drafters. General-purpose language models trained on web text or news corpora may fail to capture the semantic relationships specific to patent language, motivating the development of domain-specific representations.

Several research groups have developed patent-specific language models. Google released BERT for Patents [17], a BERT model pre-trained on over 100 million patent documents that captures domain vocabulary and phrasing. More recently, [16] introduced PaECTER (Patent Embeddings using Citation-informed TransformERs), which fine-tunes BERT for Patents using examiner-added citations as a training signal. The key insight is that patents cited together by examiners as relevant prior art share semantic similarity that should be reflected in the embedding space. PaECTER produces 1,024-dimensional embeddings and outperforms both general-purpose models and BERT for Patents on patent similarity tasks, predicting the most similar patent at rank 1.32 on average. For litigation prediction, PaECTER embeddings may capture technical characteristics relevant to infringement analysis and validity assessment that are not reflected in structured patent metadata.

1.6 LLM-Based Legal Documents Analysis

Large language models offer a promising alternative for legal information extraction. Breton et al. [8] demonstrated that LLMs can extract legal terms with limited annotated training data, leveraging pre-trained knowledge to interpret context-dependent legal language. Izzidien et al. [20] achieved F1 scores of 0.94 in classifying case types from a large UK case law corpus, substantially outperforming keyword-based approaches. [38] proposed a weak supervision approach combining LLM-generated labels with sequence labeling models to extract attributes from legal documents with limited manual annotations. Shu et al. [34] developed LawLLM, a multi-task legal language model for the US legal system that includes verdict extraction as part of its legal judgment prediction pipeline. These studies suggest that LLMs can interpret the semantic context of legal text to make extraction decisions that would require domain expertise from human annotators, precisely the capability needed to extract litigation outcomes from variable docket entry descriptions.

2 Methodology

This study combines two data sources and multiple methodological approaches to predict patent litigation outcomes. As mentioned in the previous sections, the core challenge is that while the USPTO Patent litigation Docket Reports Data provides comprehensive coverage of U.S. patent litigation since 1963, it lacks reliable outcome coding. We address this gap by using a large language model to extract litigation outcomes from the case docket entry text, then link these labeled cases to patent characteristics from the PatentsView API and construct attorney experience metrics from the litigation history recorded in the USPTO data.

The methodological pipeline proceeds in four stages. Firstly, describing the data sources and their linkages. Secondly, we detail the feature engineering process, covering party attorney experience measures and patent characteristics. Thirdly, we present the large language model based outcome extraction pipeline, including prompt design, taxonomy development and validation against hand-coded labels. Fourthly, we explain the prediction modeling framework and the evaluation approach. The Figure 1 bellow provides a schematic overview of this pipeline.

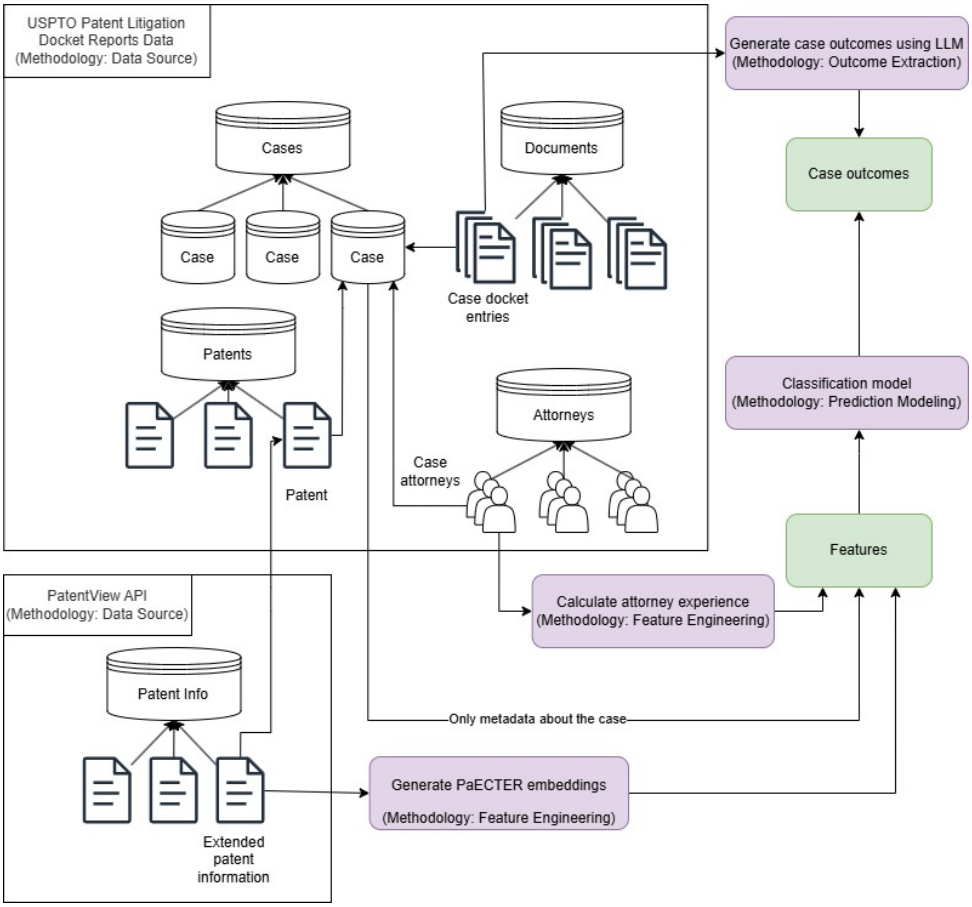

Figure 1 *Methodology*

2.1 Data Source

This study integrates two primary data sources: the USPTO Patent Litigation Docket Reports Data and patent-level data from the PatentsView.

2.1.1 USPTO Patent Litigation Docket Reports Data

The USPTO Patent Litigation Docket Reports Data os a publicly available dataset maintained by the Office of the Chief Economics at the U.S. Patent and Trademark Office [2, 4]. The dataset was originally released in 2017 and has been updated twice, most recently in 2024 to include cases filed through 2020. The underlying records are derived from the Public Access to Court Electronic Records (PACER) system and the RECAP archive.

The dataset is comprised of six interlinked files but only 3 are used in this study (cases, documents and attorneys).

The `cases` file contains metadata on 96,966 unique cases across all federal district courts . Each record includes court and district identifiers, filing and disposition dates, presiding judge assignments, jury demand indicators and a hand-coded case type field indicating the nature of the dispute (e.g., ”Patent Infringement Suit”, ”False Marking”, ”DJ Invalidity Only”) that are only present for cases from 2003.

The `documents` file contains 9.8 million docket entries with filing dates and text description of procedural events that are linked with cases. The `long_description` field forms the textual basis for the outcome extraction pipeline described in Section 2.3.

The `attorneys` file contains 1,8 million observations linking attorneys to parties in each case, including party side indicator. This file enables the construction of attorney experience metric described in Section 2.2.

2.1.2 PatentsView Data

To supplement litigation records with patent characteristics, we linked asserted patents to PatentsView using patent numbers as the join key. PatentsView is a patent data platform maintained by the USPTO that provides access to structured data on U.S patents through both an API and bulk data download [31].

From the API we retrieved structured metadata for patents appearing in the litigation dataset filed since 2003. This included temporal information such as grant dates, application filing dates, and prosecution duration. Technology classification came from the Cooperative patent Classification system, providing both section-level categories (e.g., ”Physics”, ”Electricity”) and detailed class codes. Citation data included backwards and forwards citation count. Patent complexity measures included figure counts, drawing sheet counts and detailed description lengths. Assignee information included organization names, types, classifications and country codes.

Patent text fields were obtained from the bulk data download. We extracted patent summaries, claim text and detailed descriptions for use in generating text embeddings as described in Section 2.2

2.2 Feature Engineering

Crafting features for litigation outcome prediction requires careful consideration of what drives case results. At the core, patent disputes follow the same fundamental legal structure as any litigation: a plaintiff brings a claim against a defendant, in a particular venue, over a specific object of dispute. This framework guided our feature engineering strategy. We organize predictive features around four categories: attorney characteristics, patent attributes, venue and procedural indicators, and text-based representations.

2.2.1 Attorney Experience Features

We constructed attorney experience metrics from the complete litigation history in the USPTO dataset. Patent litigation outcomes may depend not only on patent characteristics but also on the quality and experience of legal representations. We quantified attorney experience along multiple dimensions, all measured as of the case filing date to avoid data leakage and inflation of expertise of legal representation.

The USPTO attorneys file contains a `party_type` field with 182 distinct values describing each attorney's role in a case. We constructed a lookup table mapping these raw string to four normalized categories, plaintiff, defendant, inventor and other. Plaintiff-side classifications include direct plaintiffs, counter-claimants, cross-claimants, third-party plaintiffs and petitioners. Defendant-side classifications include defendants, counter-defendants, cross-defendants and respondents. Attorneys appearing on behalf of parties with both plaintiff and defendant roles in the same case were flagged ass dual-role appearances.

For each attorney-case appearance we computed cumulative experience as of the case filing date by counting all prior cases in which the attorney appeared. We constructed five experience dimensions: total prior patent cases reflecting overall litigation experience; same-district prior cases - capturing familiarity with local court procedures and judges; same-CPC-class prior cases - measuring technology-specific expertise; same-CPC-section prior cases measuring broader technology domain experience; and prior cases on each side - capturing specialization in plaintiff or defendant representation.

Because cases typically involve multiple attorneys on each side, we aggregated attorney-level metrics to the case level using three statistics: maximum experience - capturing the most experienced attorney; mean experience - reflecting overall team quality; and sum of experience measuring total team depth. We computed these aggregations separately for plaintiff-side attorneys, defendant-side attorneys, and all attorneys combined.

To capture relative legal firepower between opposing sides, we constructed experience imbalance measures. Raw imbalance was computed as the difference between plaintiff and defendant experience. Normalized imbalance was computed as the difference divided by the sum, yielding values on a -1 to +1 scale where positive values indicate plaintiff advantages. we constructed imbalance measures for total experience, technology-specific experience, and district-specific experi-

ence. Beyond experience levels, we captured team structure through attorney counts by side and the plaintiff-to-defendant attorney count ration.

2.2.2 Patent Characteristics

patent age at litigation was computed as the difference between case filing date and patent granted date. We also computed remaining patent life assuming the standard 20-year term grant date. These features capture both the maturation of patent value and stakes associated with remaining exclusivity, which was found to be indicative of patents ripe for litigation [1].

We extracted forward and backward citations counts from PatentsView. Forward citation ratio was computed as citations received divided by citations made, providing a normalized measure of patent impact. We also computed the proportion of foreign documents among backwards citations as a proxy for international technology scope.

Patent processing time was measured as days from application filing to grant. To contextualize this duration, we computed the average processing time for patents in the same CPC class and constructed both the absolute deviation and ratio relative to this benchmark. Longer-than-average prosecution may indicate examiner scrutiny or claim scope disputes relevant to litigation outcomes.

We included number of claims, number of figures, number of drawing sheets, and detailed description length as measures of patent complexity. We also constructed ratios including description characters per claim and figures per claim to capture disclosure density relative to claimed scope. Assignee characteristics included indicator variables for foreign assignees and individual assignees, capturing potential differences in litigation behavior between domestic and foreign patent holders or between corporate and individual inventors.

2.2.3 Venue and Procedural Features

We constructed binary indicator for historically significant patent venues including the Eastern District of Texas, the District of Delaware, and the Northern District of California. These districts have handled disproportionate share of patent litigation and may exhibit distinct procedural characteristics.

Case complexity indicators included jury demand, presence of related cases, and designation as part of a lead case. Jury demand may signal party expectations about case complexity or damages magnitude, while related and lead case indicators capture multi-defendant campaigns or consolidated litigation.

2.2.4 Patent Text Embeddings

Patent documents present challenges for traditional text representation methods. Bag-of-words and TF-IDF approaches struggle with patent language because patents frequently contain neologisms coined specifically for the invention being claimed, terms that would be absent from any precomputed vocabulary. Patent claims also employ stylized legal phrasing with lengthy nested struc-

tures and deliberate ambiguity that bag-of-words methods cannot capture. These characteristics motivated the use of contextual embeddings from a domain-specific language model.

We generated dense vector representations using PaECTER (Patent Embeddings using Citation-Informed TransformERs), a model fine-tuned on patent citation relationships [16]. This model was built on Google's BERT for Patents architecture, which was pre-trained on over 100 million patent documents, and produces 1,024-dimensional embeddings optimized for patent similarity tasks.

We embedded patent summaries directly. For patent claims, which vary in number across patents and contain distinct technical content, we embedded each claim separately and aggregated to the patent level using five statistics: mean embedding capturing average claim content, minimum and maximum embeddings capturing the range of claim scope, standard deviation embedding capturing claim heterogeneity, and first claim embedding capturing the typically broadest independent claim. This aggregation yielded 5,120 claim driven features per patent. All embedding dimensions were included as model features without dimensionality reduction.

2.3 Outcome Extraction

2.3.1 Deterministic Extraction Attempts

Before turning to language models, we attempted rule-based extraction using regular expressions targeting common disposition language. Patterns included explicit settlement markers (”settlement agreement,” ”parties have settled”), judgment indicators (”summary judgment granted,” ”judgment for plaintiff”), and dismissal phrases (”dismissed with prejudice,” ”case closed”). This approach seemed promising given the formulaic nature of legal filings.

Preliminary validation on a sample of 50 cases revealed fundamental limitations. Docket entries employ varied and context-dependent phrasing that resists pattern matching. A single court might record the same procedural event in multiple ways across different cases, and terminology varies substantially across districts. More problematically, the dispositive order is frequently not the final docket entry. Fee motions, cost bills, appeals, and administrative closures routinely follow the substantive resolution, meaning a simple ”look at the last entry” heuristic fails. In our validation sample, fewer than 40% of cases had the outcome-determining entry among the final docket records. These failures motivated the use of a large language model capable of interpreting semantic context across the full docket history rather than matching isolated phrases.

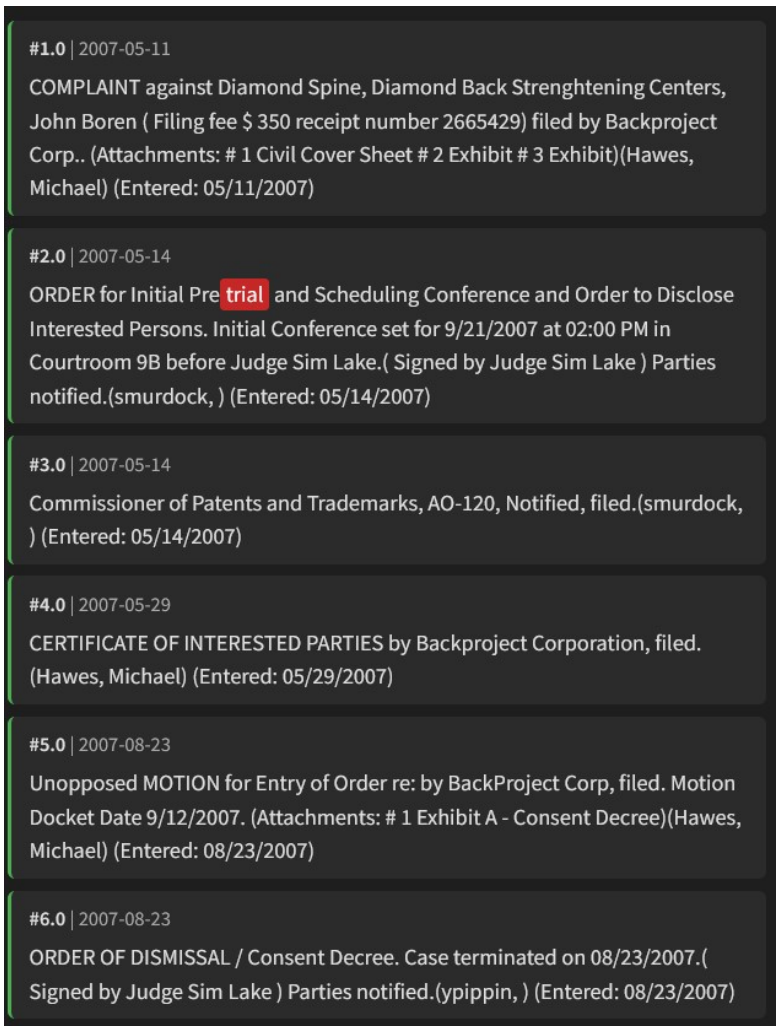

Figure 2 *Docket entries for case: 4:07-cv-01616*

Figure 2 shows docket entries for a patent case filed in May 2007. The complaint names three defendants, and standard procedural entries follow: a scheduling order, notification to the Patent Commissioner, and a certificate of interested parties. Then in August, an ”Unopposed MOTION for Entry of Order” appears with an attached ”Consent Decree,” followed immediately by an ”ORDER OF DISMISSAL / Consent Decree” terminating the case.

This is a settlement. The consent decree and unopposed motion signal negotiated resolution rather than adjudication. But a regular expression matching ”ORDER OF DISMISSAL” cannot distinguish this from a court-initiated dismissal for jurisdictional defects or failure to prosecute. The word ”dismissed” appears in both contexts. Correctly classifying this outcome requires recognizing that consent decrees reflect party agreement, that the motion was unopposed (suggesting coordination rather than adversarial proceedings), and that no entry shows the court ruling on infringement or validity. These are semantic judgments, not pattern matches.

This example is among the simpler cases in the USPTO dataset. The docket contains only six entries and the resolution is relatively transparent. Cases proceeding through discovery, claim construction, and summary judgment motions accumulate hundreds of entries. Post-resolution filings for attorney fees, cost taxation, or appeals can bury the dispositive event deep in the docket history. The extraction task scales in difficulty with procedural complexity.

2.3.2 LLM-Based Extraction Pipeline

We employed an LLM-based extraction pipeline applied to docket entry text from the documents file. Model selection required careful consideration of two constraints. Exploratory data analysis showed that docket histories for some cases exceed 4,000 entries, which excluded LawLLM [34] since its Gemma 7B [40] base limits context to 8,192 tokens. The model also needed native support for structured outputs to ensure programmatic parseability across the corpus.

We evaluated candidate models on a development sample of 50 cases with hand-coded outcomes. Non-reasoning models, including Llama 3.1 7B [18], produced inconsistent results on longer docket histories, frequently misidentifying the dispositive event when procedural entries accumulated after case resolution. Reasoning models performed substantially better, with gpt-oss-20b achieving the highest accuracy on cases exceeding 50 docket entries. We attribute this to the model's ability to trace procedural sequences and identify which events were substantively dispositive versus administrative.

We selected OpenAI's gpt-oss-20B [29], an open-weights model released under the Apache 2.0 license, for the full extraction. While proprietary reasoning models might achieve marginally higher accuracy, enabling other researchers to replicate our extraction process took priority.

For each case, we extracted all docket entries where the `long_description` field was non-empty. Entries lacking text descriptions were excluded, a limitation since some docket events are documented only through linked PDF documents in PACER rather than text summaries. The retained entries were sorted chronologically by filing date and document number to preserve the temporal sequence of litigation events. This ordering matters because the model must understand procedural progression to identify which events are dispositive versus preliminary.

Each case was processed in a fresh conversation context to prevent cross-case contamination. The extraction prompt, informed by prior work on legal information extraction [20, 38], instructed the model to act as a patent litigation analyst and return only valid JSON with no surrounding text. The prompt defined six mutually exclusive outcome categories:

- *Settlement*: parties reached agreement, identified through explicit settlement language, stipulated dismissals, consent judgments, or voluntary dismissals following apparent negotiation. This definition recognizes that settlements in federal court are typically effectuated through dismissal orders rather than recorded as such in case metadata.

- *Plaintiff win*: judgment or verdict for the patent holder, including jury verdicts for plaintiff, summary judgment on infringement, default judgments, and permanent injunctions.

- *Defendant win*: judgment or verdict for the accused infringer, including jury verdicts for defendant and summary judgments of non-infringement or invalidity.

- *Dismissed*: court-initiated dismissal without settlement, including dismissals for lack of jurisdiction, failure to prosecute, or failure to state a claim. This category applied only when no settlement indicators appeared in the docket.

- *Transferred*: cases moved to another district or consolidated into multidistrict litigation.

- *Unknown*: residual category for cases where outcome could not be determined from available text.

The prompt also required a confidence rating (high, medium, or low) and a reasoning field citing specific docket entries. Confidence ratings allow uncertain cases to be filtered from training data. The reasoning field enables manual validation and encourages the model to ground classifications in textual evidence rather than guessing. The model returned structured JSON containing the predicted outcome, dismissal type (with or without prejudice), trial indicator, and litigation stage at resolution. Secondary fields captured infringement findings, invalidity determinations, injunction status, damages amounts, key dates, presiding judges, and party names. Each field permitted null values when information was not determinable. The exact prompt is provided in Appendix Appendix B.

Temperature was set to zero for deterministic outputs. The extraction prompt was iteratively refined using few-shot examples during development [9]. The model's native context length of 131,072 tokens accommodated lengthy docket histories, as some cases accumulated over 400 entries spanning years of procedural activity. Reasoning effort was set to high. Extraction for the full dataset required approximately 140 hours of inference on a single NVIDIA RTX 4090.

2.3.3 Validation

We validated LLM-extracted outcomes against hand-coded labels on sample of 300 cases. The sample included cases from all outcome categories, with the distribution reflecting findings that more than 95% of the disputes don't reach trials[6].

Overall accuracy was 81.1%. Examination of misclassified cases revealed systematic patterns rather than random errors. The most common error involved settlements being categorized as dismissals. The majority of these cases had dismissal type recorded as "with prejudice." Because settlements in federal court are typically effectuated through stipulated dismissals without explicit mention of the underlying agreement, this pattern likely reflects cases where settlement occurred but was not documented in the docket text. We treat this as conservative labeling: it undercounts settlements rather than inflating adjudication rates.

The converse error, dismissals incorrectly labeled as settlements, occurred infrequently and was associated with dismissals coded "without prejudice." These cases can be identified programmatically and were reclassified during preprocessing. Plaintiff wins and defendant wins were correctly labeled in all manually reviewed cases, consistent with these outcomes generating unambiguous docket language such as jury verdict forms and judgment orders.

Date extraction proved more challenging. PACER dockets employ inconsistent date formats across districts and time periods, and the model struggled to determine which dates were most relevant to outcome classification. Resolution timing analysis was consequently limited, though case duration could still be approximated from filing and final docket entry dates.

Since outcome labels serve as the prediction target, measurement error in the dependent variable attenuates model performance estimates. Reported accuracies should be interpreted as lower bounds on the models' ability to predict true litigation outcomes.

2.4 Final Dataset Construction

We restricted the analysis to single-patent cases, which comprise approximately 30% of patent litigation cases filed between 2003 and 2020. This restriction serves both methodological and computational purposes. Single-patent cases allow direct attribution of patent characteristics to litigation outcomes. In multi-patent cases, aggregating patent-level features (averaging citation counts, pooling embeddings) introduces modeling assumptions that complicate interpretation, particularly for feature importance analysis using SHAP values. The restriction also reduced computational burden: outcome extraction required approximately 140 hours of LLM inference, and multi-patent cases tend toward longer docket histories that compound this cost. Feature dimensionality (7,200+ features per patent) would scale proportionally with the number of patents per case.

Step	Description	N	Dropped
1	Case-patent pairs	159,262	–
2	Matched to PatentsView patent_info	157,178	2,084
3	Matched to patent summaries	147,979	9,199
4	Non-null patent, CPC code, summary text, case type	144,829	3,150
5	Has at least one docket entry with long_description	144,233	596
6	Single-patent cases only	29,624	114,609

Table 1 *Sample construction from USPTO Patent Litigation Docket Reports Data*

Table 2 shows the distribution of cases across technology sectors. Physics and Electricity patents dominate the sample, together accounting for over 55% of cases. This concentration reflects the prevalence of software and electronics litigation in U.S. patent disputes, particularly during the study period. Human Necessities, which includes medical devices and consumer products, comprises roughly 20% of cases. Traditional industrial categories (Chemistry, Mechanical Engineering, Fixed Constructions, Textiles) are less represented, collectively accounting for approximately 14% of the sample.

CPC Section	Count
Physics	9,806
Electricity	6,720
Human Necessities	5,974
Performing Operations; Transporting	3,001
Chemistry; Metallurgy	1,475
Mechanical Engineering; Lighting; Heating; Weapons; Blasting	1,379
Fixed Constructions	991
Textiles; Paper	168

Table 2 *Distribution of single-patent cases by CPC section*

The sample is dominated by infringement suits. Table 3 presents the distribution by case type, where type 1 denotes patent infringement actions and type 2 denotes declaratory judgment actions seeking non-infringement rulings. Infringement cases account for 90.7% of single-patent litigation, slightly higher than their 87.5% share of all patent cases. This difference arises because multi-patent

cases are more common in certain case types: only 3.2% of type 10 cases (which involve ANDA pharmaceutical litigation) appear as single-patent disputes, compared to 44% of standard infringement actions.

Case Type	All Cases	% of All	Single-Patent	% of Single	Single Ratio
1	60,916	87.46%	26,859	90.67%	44.09%
2	3,754	5.39%	1,494	5.04%	39.80%
3	1,169	1.68%	456	1.54%	39.01%
5	1,026	1.47%	326	1.10%	31.77%
6	532	0.76%	167	0.56%	31.39%
4	290	0.42%	110	0.37%	37.93%
11	315	0.45%	96	0.32%	30.48%
15	236	0.34%	40	0.14%	16.95%
10	752	1.08%	24	0.08%	3.19%
12	75	0.11%	23	0.08%	30.67%
9	485	0.70%	16	0.05%	3.30%
8	53	0.08%	10	0.03%	18.87%
7	34	0.05%	3	0.01%	8.82%
14	2	0.00%	0	0.00%	–
13	8	0.01%	0	0.00%	–

Table 3 *Distribution of cases by case type. Single Ratio indicates the percentage of cases in each type that are single-patent disputes.*

2.5 Prediction Modeling

Rather than framing patent ligation outcome prediction as a single multiclass problem, this study decomposes it into three sequential binary classifications tasks corresponding to distinct decision points in the litigation process. This staged approach draws on Cremers [12], who modeled patent disputes as a decision tree where cases filter through successive stages. The first stage, survival, distinguishes cases that are dismissed from those that proceed further in the litigation process. The second stage, settlement, applies only to cases that survived dismissal and separates settlements from cases reaching adjudication. The third stage, adjudication, applies only to the small subset of cases decided on the merits and distinguishes plaintiff victories from defendant victories.

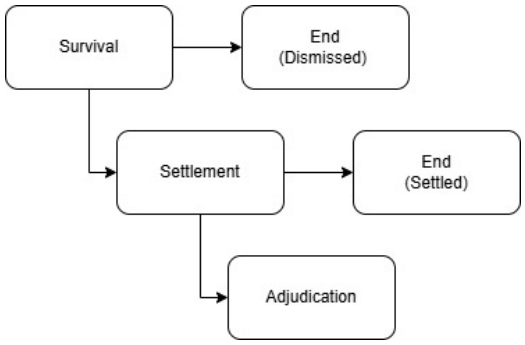

Figure 3 *Litigation process decision tree*

This decomposition offers two advantages. Firs, it addresses the sever class imbalance inherent

in litigation data. Adjudicated outcomes represent fewer than 5% cases, making direct multiclass prediction problematic. By restricting the adjudication model to cases that actually reached judgment we obtain more balanced training data and avoid the model simply learning to predict the dominant class. Second, the staged structure reflects how litigation actually unfolds. The factors driving early dismissal may differ from those determining trial outcomes. Separating these stages allows the models to learn stage-specific patterns.

The dataset was partitioned into training, validation and test sets using 60/20/20 split with random shuffling. The test set remained held out through model development and was used only for final performance estimation. The validation set served two purposes: hyperparameter tuning and classification threshold optimization. Hyperparameters were optimized using randomized search with 150 iterations per model and stage combination, scored by ROC-AUC. For stages with class imbalance exceeding a ratio of 1.5, the default threshold of 0.5 was replaced with a threshold optimized on validation data to maximize the F1 score. This threshold was then applied unchanged to the test set to obtain the final performance estimates. Bootstrap resampling with 1,000 iterations was used to construct 95% confidence intervals for AUC and PR-AUC.

We compared four model classes. Logistic regression served as an interpretable linear baseline. Random forest and XGBoost represented tree-based ensemble methods that can capture nonlinear relationships and feature interactions. Prior work in patent litigation prediction found that tree-based methods outperformed linear models, particularly when predictors exhibit complex interactions [37]. We also included base-rate dummy classifiers to establish floor performance levels. For all non-baseline models, class imbalance was addressed through balanced sample weighting during training.

To asses the contribution of patent text representations, each model was trained in two configurations: using only structured features (patent characteristics, attorney metrics, venue indicators) and using features augmented with PaECTER embeddings derived from patent summaries and claims. Comparing performance across these configurations isolates the marginal value of semantic information extracted from patent documents.

Global feature importance was computed as the mean absolute SHAP value for each feature across all test samples, following the standard aggregation approach lundberg2020local. This measure captures the average magnitude of each feature's contribution to predictions regardless of direction.

2.6 Implementation

All data processing and analysis was conducted in Python. Patent and litigation data were stored in a PostgreSQL database with the pgvector [32] extension (pgvector/pgvector:pg18) running in a Docker container. Feature engineering and modeling used pandas, scikit-learn, and XGBoost with GPU acceleration. Embeddings were generated using the sentence-transformers library. The LLM for outcome extraction was deployed locally using LM Studio (version 0.3.32) as an inference server and accessed via API.

3 Results and Analysis

This section present the finding from our outcome extraction pipeline and predictive modeling experiments. We begin by examining the quality and distribution of LLM-extracted outcomes, which form the foundation for all subsequent analysis. Understanding the characteristics of these labeled outcomes is important because systematic biases or gaps in the extraction process will propagate through the prediction models trained on this data.

After establishing the outcome distributions, we turn to the prediction task itself. As described in the methodology, the severe class imbalance in litigation outcomes required us to restructure the prediction problem into three binary classification tasks corresponding to distinct stages oof the litigation process: survival(whether a case proceeds past dismissal), settlement (whether surviving cases settle or reach adjudication), and adjudication (whether adjudicated cases favor the plaintiff or defendant). For each stage we evaluate models with and without patent text embeddings to isolate their impact.

3.1 Large Language Model Extracted Outcomes

3.1.1 Outcome Distribution

The extraction pipeline classified 29,084 single-patent cases filed between 2003 and 2020. Table 4 presents the distribution across the six outcome categories. Not all of the single patent cases analyzed by the LLM were successfully processed, context limits were reached and the 540 cases with the longest docket entries weren't processed.

Outcome	N	%
Dismissed	15,451	53.1
Settlement	9,841	33.8
Unknown	1,440	5.0
Transferred	1,156	4.0
Defendant win	699	2.4
Plaintiff win	497	1.7
Total	29,084	100.0

Table 4 *Distribution of LLM-extracted litigation outcomes*

Dismissals dominate the sample, accounting for over half of all cases. Settlements comprise another third. Adjudicated outcomes, where a court or jury rendered judgment on the merits, total just 4.1% of cases. This distribution aligns with prior findings that approximately 95% of patent suits terminate before adjudication [23]. Transferred cases reflect venue changes, often to multidistrict litigation, and are excluded from prediction modeling since their ultimate resolution occurs in a different proceeding. The unknown category, examined in the following section, comprises cases where the docket text did not permit reliable classification.

3.1.2 Classification Confidence by Outcome

The 1,440 cases classified as unknown warrant examination to distinguish genuine ambiguity from pipeline failures. The LLM provided confidence ratings alongside each classification, and the pattern for unknown cases differs sharply from other outcomes. While classified outcomes received high confidence ratings in 94-99% of cases, unknown classifications received low confidence ratings in 81.8% of cases. This asymmetry suggests the model appropriately recognized insufficient information rather than forcing an uncertain classification.

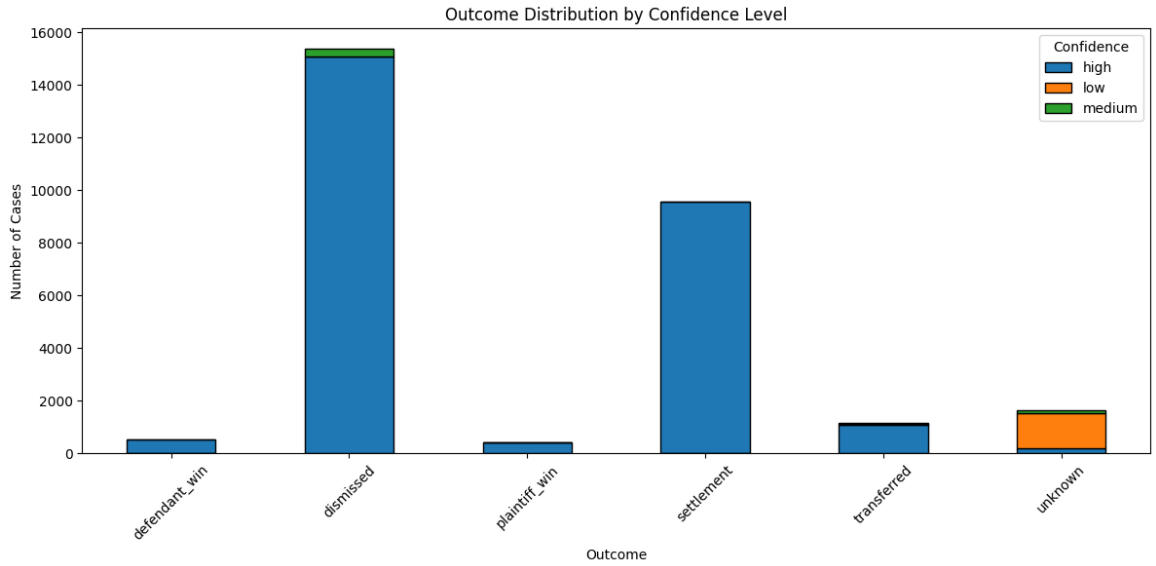

Figure 4 *Outcome Distribution by Confidence Level*

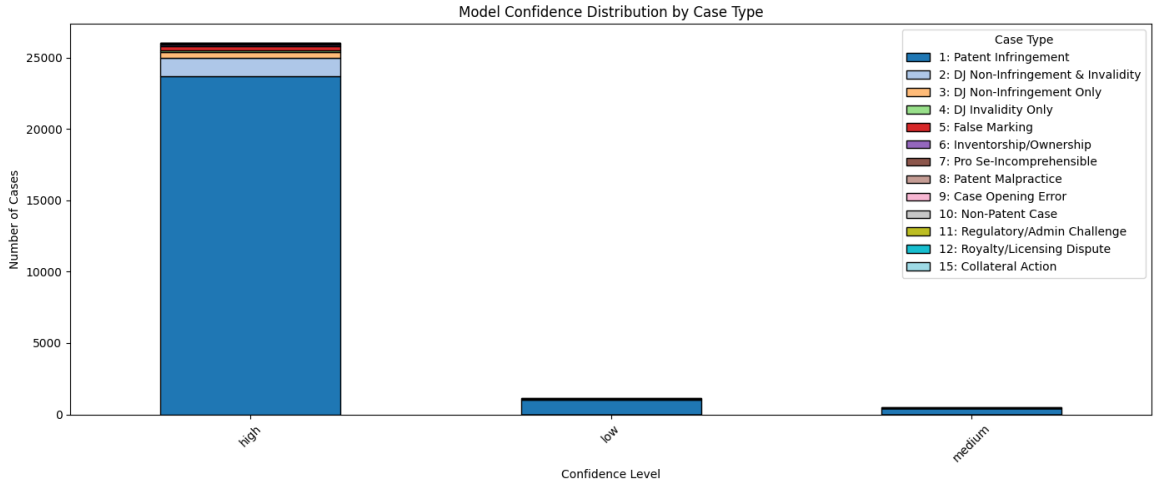

Figure 5 *Model Confidence Distribution by Case Type*

Examining the model's reasoning for unknown classifications reveals several categories. The largest group, comprising 963 cases (66.9%), consisted of cases where docket entries existed but did not clearly indicate a final disposition. These may include cases with procedural activity that trailed off without explicit closure, or situations where resolution occurred through channels not reflected in the docket text. Incomplete records accounted for 363 cases (25.2%), representing dockets that

appeared truncated or missing entries, possibly due to transfers or gaps in PACER data capture. Parse errors affected 193 cases (13.4%), reflecting technical failures where the LLM encountered unusual formatting or docket structures. An additional 128 cases (8.9%) contained only early filings such as the initial complaint and summons with no subsequent activity, likely representing cases still pending at the time of data collection. A negligible number of cases, just 4 total, resulted from missing docket entries or processing timeouts.

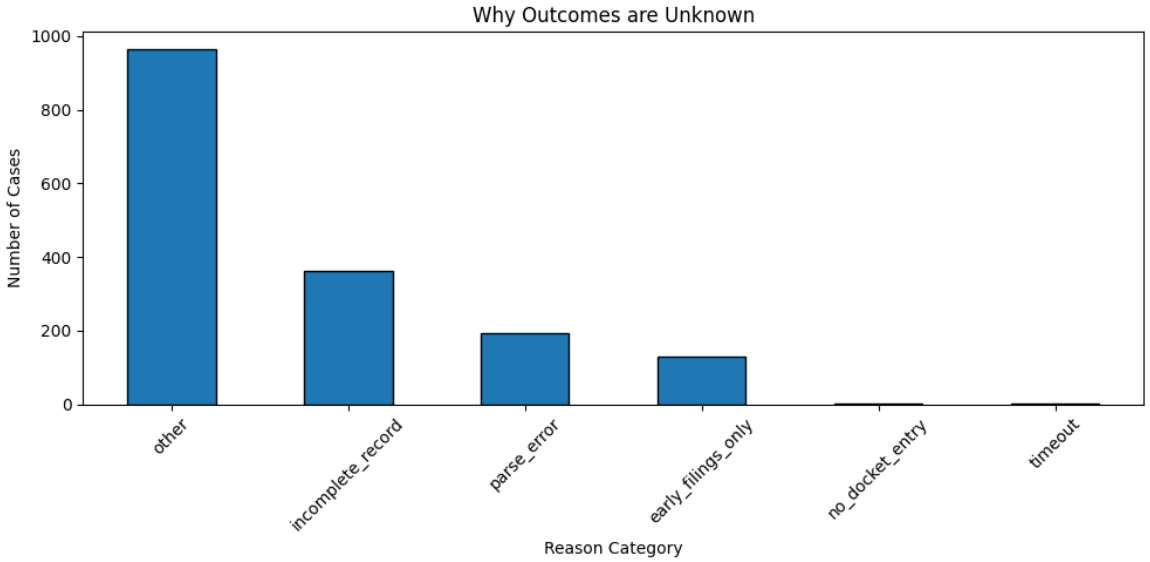

Figure 6 *Reasons associated with Unknown outcomes*

Pipeline failures in the strict sense (parse errors, timeouts, missing entries) account for approximately 14% of unknown cases. The remainder reflect limitations in the underlying data rather than methodological shortcomings. For prediction modeling, we exclude unknown cases from the analysis sample, as they cannot serve as reliable outcome labels.

3.1.3 Procedural Complexity by Outcome

The number of docket entries per case varied substantially and correlated with outcome type in ways that provide implicit validation of the extracted labels. Table 5 presents summary statistics for document counts by outcome category.

Outcome	N	Mean	Median	Std
Defendant win	699	153.1	128.0	99.5
Plaintiff win	497	109.9	64.0	106.8
Settlement	9,841	63.9	40.0	68.1
Transferred	1,156	62.4	41.0	63.9
Unknown	1,440	64.4	29.0	84.0
Dismissed	15,451	39.3	21.0	54.3

Table 5 *Docket entry counts by outcome category*

Adjudicated cases required considerably more procedural activity than cases ending in dismissal or settlement. Defendant wins had a median of 128 docket entries, the highest among all

outcome categories. Plaintiff wins had a median of 64 entries. By contrast, dismissed cases had a median of only 21 entries, and settlements fell in between at 40 entries. The gap is substantial: defendant wins involved roughly six times as many docket entries as dismissed cases.

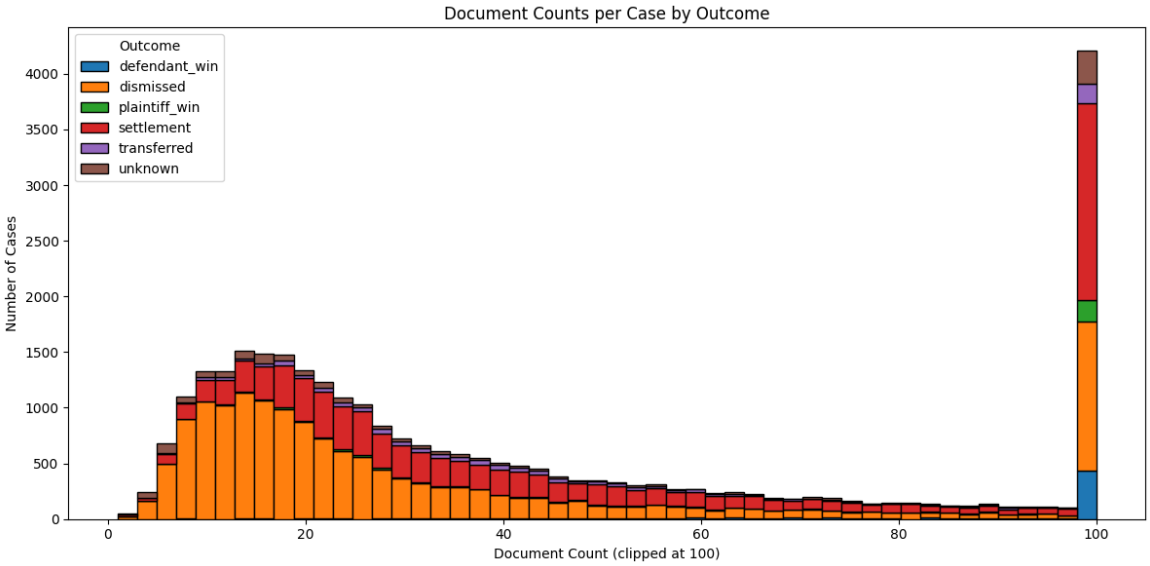

Figure 7 *Document counts per case by outcome*

This pattern reflects the structure of patent litigation. Dismissals can occur before discovery begins, requiring minimal procedural activity. Settlements can occur at any point but often follow enough discovery for parties to evaluate their positions. Adjudicated outcomes, by definition, require the case to survive all intermediate stages, accumulating motions, discovery disputes, claim construction proceedings, and potentially trial. The higher document count for defendant wins compared to plaintiff wins may reflect different paths to victory: defendants often prevail through summary judgment on invalidity or non-infringement after full discovery, while plaintiff wins may include default judgments or cases where liability was clearer earlier in the process.

The extracted outcome labels correlate sensibly with procedural complexity provides evidence that the LLM classification captures meaningful distinctions rather than noise.

3.1.4 Temporal Trends

Outcome proportions shifted markedly over the 18-year period covered by the dataset. Figure 8 displays case counts by outcome and filing year.

In the early years (2003-2009), dismissals and settlements occurred at roughly equal rates, each accounting for approximately 41-45% of cases. By 2015, dismissals had risen to 68% while settlements had fallen to 21%. The trend partially reversed in subsequent years, with dismissals stabilizing around 53-58% and settlements recovering to 29-31% by 2018-2020. Note that the figure displays absolute counts, which also reflect changes in litigation volume over the period; the percentages cited here are calculated relative to annual totals.

Adjudicated outcomes became less common over time. Plaintiff win rates declined from approximately 3% in 2003-2005 to around 1% by 2015-2020. Defendant win rates similarly dropped

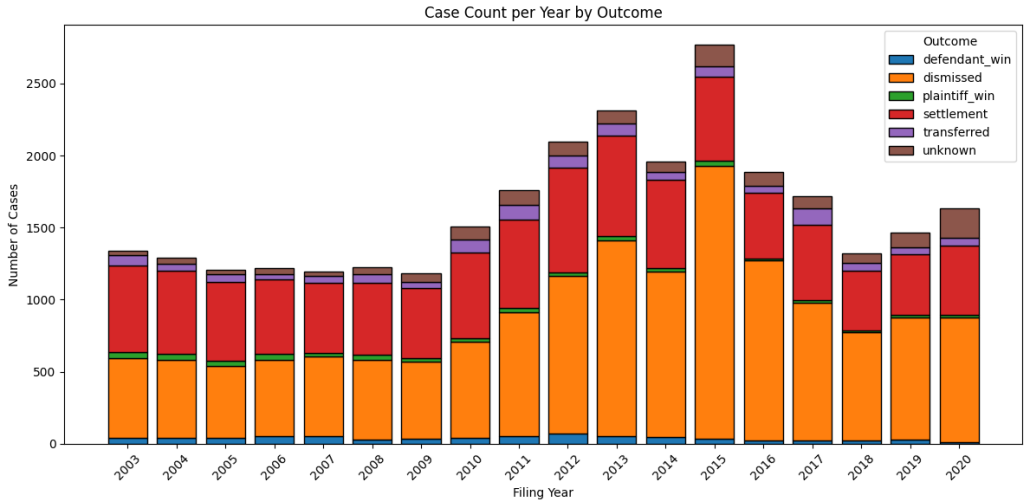

Figure 8 *Case counts by outcome and filing year, 2003-2020*

from 3-4% to 1-2%. This decline follows from the increase in early dismissals: as more cases terminate before reaching the merits, fewer remain to produce adjudicated outcomes.

The proportion of unknown outcomes increased in later years, reaching 12.4% for cases filed in 2020. This likely reflects incomplete data: cases filed near the end of the observation window may have remained pending when the USPTO dataset was compiled. The same caveat applies, to a lesser degree, to 2018-2019 filings. The apparent recovery in settlement rates during these years should be interpreted cautiously, as some pending cases classified as unknown may ultimately resolve through settlement.

3.1.5 Variation by Technology Sector

Outcome distributions varied across technology sectors. Figure 9 displays outcome distributions by CPC section.

Patents in Physics and Electricity exhibited the highest dismissal rates at 59.3% and 61.4% respectively, compared to 40-47% for other sectors. These two sections also had the lowest plaintiff win rates, both under 1%. Chemistry patents showed a more balanced profile with 40.5% dismissals, 40.4% settlements, and a plaintiff win rate of 3.8%.

At the class level, Organic Chemistry stood out with the highest plaintiff win rate (5.15%) and lowest dismissal rate (37.8%) among major patent categories. Computing, Electric Communication, and Signalling occupied the opposite extreme, with dismissal rates exceeding 62% and plaintiff wins below 0.7%.

Chemistry and mechanical patents showed no comparable shift. Their outcome distributions remained relatively stable across the study period. This divergence suggests the aggregate temporal trend observed in Section 3.1.4 was driven primarily by changes in software and electronics litigation rather than a uniform shift across all patent categories.

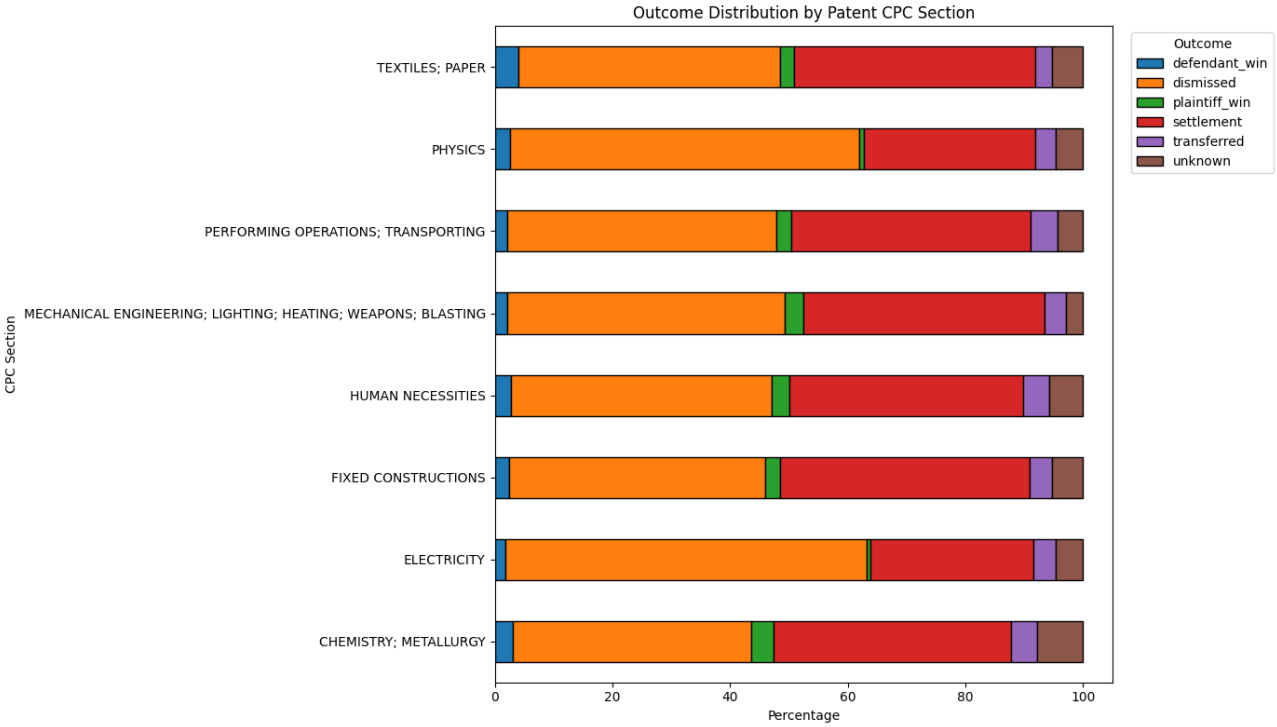

Figure 9 *Outcome distribution by patent CPC section*

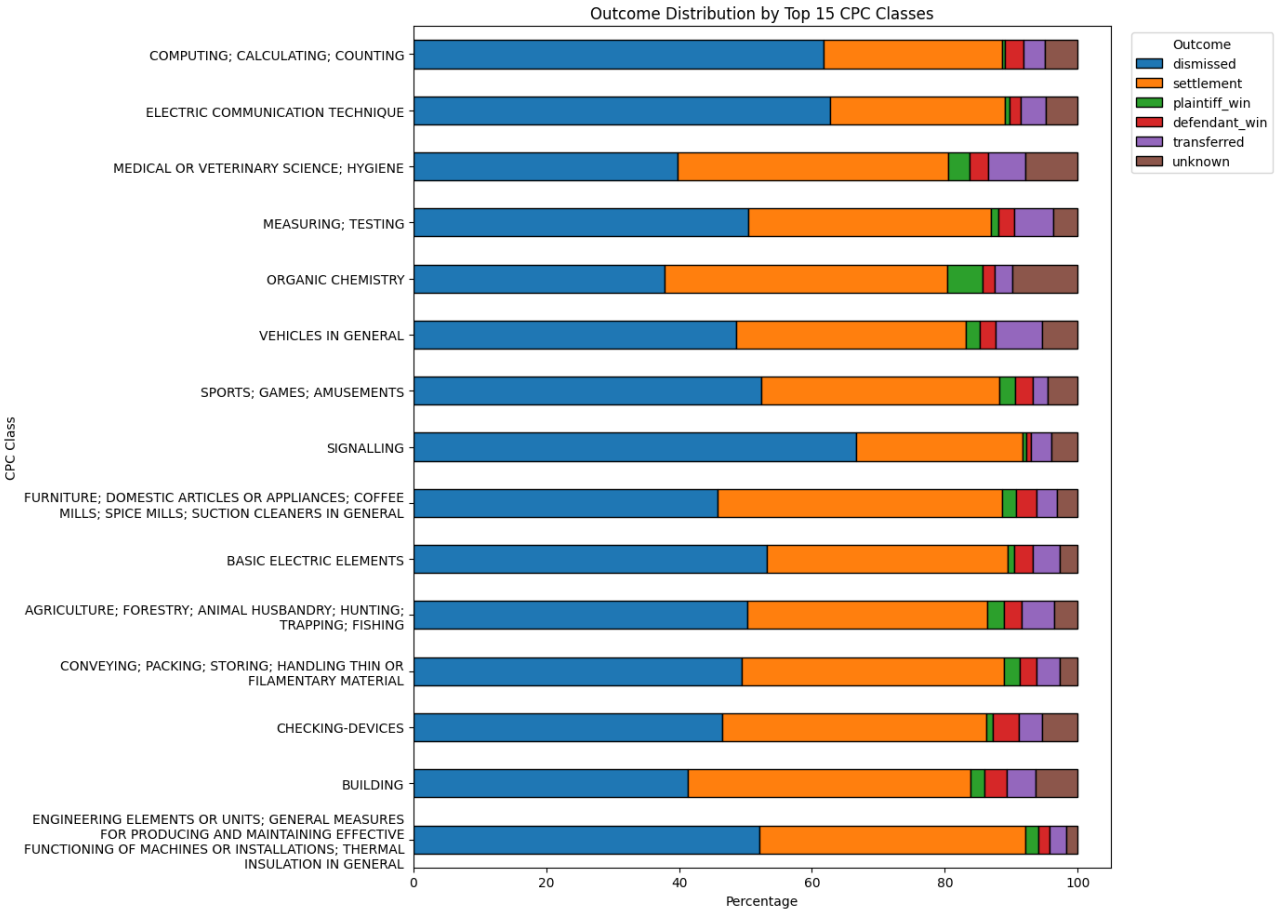

Figure 10 *Outcome distribution by top 15 CPC classes*

Figure 11 *Temporal outcome trends for Physics and Electricity patents, 2003-2020*

The aggregate dismissal increase documented in the previous section was not uniform across technology areas. Figure 11 shows temporal trends for Physics and Electricity patents specifically.

3.1.6 Venue Patterns

Patent litigation concentrated in a small number of district courts, and venue selection shifted dramatically over the study period. Figure 12 displays filing counts for major patent venues from 2003 to 2020.

The Eastern District of Texas grew from 20 cases in 2003 to 1,576 cases in 2015, becoming the dominant patent venue. Filings then declined sharply, falling to 89 cases by 2019. The District of Delaware absorbed much of this displaced volume, growing from approximately 200 cases in 2015 to 456 in 2019. California districts remained relatively stable throughout, each handling 100-200 cases annually.

Outcome distributions differed across major venues. Eastern Texas shows a 71.1% dismissal rate and only 20.2% settlement rate, while Northern Illinois exhibits the opposite pattern with 47.6% dismissals and 43.2% settlements. The California districts fall between these extremes, with dismissal rates around 46-50% and settlement rates of 33-39%.

The high dismissal rate in Eastern Texas reflects case composition rather than venue-specific effects. During the peak filing years (2012-2016), 91% of Eastern Texas cases involved Physics or Electricity patents, up from 61% in the earlier period (2003-2010). This concentration of software

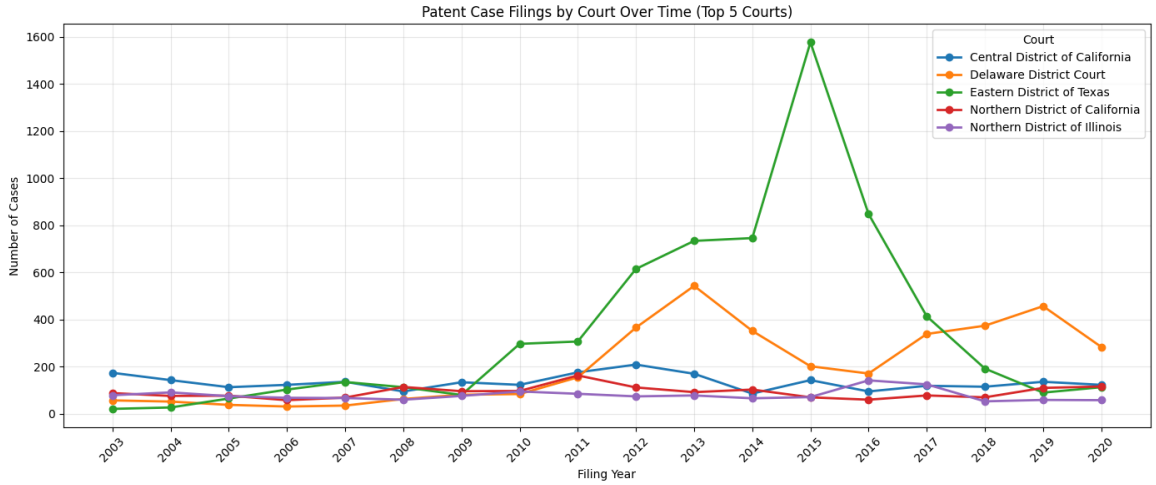

Figure 12 *Patent case filings by court over time*

and electronics patents meant litigants there were disproportionately exposed to the technology-specific dismissal patterns documented in the previous section. Delaware maintained a more diversified portfolio with 67% Physics/Electricity patents and meaningful shares of Chemistry (8%) and Human Necessities (18%). Central California was the most balanced with only 47% in software-related categories.

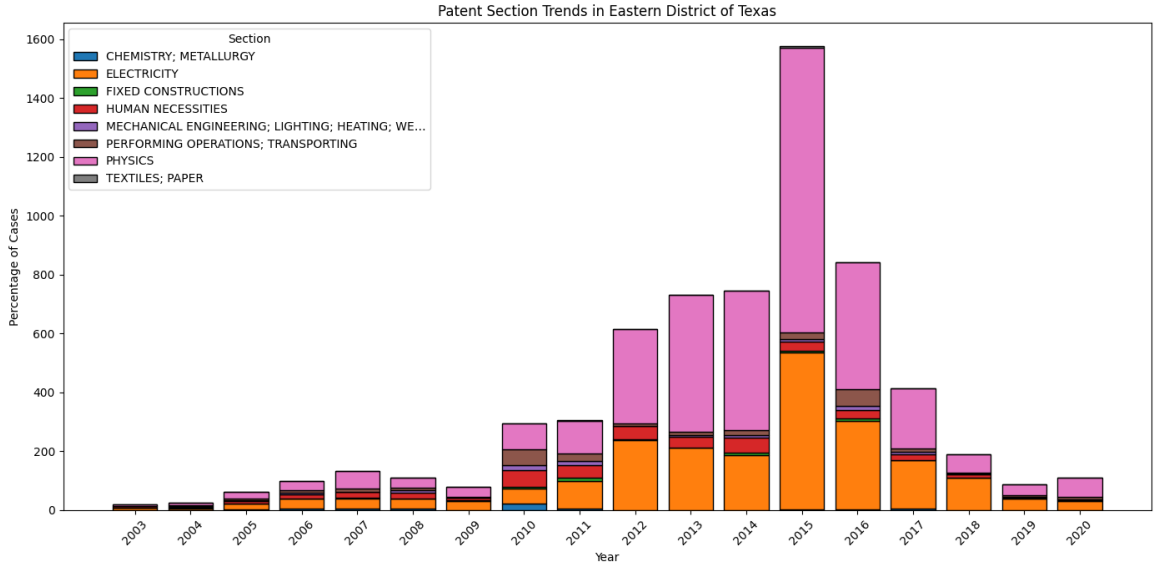

Figure 13 *Patent technology composition in the Eastern District of Texas over time*

A caveat applies to distinguishing dismissals from settlements across all venues. Validation revealed that approximately 15% of extraction errors involved cases classified as dismissed that may have actually settled. This ambiguity reflects genuine interpretive difficulty rather than purely methodological failure. Settlements in federal court are often effectuated through stipulated dismissals without explicit mention of the underlying agreement in the docket record. A dismissal with prejudice, absent settlement language, could reasonably be classified either way. This ambiguity likely inflates dismissal rates and deflates settlement rates across all venues, though it does not systematically bias comparisons between them.

3.2 Model Results

3.2.1 Experimental Setup and Data Summary

After excluding cases with unknown or transferred outcomes and all other cases where the LLM assigned medium or low confidence to the extraction. This confidence filter removed cases where docket text left outcome classification ambiguous, particularly dismissals that may have resulted from undocumented settlements. The resulting sample contained 25,730 cases. Table 6 summarizes the data at each litigation stage.

Stage	Train	Val	Test	Positive Rate	Imbalance Ratio
Survival	15,438	5,146	5,146	42.04%	1.38
Settlement	6,490	2,179	2,176	11.29%	7.85
Adjudicated	733	252	245	40.79%	1.45

Table 6 *Sample sizes and class distribution by litigation stage. Imbalance ratio is the majority class count divided by the minority class count.*

The survival and adjudicated stages are relatively balanced, while the settlement stage exhibits substantial skew with an imbalance ratio of 7.85. The sample size decreases sharply across stages, reflecting the litigation funnel where most disputes terminate before judgment. The adjudicated stage contains only 245 test cases, which affects the precision of performance estimates at this stage.

Model performance was evaluated using multiple metrics to capture different aspects of predictive quality. ROC-AUC measure discrimination, the model's ability to rank positive cases above negative cases, regardless of the classification threshold. We report 95% bootstrap confidence intervals for AUC to quantify uncertainty, using 1,000 bootstrap iterations. Precision-Recall AUC (PR-AUC) when classes are imbalanced, as at the settlement stage. The Matthews Correlation Coefficient (MCC) summarizes confusion matrix performance in a single stage value bounded between -1 and 1, with 0 representing random prediction.

For probability calibration, we report the Brier Skill Score (BSS), which compares model performance against predicting the base rate for all cases. A BSS of 0 indicates no improvement over the base rate baseline, positive values indicate better calibration, and negative values indicate that the model's probability estimates are worse than simply predicting the prevalence. Lastly, lift metrics quantify practical utility by measuring how much better the model performs when focusing on its highest-confidence predictions compared to random selection.

3.2.2 Model Performance Comparison

Table 7 shows classification performance across all model stages.

Stage	Model	AUC	PR-AUC	MCC	BSS	$F1_{neg}$	$F1_{pos}$
Survival	Baseline	0.500	0.423	0.000	0.000	0.732	0.000
	Logistic Regression	0.622	0.517	0.184	0.022	0.591	0.572
	Random Forest	0.767	0.679	0.386	0.208	0.743	0.642
	XGBoost	0.771	0.684	0.385	0.112	0.648	0.680
Settlement	Baseline	0.500	0.113	0.000	0.000	0.940	0.000
	Logistic Regression	0.582	0.145	0.091	-1.457	0.445	0.226
	Random Forest	0.635	0.186	0.122	-1.528	0.691	0.248
	XGBoost	0.716	0.310	0.210	0.057	0.892	0.310
Adjudicated	Baseline	0.500	0.400	0.000	0.000	0.750	0.000
	Logistic Regression	0.598	0.510	0.138	-0.024	0.498	0.559
	Random Forest	0.778	0.740	0.429	0.235	0.784	0.641
	XGBoost	0.767	0.712	0.397	0.180	0.768	0.628

Table 7 *Model performance across litigation stages. BSS = Brier Skill Score. $F1_{neg}$ and $F1_{pos}$ are F1 scores for the negative and positive classes respectively.*

In the survival stage, tree-based ensemble methods substantially outperformed logistic regression in predicting whether cases survived dismissal. XGBoost achieved the highest AUC of 0.771 (95% CI: 0.757-0.783), with Random Forest close behind at 0.767. Logistic regression lagged at 0.622. The MCC values of 0.386 for both ensemble methods indicated moderate predictive power beyond chance. Random Forest achieved better probability calibration (BSS = 0.208) than XGBoost (BSS = 0.112), despite similar discrimination. The F1 score show balanced performance across classes for XGBoost (0.648 and 0.680), while Random Forest favored the negative class (0.743 vs. 0.642).

Predicting which cases reach adjudication provided more difficult. XGBoost achieved an AUC of 0.716 (95% CI: 0.684-0.748), outperforming Random Forest (0.635) and logistic regression (0.582). The severe class imbalance required threshold adjustment: XGBoost used and optimized threshold of 0.21 rather the default 0.50.

The F1 score reveal the challenge of minority class prediction at this stage. Even XGBoost, the best performing model, achieved only 0.310 for the positive class (cases reaching adjudication) compared to 0.892 for the negative class (settlement). Logistic regression and Random Forest produced negative Brier Skill Scores (-1.457 and -1.528) meaning their probability estimates performed worse than predicting the base rate. XGBoost maintained a positive BSS of 0.057, indicating its probabilities provided some information beyond the base rate.

Among cases reaching adjudication, Random Forest achieved the best discrimination with an AUC of 0.778 (95% CI: 0.718-0.834), slightly exceeding XGBoost at 0.757. The MCC of 0.429 for Random Forest represents the strongest predictive signal across all stage-model combinations. Both ensemble methods achieved positive Brier Skill scores and Reasonable F1 balance across classes.

The wide confidence interval at this stage (0.718-0.834, a span of 0.116) reflects the small test

sample of 245 cases. While the point estimates suggest meaningful predicative power, the uncertainty is too large to support a reliable ranking among the ensemble methods.

Logistic regression consistently underperformed ensemble methods across all stages, with the largest gap at the settlement stage (0.582 vs. 0.716). This suggests litigation outcomes depend on feature interactions that linear models cannot capture. Random Forest and XGBoost performed similarly at the survival and adjudication stages but diverged at the settlement stage, where XGBoost can place more emphasis on the cases that are repeatedly misdirected, improving identification of the minority that ultimately resist settlement.

3.2.3 Feature Importance Analysis

We analyzed feature importance using SHAP values computed on the best performing models at each stage: XGBoost for survival and settlement, Random Forest for Adjudicated. SHAP values decompose each prediction into feature contributions. We used mean absolute SHAP across the test set as our importance measure, which captures how much each feature contributes to predictions regardless of direction.

To check whether the rankings are stable, we recomputed importance on subsamples and measured rank correlation. Stability was high at all stages (Spearman ρ= 0.9946 for survival, 0.993 for settlement, 1.0 for adjudicated). The perfect correlation at the adjudicated stage reflects the small sample (245 cases) and the dominance of technology classification, which holds top rank in every subsample.

Table 8 shows the ten most important features at each stage.

For the survival stage, The top predictors are predominantly procedural and contextual rather than patent characteristics. Attorneys count, jury demand, and venue indicators occupy the top positions. Technology classification and patent text features appear in the top ten but contribute less than litigation context variables. This pattern suggests that early case outcomes depends more on who is litigating and where tan on the technical merits of the patent.

The settlement stages features shifted towards patent characteristics. Related cases, patent complexity measures (description length, prosecution duration), and citation counts. Attorney variables remain present but are less prominent than at the survival stage. This shift suggest that once a case makes it past the early procedural hurdles, settlement decisions depend more on how strong and complex the patent appears than on who the lawyers are.

In the adjudicated stage, technology classification dominates with importance (0.402) nearly double the next feature. This might reflect the large variation in plaintiff win rates across technology sectors mentioned in section 3.1.6. Attorney related variables (total attorneys, defendant experience, attorney count imbalance, local counsel rations) occupy most of the remaining top positions.

Patterns across stages are consistent. Attorney counts and district identifiers appear at every stage, indicating that litigation structure and venue remain predictive through the process. patent characteristics become more prominent at the settlement stage. Technology classification grows in importance as cases move forward, and it is the dominant signal at the adjudication stage.

Stage	Feature	Mean |SHAP|
Survival	total_attorneys	0.327
	jury_demand	0.254
	is_edtx	0.195
	district_id	0.131
	plaintiff_side_attorney_count	0.104
	cpc_code	0.094
	plaintiff_defendant_attorney_ratio	0.057
	cpc_avg_processing_days	0.056
	description_chars_per_claim	0.039
	patent_remaining_life_years	0.034
Settlement	has_related_case	0.319
	description_chars_per_claim	0.163
	total_attorneys	0.145
	cpc_avg_processing_days	0.144
	summary_text_length	0.140
	patent_age_years_at_filing	0.134
	num_times_cited_by_us_patents	0.133
	patent_processing_days	0.132
	district_id	0.129
	detail_desc_length	0.125
Adjudicated	cpc_code	0.402
	total_attorneys	0.308
	total_prior_defendant_side_cases	0.218
	attorney_count_imbalance	0.178
	patent_remaining_life_years	0.126
	jury_demand	0.126
	foreign_citation_pct	0.125
	plaintiff_local_ratio	0.124
	district_id	0.120
	team_local_counsel_ratio	0.097

Table 8 *Top 10 features by mean absolute SHAP value at each litigation stage.*

3.2.4 Embedding Contribution

Adding PaECTER embeddings to the features set reduced the predictive performance across all litigation stage. Table 9 compares XGBoost results with and without text embeddings.

The settlement stage showed the largest degradation, with Brier Skill Score falling to -0.070, which worse than predicting the base rate. The 5,120 embedding dimensions relative to 6,490 training cases likely created overfitting, and structured patent features may already capture the litigation-relevant signal in patent text.

3.2.5 Temporal Validation

To assess whether models trained on historical cases generalize to future litigation, we conducted temporal validation using a 2016 cutoff. This was done because of the Eastern District of

Stage	AUC (no emb)	AUC (with emb)	Δ
Survival	0.771	0.765	-0.006
Settlement	0.716	0.688	-0.028
Adjudicated	0.767	0.754	-0.013

Table 9 *Effect of PaECTER embeddings on XGBoost performance. Δ reports the change in AUC when embeddings are included (with emb $-$ no emb).*

Texas historic litigation data present in Table 12. Cases filed before 2016 served as training data while cases filed in 2016 and later served as the test set. Table 10 presents the results.

Stage	Model	Train N	Test N	AUC	MCC
Survival	XGBoost	16,282	4,302	0.730	0.331
	Logistic	16,282	4,302	0.588	0.116
Settlement	XGBoost	7,064	1,602	0.562	0.042
	Logistic	7,064	1,602	0.536	0.050
Adjudicated	XGBoost	841	137	0.811	0.480
	Logistic	841	137	0.623	0.202

Table 10 *Temporal validation results using a pre-2016 training set and post-2016 test set.*

Settlement prediction suffered the largest degradation, with XGBoost AUC dropping from 0.716 to 0.562, while survival prediction declined more modestly from 0.771 to 0.730. Adjudication prediction maintained performance (AUC 0.811), possibly because adjudicated outcomes produce less ambiguous docket text than settlements or dismissals.

4 Discussion

Attorney-related features dominated survival prediction. Total attorney count, plaintiff-side attorney count, and attorney ratios ranked among the top predictors, with venue and procedural indicators also contributing. Patent characteristics appeared among important features but contributed less than litigation representing features. We could argue that early case survival strongly depends on procedural execution: resourced, experienced counsel may file more robust complaints and respond more effectively to early motions. An alternative interpretation could be that attorney counts proxy for case value and party resources rather than causing survival directly.

Patent characteristics gained prominence at the settlement stage. Related case indicators, filing duration, citation counts, and complexity ranked highly. The best performing model captured limited predictive signal, with XGBoost achieving an AUC of 0.716, implying that substantial outcome-relevant variation is not explained by the current feature set. This results is informative at this stage because it motivates the inclusion of additional predictors that would represent the defending side - giving contrasting context for the plaintiff (patent holder).

For the adjudication stage, CPC code was the most important feature, indicating that the adjudication model relied heavily on sectoral differences in plaintiff success rates. This was already passively hinted in the LLM-extracted outcomes whilst investigating outcome variance in different cpc sections: chemistry, which had 5% of cases that plaintiffs won, in contrast to 1% of cases in the electronics section. An explanation could be that different patent sections have different litigation processes. Attorney experience and other litigation context variables were also dominant in this stage.

Temporal validation revealed that patent litigation dynamics shifted substantially during the study period. When training on pre-2016 cases and testing on later filings, settlement prediction dropped from 0.716 to 0.562. Survival prediction degraded from 0.771 to 0.730. Only the adjudication stage maintained performance, with AUC actually increasing slightly to 0.811. This increase could be attributed to the smaller test set.

This degradation is not a methodological artifact. It reflects documented changes in the litigation landscape.

The most prominent example, and the foundation for choosing the temporal split range was found through analyzing the top sections outcome distributions throughout the filing years (see Figure 11) and the top case filings by venue (see Figure 12). In particular, the Eastern District of Texas, which handled 1,576 single-patent cases in 2015 saw filings collapse to 89 cases by 2019. An exploratory review of secondary sources and practitioner commentary suggests that this period coincides with major venue and doctrine developments in U.S. patent litigation. However, formally attributing the observed shift is outside the scope of this study.

Adjudication prediction maintained performance across the temporal split (AUC 0.811). Adjudicated outcomes produce less ambiguous docket text than settlements or dismissals, which may explain the stability.

These finding have implications for applied use. Prediction models trained on historical lit-

igation data require periodic retraining as legal doctrine, venue rules, and filing patterns evolve. A model calibrated to pre-2016 dynamics would misled users making decisions in 2018. The magnitude of drift, particularly for the settlement predictions, indicates that litigation outcome models cannot be treated as static tools.

5 Limitations

Several constraints bound the conclusions that can be drawn.

The reliance on docket summary text limits outcome granularity. Settlements appear only when explicitly mentioned: some can be assumed through stipulated dismissal indistinguishable from other dismissal without access to underlying documents. The validation results confirmed this ambiguity. Accessing more comprehensive case information would require retrieving full docket documents from PACER, which puts a monetary barrier. There are some platforms that provide limited court documents for free . Attempts to validate through free sources did not show promise, because CourtListener provides only limited access to underlying docket documents [14].

The restriction to single-patent cases exclude 70% of patent litigation. Multi-patent cases may differ systematically in complexity, stakes, and outcome patterns. Aggregating features across multiple patents introduces modeling assumptions that complicate interpretation. The findings may not extend to portfolio litigation strategies employed by large technology companies. Whilst investigating multi-patent cases it was found that the created paper trail, in such cases, is generally longer than in single-patent cases.

Patent text embeddings failed to improve prediction. The embedding dimensions relative to training sample size likely created overfitting. Structured patent features may already capture the litigation relevant signal present in patent summaries and claims. This negative result suggests that either raw patent text, at least as represented by current embeddings approaches, does not add predictive value beyond what the metadata contains.

The study covers only U.S. federal district court litigation. Patent disputes in other jurisdictions, particularly Europe, are less well characterized at scale, in part due to fragmented institutional structure and multilingual documentation that complicate dataset construction and harmonized measurement. Thus, weather the patterns observed here generalize across legal systems remains unknown.

6 Future Work

Improving outcome extraction remains the biggest hurdle for this subject. Newer language models with expanded context windows and stronger reasoning capabilities should handle complex docket histories more reliably. But developing dedicated extraction methods should also be welcomed, since docket entries exist not only in the patent litigation domain, solving this could potentially be applied to other fields. Furthermore, if using large language models to extract the outcomes of litigation processes, the outcome taxonomy could be atomized further for this specific domain.

The feature set captures attorney characteristics for both sides but lacks defendant party information beyond legal representation. Company size, financial resources, and whether the defendant is a repeat target of patent litigation likely influence both settlement behavior and litigation outcomes. Linking defendant names to corporate databases would provide this contrasting context.

Temporal validation showed performance degradation when training on pre-2016 cases and testing on later filings. What caused this degradation, and how to anticipate similar shifts, remains unclear.

7 Conclusions

Patent litigation cases contained in the USPTO dataset don't have case outcomes encoded in accessible way (where present - they are mostly encoded via PACER identifiers). We used an LLM to extract the outcomes from the case docket text, that contained the summaries of the actions made in court. We then linked patent characteristics, built attorney metrics for the representing parties and used them to predict the extracted outcomes.

The hierarchical classification approach revealed distinct patterns for each outcome grouping. Attorney characteristics, particularly total attorney count and plaintiff-side representation were the most prevalent in the dismissal versus non-dismissal distinction along with venue indicators. When separating settlements from adjudication outcomes, related case indicators ranked highest, followed by patent characteristics like description length, processing duration, and citation counts, though attorney features remain present. CPC code was the strongest predictor for plaintiff versus defendant wins. Random Forest achieved the best results for adjudication (AUC 0.778), XGBoost for survival (0.771) and XGBoost for settlement (AUC 0.716). Patent text embeddings did not help at any level.

The results also point to limits that matter for interpretation and applied use. Temporal validation indicates performance degradation when models trained on earlier filings are applied to later cases, particularly for settlement prediction, which suggests instability in the between features and outcomes over time.

The extraction approach produced 25,730 labeled cases from public data. Prior academic work relied on hand-coded samples of a few thousand cases or proprietary sources. The dataset produced here makes outcome prediction makes outcome prediction research feasible without proprietary access and without prohibitive manual labeling effort.

Acknowledgments

The author utilized Claude Opus 4.5 [5], a large language model, to enhance text clarity in sections addressing scientific concepts and for formatting tables.

The prompt for clarifying text: You are my research writing assistant, don't use exaggerated language, don't enhance my findings. Please push back on my findings.

The prompt for table formatting: Please provide the correctly formatted table from this data.

References

[1] K. Ahn, A. Trujillo, J. Gibbons, C. L. Bennett, G. Anderson. “Settled: Patent characteristics and litigation outcomes in the pharmaceutical industry.” English (US). In: *International Review of Law and Economics* 76 (2023). ISSN: 0144-8188. https://doi.org/10.1016/j.irle.2023.106169.

[2] A. T. Alan C. Marco, A. Toole. “The USPTO Patent Litigation Patent Litigation Data from US District Court Electronic Records (1963-2015).” In: *SSRN Electronic Journal* (2017). URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2942295.

[3] J. R. Allison, M. A. Lemley, K. A. Moore, R. D. Trunkey. “Valuable Patents.” In: *Georgetown Law Journal* 92 (2004), pages 435–479. URL: https://heinonline.org/HOL/Page?handle=hein.journals/glj92&id=453.

[4] R. M. Andrew Toole, T. M. Sichelman. “Technical Documentation for Patent Litigation Docket Reports Data, 1963-2020.” In: *SSRN Electronic Journal* (2024). URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4780166.

[5] Anthropic. URL: https://claude.ai/%7D.

[6] D. L. Atkins, T. Fishman. *Patent Litigation Strategies Handbook*. n.d. URL: https://www.americanbar.org/products/inv/book/447746348/.

[7] T. Bar, J. Kalinowski. “Patent validity and the timing of settlements.” In: *International Journal of Industrial Organization* 67 (2019), page 102535. ISSN: 0167-7187. https://doi.org/https://doi.org/10.1016/j.ijindorg.2019.102535. URL: https://www.sciencedirect.com/science/article/pii/S0167718719300633.

[8] J. Breton, M. M. Billami, M. Chevalier, H. T. Nguyen, K. Satoh, C. Trojahn, M. M. Zin. “Leveraging LLMs for legal terms extraction with limited annotated data.” In: *Artificial Intelligence and Law* (2025). ISSN: 1572-8382. https://doi.org/10.1007/s10506-025-09448-8. URL: https://doi.org/10.1007/s10506-025-09448-8.

[9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, et al. “Language Models are Few-Shot Learners.” In: *CoRR* abs/2005.14165 (2020). URL: https://arxiv.org/abs/2005.14165.

[10] Y.-c. Chang, K.-P. Chen, C.-C. Lin. *Attorney and Judge Experience in Torts Litigation: An Empirical Study*. Draft, accessed via NYU Law website. 2016. URL: https://www.law.nyu.edu/sites/default/files/upload_documents/Yun-Chien_Chang_Attorney_and_Judge_Experience_in_Torts_Litigation_160125-1.pdf.

[11] S.-H. Chen, C.-Y. Lai. “Patent Litigation Prediction Using Machine Learning Approaches.” In: *Artificial Intelligence and Law*. Springer Nature Switzerland, 2023, pages 389–395. URL: https://link.springer.com/chapter/10.1007/978-3-031-36004-6_53.

[12] K. Cremers. “Settlement during patent litigation trials. An empirical analysis for Germany.” In: *Journal of Technology Transfer* 34 (2009), pages 182–195. URL: https://link.springer.com/article/10.1007/s10961-007-9066-7.

[13] N. Z. Dina, S. D. Ravana, N. Idris. “Legal Judgment Prediction using Natural Language Processing and Machine Learning Methods: A Systematic Literature Review.” In: *Sage Open* 15.2 (2025), page 21582440251329663. https://doi.org/10.1177/21582440251329663. URL: https://doi.org/10.1177/21582440251329663.

[14] Free Law Project. *RECAP Archive (CourtListener)*. https://www.courtlistener.com/recap/. Accessed 2026-01-03.

[15] M. Galanter. “Why the ”Haves” Come out Ahead: Speculations on the Limits of Legal Change.” In: *Law Society Review* 9.1 (1974), pages 95–160. ISSN: 00239216, 15405893. URL: http://www.jstor.org/stable/3053023 (viewed 2025-12-16).

[16] M. Ghosh, M. E. Rose, S. Erhardt, E. Buunk, D. Harhoff. “PaECTER: Patent-level Representation Learning using Citation-informed Transformers.” In: (2025). URL: https://arxiv.org/abs/2402.19411.

[17] Google Patents Public Data. *BERT for Patents*. 2019. URL: https://github.com/google/patents-public-data/blob/master/models/BERT%20for%20Patents.md.

[18] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, et al. *The Llama 3 Herd of Models*. 2024. URL: https://arxiv.org/abs/2407.21783.

[19] B. Hamzehzadeh. “Repeat Player vs. One-Shotter: Is Victory all that Obvious.” In: *Hastings Business Law Journal* 6.1 (2010), pages 239–270. URL: https://repository.uclawsf.edu/hastings_business_law_journal/vol6/iss1/6.

[20] A. Izzidien, H. Sargeant, F. Steffek. “LLM vs. Lawyers: Identifying a Subset of Summary Judgments in a Large UK Case Law Dataset.” In: No. 10/2024 (2024). https://doi.org/https://doi.org/10.2139/ssrn.4746305. URL: https://ssrn.com/abstract=4746305.

[21] D. M. Katz, M. J. Bommarito II, J. Blackman. “A general approach for predicting the behavior of the Supreme Court of the United States.” In: *PLOS ONE* 12 (2017), pages 1–18. https://doi.org/10.1371/journal.pone.0174698. URL: https://doi.org/10.1371/journal.pone.0174698.

[22] J. O. Lanjouw, M. Schankerman. “Characteristics of Patent Litigation: A Window on Competition.” In: *RAND Journal of Economics* 32.1 (2001), pages 129–151. URL: https://www.jstor.org/stable/2696401.

[23] J. O. Lanjouw, M. Schankerman. “Patent Quality and Research Productivity: Measuring Innovation with Multiple Indicators.” In: *The Economic Journal* 114.495 (2004), pages 441–465. URL: https://doi.org/10.1111/j.1468-0297.2004.00216.x.

[24] J. O. Lanjouw, M. Schankerman. “Protecting Intellectual Property Rights: Are Small Firms Handicapped?” In: *Journal of Law and Economics* 47.1 (2004), pages 45–74. https://doi.org/10.1086/380476. URL: https://chicagounbound.uchicago.edu/jle/vol47/iss1/3/.

[25] B. J. LOVE. “AN EMPIRICAL STUDY OF PATENT LITIGATION TIMING: COULD A PATENT TERM REDUCTION DECIMATE TROLLS WITHOUT HARMING INNOVATORS?” In: *University of Pennsylvania Law Review* 161 (2013), pages 1309–1359. URL: https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=1388&context=penn_law_review.

[26] K. T. McGuire. “Repeat Players in the Supreme Court: The Role of Experienced Lawyers in Litigation Success.” In: *The Journal of Politics* 57.1 (1995), pages 187–196. https://doi.org/10.2307/2960277. URL: https://www.journals.uchicago.edu/doi/10.2307/2960277.

[27] M. Medvedeva, M. Wieling, M. Vols. “Rethinking the field of automatic prediction of court decisions.” In: *Artificial Intelligence and Law* (2022). URL: https://doi.org/10.1007/s10506-021-09306-3.

[28] M. Medvedeva, M. Vols, M. Wieling. “Using Machine Learning to Predict Decisions of the European Court of Human Rights.” In: *Artificial Intelligence and Law* 28 (2020), pages 237–266. URL: https://link.springer.com/article/10.1007/s10506-019-09255-y.

[29] OpenAI, : S. Agarwal, L. Ahmad, et al. *gpt-oss-120b gpt-oss-20b Model Card*. 2025. URL: https://arxiv.org/abs/2508.10925.

[30] Patent Progress. *Overview of a Patent Litigation*. n.d. URL: https://patentprogress.org/basics/overview-of-a-patent-litigation/.

[31] PatentsView. *PatentsView Data Download Tables*. Accessed: 2025. 2025. URL: https://patentsview.org/download/data-download-tables.

[32] *pgvector Documentation*. https://access.crunchydata.com/documentation/pgvector/latest/pdf/pgvector.pdf. Accessed 2026-01-03.

[33] D. L. Schwartz. “Courting Specialization: An Empirical Study of Claim Construction Comparing Patent Litigation Before Federal District Courts and the International Trade Commission.” In: *William & Mary Law Review* 51.1 (2009), pages 121–196. URL: https://wmlawreview.org/sites/default/files/Schwartz_final.pdf.

[34] D. Shu, H. Zhao, X. Liu, D. Demeter, M. Du, Y. Zhang. “LawLLM: Law Large Language Model for the US Legal System.” In: (2024), pages 4882–4889. https://doi.org/https://doi.org/10.1145/3627673.3680020. URL: http://dx.doi.org/10.1145/3627673.3680020.

[35] D. Somaya. “Strategic Determinants of Decisions Not to Settle Patent Litigation.” In: *Strategic Management Journal* 24.1 (2003), pages 17–38. URL: https://doi.org/10.1002/smj.281.

[36] D. Songer, A. Kuersten, E. Kaheny. “Why the Haves Don’t Always Come Out Ahead: Repeat Players Meet Amici CCuriae for the Disadvantaged.” In: *Political Research Quarterly* 53.3 (2000), pages 537–556. https://doi.org/10.1177/106591290005300305. URL: https://doi.org/10.1177/106591290005300305.

[37] H. O. Steffen Juranek. “Predicting Patent Litigation.” In: *International Review of Law and Economics* (2024). URL: https://doi.org/10.1016/j.irle.2024.106228..

[38] D. R. Subinay Adhikary Procheta Sen, K. Ghosh. “A case study for automated attribute extraction from legal documents using large language models.” In: *Artificial Intelligence and Law* (2024). URL: https://link.springer.com/article/10.1007/s10506-024-09425-7.

[39] Supreme Court of the United States. *Markman v. Westview Instruments, Inc., 517 U.S. 370 (1996)*. 1996. URL: https://supreme.justia.com/cases/federal/us/517/370/.

[40] G. Team, T. Mesnard, C. Hardin, R. Dadashi, et al. *Gemma: Open Models Based on Gemini Research and Technology*. 2024. URL: https://arxiv.org/abs/2403.08295.

Appendix A Patent Document Metadata Plots

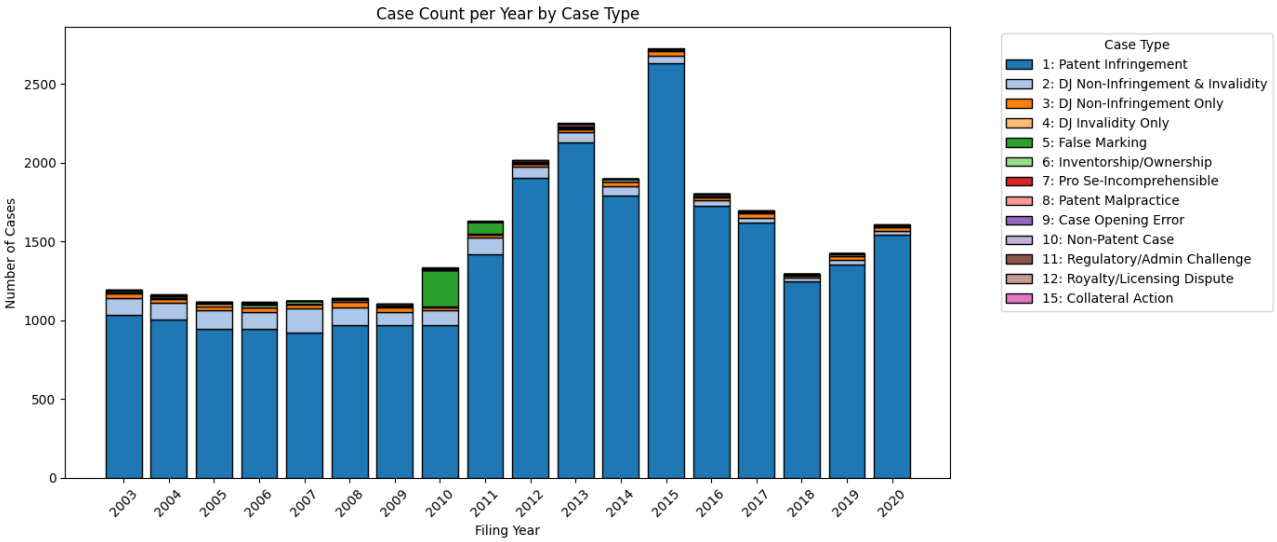

Figure 14 *Case Count per Year by Case Type*

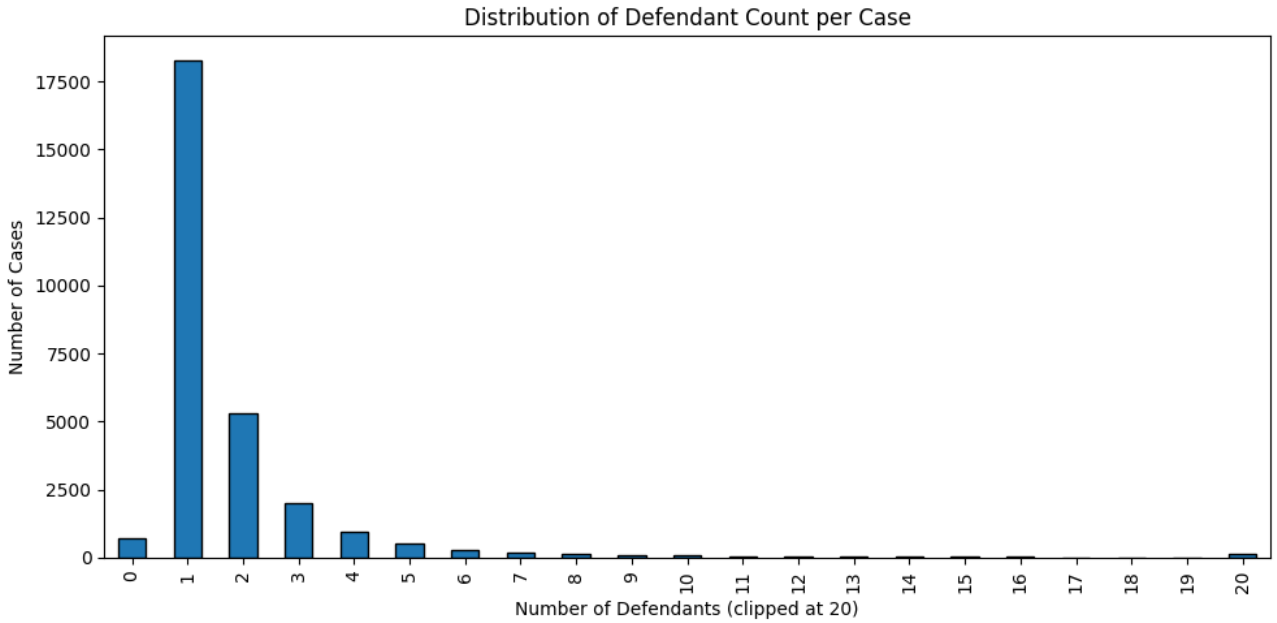

Figure 15 *Distribution of Defendant Count per Case*

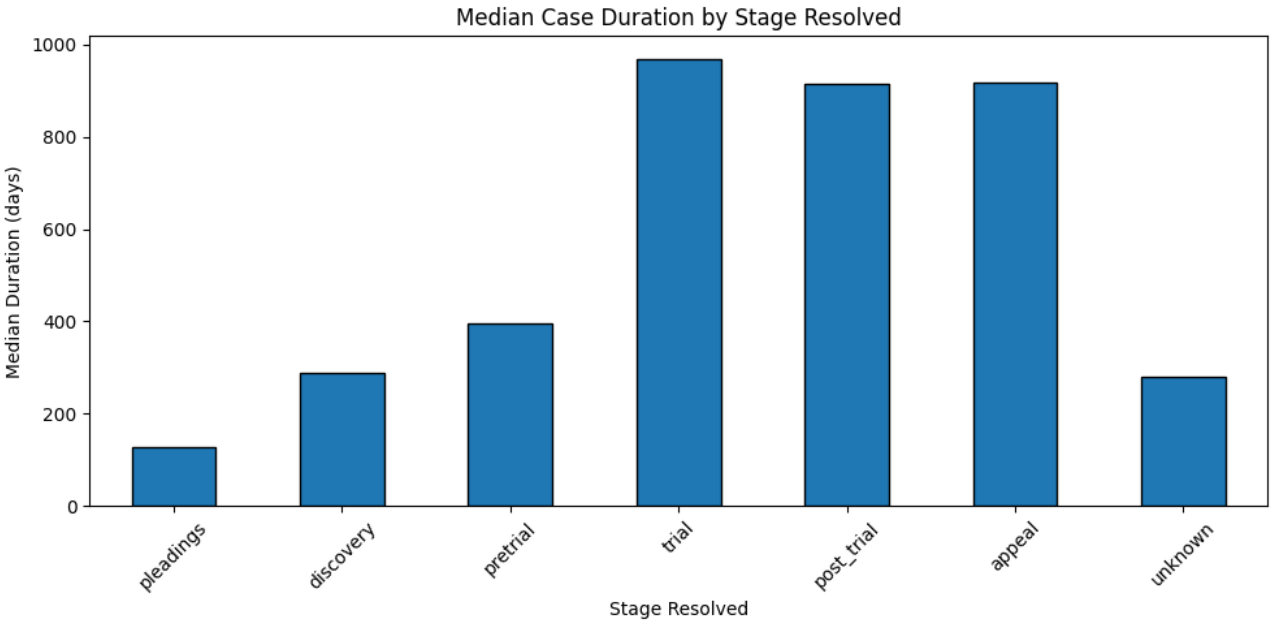

Figure 16 *Median Case Duration by Stage Resolved*

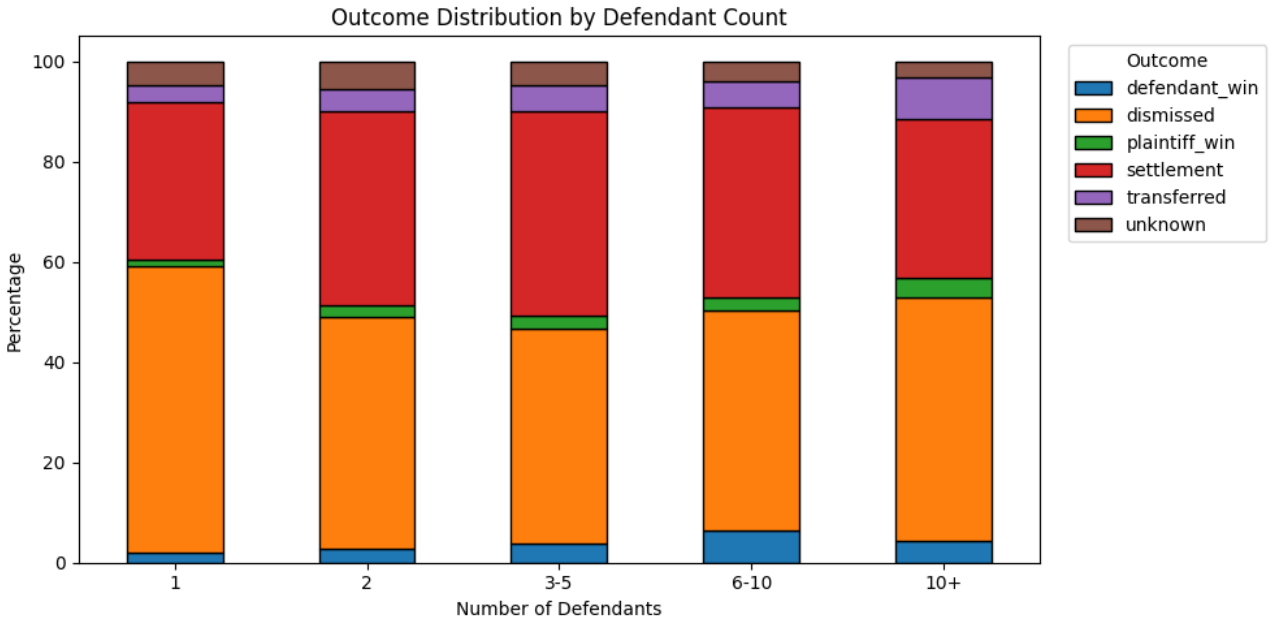

Figure 17 *Outcome Distribution by Defendant Count*

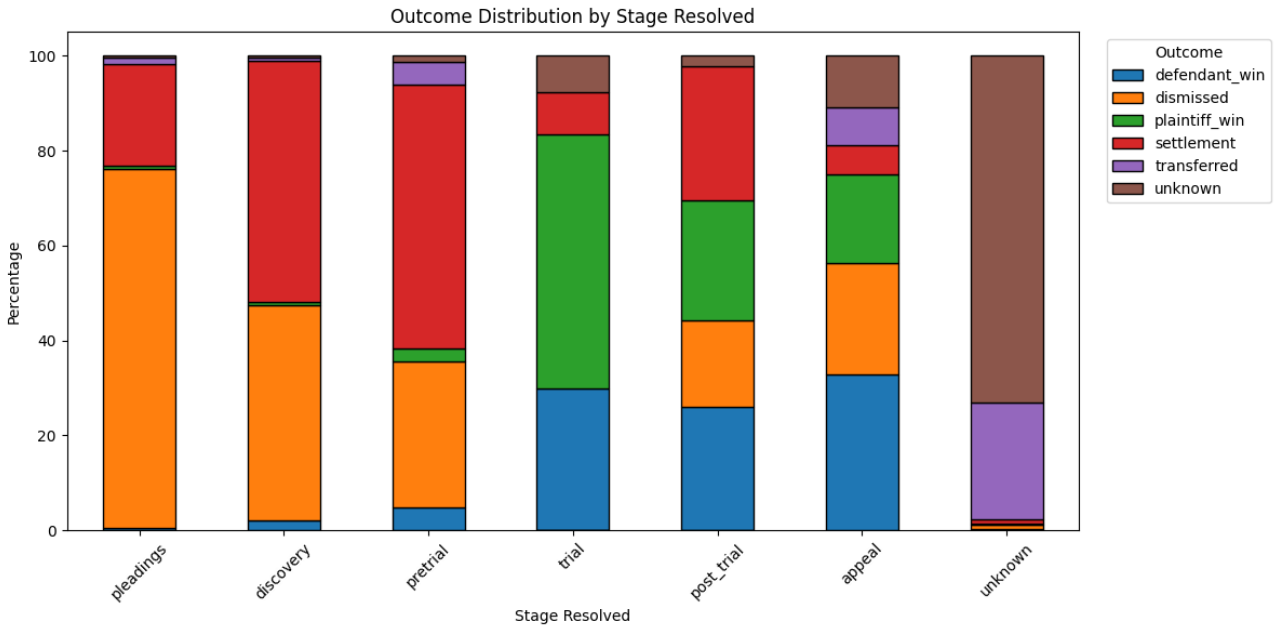

Figure 18 *Outcome Distribution by Stage Resolved*

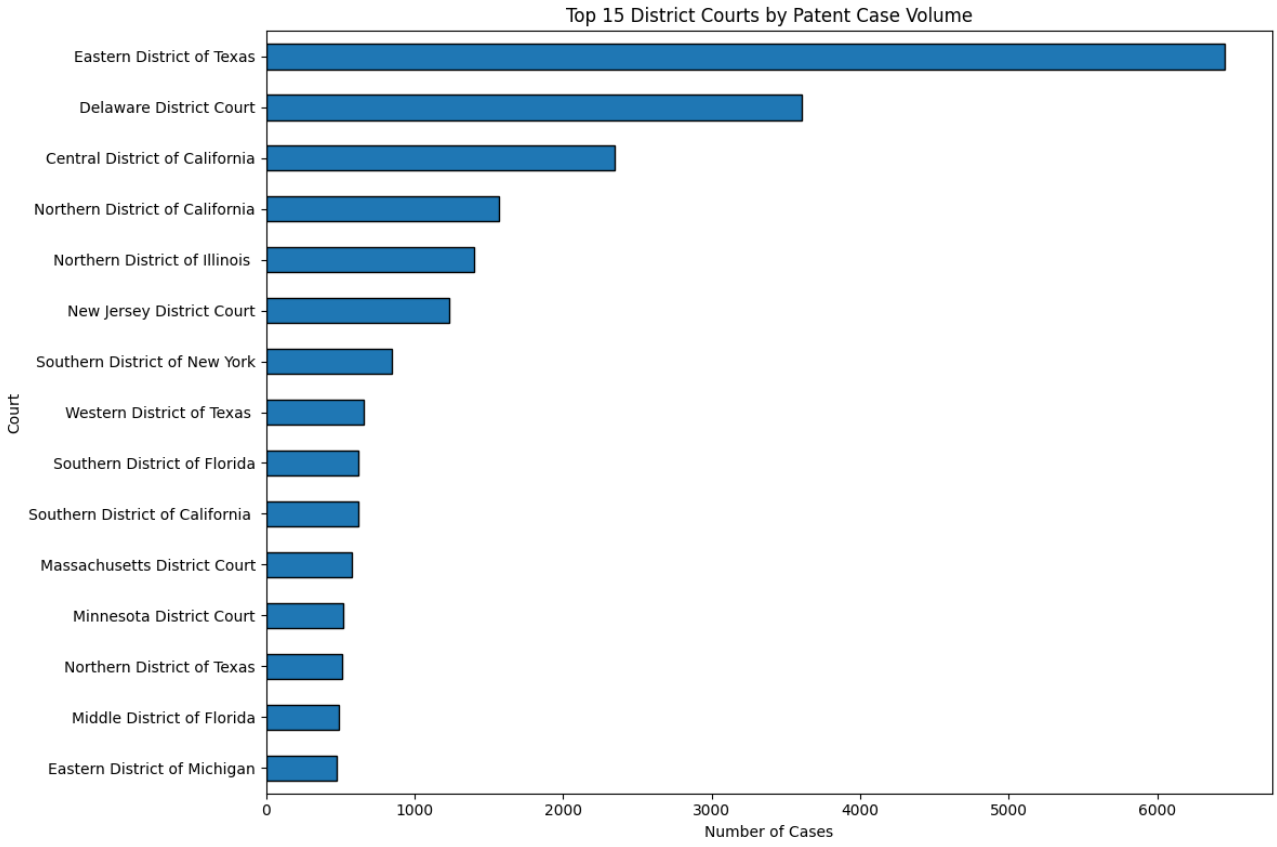

Figure 19 *Top 15 District Courts by Patent Case Volume*

Appendix B Prompt for Docket Outcome Extraction

```
You are a patent litigation analyst. Extract case outcomes from court docket entries.
```

```
Return ONLY a valid JSON object with these fields:

{
  "outcome": "settlement" | "plaintiff_win" | "defendant_win" | "dismissed" | "transferre
  "dismissal_type": "with_prejudice" | "without_prejudice" | null,
  "went_to_trial": true | false,
  "stage_resolved": "pleadings" | "discovery" | "pretrial" | "trial" | "post_trial" | "ap
  "infringement_found": true | false | null,
  "patent_invalidated": true | false | null,
  "injunction_granted": true | false | null,
  "damages_awarded": number | null,
  "filing_date": "YYYY-MM-DD" | null,
  "resolution_date": "YYYY-MM-DD" | null,
  "trial_date_scheduled": "YYYY-MM-DD" | null,
  "judge": "string" | null,
  "magistrate": "string" | null,
  "plaintiff": "string" | null,
  "defendants": ["string"] | null,
  "patent_numbers": ["string"] | null,
  "confidence": "high" | "medium" | "low",
  "reasoning": "string"
}

---

OUTCOME DEFINITIONS:

settlement: Parties reached agreement. Indicators:
- "settlement agreement", "settled", "stipulation of dismissal"
- Dismissal (with or without prejudice) following settlement correspondence
- Consent judgment or consent decree
- Voluntary dismissal after negotiation period

plaintiff_win: Judgment or verdict for plaintiff. Indicators:
- Jury verdict for plaintiff
- Summary judgment granted to plaintiff
- Default judgment against defendant
- Permanent injunction granted

defendant_win: Judgment or verdict for defendant. Indicators:
- Jury verdict for defendant
- Summary judgment granted to defendant
- Summary judgment of non-infringement
```

- Summary judgment of invalidity

dismissed: Court dismissed without settlement. Indicators:
- Dismissed for lack of jurisdiction
- Dismissed for failure to prosecute
- Dismissed as frivolous
- No settlement language present

transferred: Case moved to another court. Indicators:
- Transfer order to another district
- MDL transfer

STAGE DEFINITIONS:

pleadings: Resolved before discovery began (Rule 12 motions, early dismissal)
discovery: Resolved during discovery phase
pretrial: Discovery complete, trial scheduled but not commenced
trial: Resolved during or immediately after trial
post_trial: Resolved after verdict (JMOL, new trial motion, remittitur)
appeal: Resolved at appellate level (Federal Circuit, Supreme Court)

EXTRACTION RULES:

Dates:
- Docket dates use MM/DD/YY format (e.g., 2/13/03 = 2003-02-13)
- filing_date: Date of original complaint
- resolution_date: Date of final dispositive order

Parties:
- Judge: The presiding district judge at resolution
- Magistrate: The final assigned magistrate (use most recent if reassigned)
- Plaintiff: Primary plaintiff or first-named plaintiff
- Defendants: All named defendants as array

Patent numbers:
- Extract any patent numbers mentioned (format: X,XXX,XXX or USXXXXXXX)
- Use null if none found

Damages:

- Extract dollar amount if specified
- Use null if "damages to be determined" or not mentioned

Confidence:
- high: Clear outcome signals, unambiguous resolution
- medium: Outcome determinable but some ambiguity
- low: Outcome inferred with significant uncertainty

Reasoning:
- Cite specific docket entries that support your outcome determination
- Keep to 1-2 sentences

CRITICAL:
- Respond with valid JSON only
- No markdown, no backticks, no text outside JSON
- Use null for fields that cannot be determined from the docket
- Do not invent or assume information not present

Appendix C GitHub Repository

Repository: PanBurst/masters_thesis_code