SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY

DEPARTMENT OF MATHEMATICS

TECHNICAL UNIVERSITY OF MUNICH

Master's Thesis in Mathematics in Data Science

Trend Analysis of NLP Utilization in the Legal Domain within the DACH Region

Henrik Sergoyan

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY

DEPARTMENT OF MATHEMATICS

TECHNICAL UNIVERSITY OF MUNICH

Master's Thesis in Mathematics in Data Science

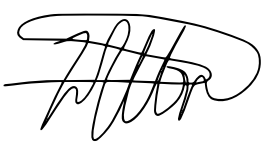
Trend Analysis of NLP Utilization in the Legal Domain within the DACH Region

Trendanalyse der NLP-Nutzung im Rechtsbereich in der DACH-Region

Author:	Henrik Sergoyan
Supervisor:	Prof. Dr. Florian Matthes
Advisor:	Juraj Vladika, M.Sc; Stephen Meisenbacher, M.Sc.
Submission Date:	09.05.2025

I confirm that this master's thesis in mathematics in data science is my own work and I have documented all sources and material used.

Munich, 09.05.2025 Henrik Sergoyan



AI Assistant Usage Disclosure

Introduction

Performing work or conducting research at the Chair of Software Engineering for Business Information Systems (sebis) at TUM often entails dynamic and multi-faceted tasks. At sebis, we promote the responsible use of *AI Assistants* in the effective and efficient completion of such work. However, in the spirit of ethical and transparent research, we require all student researchers working with sebis to disclose their usage of such assistants.

Use of *AI Assistants* for Research Purposes

I have used AI Assistant(s) for the purposes of my research as part of this thesis.

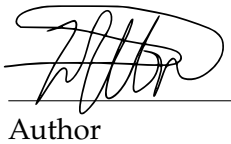
Yes ☑ No ☐

Explanation: Throughout the preparation of this thesis, I employed the ChatGPT as an auxiliary tool for a broad range of scholarly writing tasks. Specifically, it supported iterative refinement of phrasing, grammar, and punctuation; bidirectional translation between English and German; schematic outlining of the Methodology and Results chapters; and the restructuring of lengthy tables so that they conform to page limits while remaining intelligible. I further consulted the model to probe alternative argument flows and to test the logical coherence of complex passages, after which I critically evaluated, edited, and—where necessary—substituted the generated text with my own formulations. All empirical analyses, interpretations, and conclusions originate from my independent research effort, and the final wording, structure, and scholarly contribution are entirely my intellectual property.

I confirm in signing below, that I have reported all usage of AI Assistants for my research, and that the report is truthful and complete.

Munich, 09.05.2025

Location, Date



Author

Acknowledgments

I am profoundly grateful to **Prof. Dr. Florian Matthes** for granting me the opportunity to write my thesis at his chair. My sincere thanks also go to my advisors, **Juraj Vladika** and **Stephen Meisenbacher**, whose unwavering belief in my abilities, together with their extensive expertise in the field, paved the way for my research to flourish. Their constructive feedback were instrumental in refining my approach and elevating the overall quality of my work.

On a personal note, I extend my heartfelt gratitude to my beloved wife, **Seda**, whose steadfast support during the arrival of our newborn son, **Tigran**, allowed me to devote my full attention to this thesis. Her patiance, perseverance and love are and will always remain my greatest inspiration.

Abstract

Context: Natural language processing (NLP) has gained increasing prominence in modern legal work, in large part due to rapid advancements in large language models (LLMs). While academic literature and professional communities alike have highlighted various NLP-driven solutions for legal domains, the lack of a systematic *and* scalable trend analysis persists. As a result, it remains unclear how research priorities evolve and whether they sufficiently align with the most prominent use-case categories in legal practice—particularly those salient within the DACH region. Existing inquiries often center on isolated tasks or case studies, overlooking broader insights into overall publication patterns, category-level coverage, and the day-to-day requirements of practitioners.

Aim: Building on prior work by SEBIS research group which identified core legal NLP use-cases, this thesis provides a detailed examination of contemporary trends and practitioner priorities of those use-cases. Specifically, it seeks to (i) categorize a large corpus of academic publications using a rigorous pipeline, (ii) analyze how these categorization results inform emerging patterns in NLP-driven legal AI, and (iii) validate and contextualize these findings through semi-structured interviews with industry experts in the DACH region.

Approach: A large language model was fine-tuned for a two-step classification pipeline. First, it determines whether each paper discusses at least one legal NLP use-case, achieving a 97% F_1 score in distinguishing relevant from non-relevant publications. For those deemed relevant, the same model identifies which among 31 predefined sub-use-cases apply, yielding an overall 81.3% F_1 score. Subsequently, 12 semi-structured interviews were conducted with attorneys and in-house counsel to rank seven broad legal AI categories and provide insights into day-to-day sub-use-case demands, thereby bridging academic findings with professional realities.

Results: Of the 3,578 screened papers, 988 were classified as discussing legal NLP and analysis of those qualified papers reveals that *Legal Research & Information Management* emerges most prominently, occurring in 53.7% of documents. *Information Processing & Extraction* follows at 52.4%, underscoring the field's enduring focus on advanced retrieval and entity extraction workflows. However, interviews tell a complementary story, where practitioners placed *Document Generation and Assistance* at the apex of their priorities (average rank of 1.42), often citing efficiency gains in contract drafting. They similarly ranked *Legal Research and Information Management* highly (average rank of 3.67), yet underscored gaps between the sub-use-cases academics spotlight (e.g., broad research automation) and the more nuanced tasks they require (e.g., e-Discovery, legislative tracking).

Conclusion: By blending an automated literature review—fueled by large language models—with targeted industry consultations, this thesis elucidates a rapidly growing legal NLP research landscape and reveals the intricacies of translating those scholarly endeavors into practical solutions. The findings indicate that, while academia covers an expansive set of methodological frontiers, legal practitioners emphasize user-friendly, compliance-aligned technologies. This disparity highlights the need for ongoing discourse between researchers and frontline professionals, ensuring that advancements in NLP serve the most pressing objectives within the evolving legal ecosystem of the DACH region.

Kurzfassung

Kontext: Die Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) gewinnt in der modernen Rechtsarbeit zunehmend an Bedeutung, was in hohem Maße auf rasche Fortschritte bei Large Language Models (LLMs) zurückzuführen ist. Obwohl Literatur und Praxis bereits vielfältige NLP-basierte Lösungen für juristische Anwendungsgebiete hervorgehoben haben, fehlt es weiterhin an einer systematischen *und* skalierbaren Trendanalyse. Infolgedessen bleibt unklar, wie sich Forschungsschwerpunkte entwickeln und inwieweit sie den bedeutendsten Use-Case-Kategorien in der juristischen Praxis—insbesondere im DACH-Raum—gerecht werden. Vorhandene Untersuchungen konzentrieren sich häufig auf einzelne Aufgaben oder Fallstudien und vernachlässigen dabei umfassende Erkenntnisse über Publikationsmuster auf Kategorieebene sowie die konkreten Anforderungen beruflicher Akteure.

Zielsetzung: Aufbauend auf früheren Arbeiten der SEBIS-Forschungsgruppe, die zentrale juristische NLP-Use-Cases identifiziert haben, liefert diese Arbeit eine detaillierte Untersuchung aktueller Trends und Praktikerprioritäten in diesen Anwendungsbereichen. Konkret soll (i) ein umfangreicher Korpus wissenschaftlicher Publikationen mithilfe einer rigorosen Pipeline kategorisiert, (ii) auf Basis der Resultate ermittelt werden, welche Muster sich im NLP-gestützten Rechtsumfeld abzeichnen, und (iii) durch halbstrukturierte Interviews mit Fachexperten aus dem DACH-Raum eine Validierung sowie Kontextualisierung der Ergebnisse erfolgen.

Vorgehensweise: Ein Large Language Model wurde für einen zweistufigen Klassifikationsprozess feinabgestimmt. Zunächst wird ermittelt, ob eine Publikation zumindest einen juristischen NLP-Use-Case behandelt, was eine F_1-Score von 97% in der Unterscheidung zwischen relevanten und nicht-relevanten Arbeiten ergibt. Für als relevant eingestufte Artikel identifiziert dasselbe Modell anschließend, welche von 31 vordefinierten Sub-Use-Cases zutreffen, was zu einer Gesamt-F_1-Score von 81,3% führt. Anschließend wurden 12 halbstrukturierte Interviews mit Rechtsanwälten und Unternehmensjuristen durchgeführt, um sieben übergeordnete Kategorien juristischer KI zu bewerten und Anforderungen für Sub-Use-Cases aus dem Arbeitsalltag zu ermitteln. Auf diese Weise werden die Ergebnisse aus der akademischen Forschung mit den praktischen Gegebenheiten verknüpft.

Ergebnisse: Von insgesamt 3.578 gescreenten Arbeiten wurden 988 als juristisch relevant eingestuft. Die Analyse dieser Publikationen zeigt, dass *Legal Research & Information Management* am häufigsten vorkommt (in 53,7% der Dokumente), gefolgt von *Information Processing & Extraction* mit 52,4%, was das anhaltende Interesse an komplexen Retrieval- und Entity-Extraktionsprozessen verdeutlicht. Die Interviews zeichnen jedoch ein ergänzendes Bild: Praktiker setzen *Document Generation and Assistance* an die Spitze ihrer Prioritäten

(durchschnittlicher Rang von 1,42) und betonen häufig Effizienzgewinne bei der Vertragsgestaltung. Zwar bewerteten sie *Legal Research and Information Management* ebenfalls hoch (durchschnittlicher Rang von 3,67), wiesen aber auf Unterschiede zwischen den in der Forschung hervorgehobenen Sub-Use-Cases (z.B. breit angelegte Forschungsautomatisierung) und ihren konkreten Anforderungen (z.B. e-Discovery, Gesetzgebungs-Tracking) hin.

Fazit: Durch die Verknüpfung eines automatisierten Literaturreviews—auf Basis großer Sprachmodelle—mit gezielten Branchenbefragungen verdeutlicht diese Arbeit den rasant wachsenden Forschungsstand im Bereich juristischer NLP-Anwendungen und macht die Herausforderungen bei der Überführung dieser Erkenntnisse in praxistaugliche Lösungen sichtbar. Die Ergebnisse zeigen, dass die wissenschaftlichen Publikationen ein breites methodisches Spektrum abdecken, während Rechtsanwender verstärkt Wert auf nutzerfreundliche, compliance-orientierte Technologien legen. Diese Diskrepanz unterstreicht den Bedarf an einem kontinuierlichen Dialog zwischen Forschern und beruflichen Anwendern, damit Fortschritte im NLP passgenau auf die dringlichsten Ziele im sich wandelnden Rechtsumfeld des DACH-Raums abgestimmt werden.

Contents

1. Introduction

This introduction begins by highlighting the significance and expanding role of Natural Language Processing (NLP) in modern industries, emphasizing its growing impact on legal workflows. It then articulates the central research questions guiding the present work, establishing the analytical framework adopted in subsequent chapters. Lastly, a brief overview of the thesis structure is provided.

1.1. Significance of NLP in the Legal Industry

NLP, a prominent subset of Artificial Intelligence (AI), has transformed multiple sectors by enabling efficient automation of complex language-driven tasks. Industries such as finance, healthcare, manufacturing, and customer relations have experienced significant benefits from NLP, ranging from improved accuracy and productivity to enhanced customer experiences [1]. The legal sector, inherently dependent on extensive textual analysis, is increasingly leveraging NLP capabilities for tasks including contract evaluation, automated legal research, predictive analytics, and document generation.

The integration of NLP within the legal domain presents unique opportunities and challenges, particularly with the advancement of generative models and large language models (LLMs). These technologies offer legal practitioners robust tools for automating routine processes and providing detailed analytical insights, thereby reshaping the landscape of legal practice and enhancing efficiency.

Recognizing the evolving role of NLP technologies, Chair of Software Engineering for Business Information Systems (*sebis*) at Technical University of Munich (TUM) developed the Legal AI Use Case Radar to map and systematically analyze NLP use cases in legal contexts [2]. This thesis advances previous efforts by introducing an innovative, automated literature classification method alongside structured expert analyses, specifically targeting the legal sector within the DACH region.

1.2. Research Questions

To address the interplay between academic research and real-world practitioner needs in legal NLP, the following research questions guide this thesis:

RQ1: *How can an automated, scalable pipeline effectively categorize academic literature into predefined legal AI use cases?*

RQ2: *Which legal AI use cases have received the most attention in academic research, and how have these patterns evolved?*

RQ3: *Which legal AI use cases do practitioners identify as most relevant to their professional practice, and what factors influence these perceptions?*

Collectively, these questions frame the methodological design and the subsequent analyses reported in this work. By focusing on both the technical feasibility of a scalable literature review pipeline and the alignment (or divergence) between research priorities and practitioner viewpoints, the thesis aspires to offer a multifaceted perspective on NLP's role in the contemporary legal sector.

1.3. Outline

This thesis is organized into seven chapters. Chapter 2 introduces foundational concepts required for understanding NLP's application to legal tasks. Chapter 3 examines existing scholarship on NLP technologies and Legal AI adoption trends. Next, Chapter 4 outlines the methodology for literature classification and expert interviews, and Chapter 5 presents the primary results from each methodological step. Chapter 6 then interprets these findings, situating them within broader academic and industry contexts. Finally, Chapter 7 concludes with a summary of key insights and offers prospects for future research directions.

2. Fundamentals

To ensure clarity and accessibility for readers with diverse backgrounds, this chapter defines core terminology and foundational concepts pertinent to NLP in the legal domain. We begin by presenting the basic principles and recent evolutions of Natural Language Processing, then highlight how these elements shape the use-case frameworks explored in later chapters.

2.1. Natural Language Processing: Concepts and Techniques

Natural Language Processing (NLP) is a subfield of Artificial Intelligence (AI) centered on enabling machines to interpret, analyze, and generate human language. Drawing upon computational linguistics, computer science, and machine learning, NLP automates tasks such as text classification, information extraction, and sentiment analysis [1]. By converting unstructured text (e.g., sentences, paragraphs) into computationally tractable formats, NLP systems can identify patterns and derive insights from large corpora of documents.

Advances in Generative AI and Large Language Models (LLMs). In recent years, NLP has progressed considerably due to advances in generative AI (GenAI) and large language models (LLMs) [3]. GenAI encompasses technologies designed to create new content (text, images, or even code) from existing data, while LLMs—exemplified by models like GPT-4—are powerful neural architectures trained on vast textual datasets. Such models excel in understanding linguistic context, which enables them to produce coherent paragraphs, summarize lengthy documents, and engage in structured dialogues. These features are particularly relevant to legal workflows, where precision and context are paramount.

Prompt Engineering. When interacting with LLMs, *prompt engineering* is often employed to guide model outputs [4]. This practice entails crafting text prompts that supply context, examples, or structured instructions to the model. Through iterative refinement of prompt phrasing and content, users can direct the model's focus, minimize ambiguity, and attain more accurate results. In complex domains like law, where terminology and formatting can vary significantly, well-designed prompts can substantially improve the model's clarity and consistency.

Fine-Tuning for Domain-Specific Tasks. Beyond prompt engineering, *fine-tuning* a pre-trained LLM involves retraining its parameters on more targeted datasets, adapting the model's broader linguistic knowledge to a specific domain [5]. In legal contexts, this might entail refining the model with case-law excerpts, regulatory texts, or annotated legal documents,

thereby enhancing its grasp of specialized vocabulary and stylistic conventions. Fine-tuned models often outperform generic ones on tasks such as classification, clause identification, or legal summarization, as they acquire a more nuanced understanding of relevant terms and document structures. Effective fine-tuning typically requires curating representative samples, maintaining robust validation protocols, and ensuring that any domain-specific biases—such as regional legal practices or specialized terminology—are adequately addressed.

2.2. Legal Tech

Legal Tech refers to the range of technological solutions and software platforms aimed at automating or enhancing various legal processes, such as contract drafting, document review, compliance checks, and legal research. Although Legal Tech tools can take many forms—from simple practice management systems to sophisticated artificial intelligence (AI) applications—they share a common objective of improving efficiency and accuracy in legal workflows [6].

Core Characteristics. Legal Tech generally exhibits three defining features:

1. **Domain-Specific Functionality:** Tools are purpose-built for legal tasks, addressing unique aspects such as statutes, regulations, and specialized legal terminology.
2. **Workflow Integration:** Many platforms are designed to fit into established law-firm or corporate legal practices, supporting operations like case management, contract lifecycle tracking, or risk assessment.
3. **Potential for Automation:** By leveraging natural language processing (NLP) and machine learning, modern Legal Tech can automate tasks that previously demanded extensive manual effort, such as bulk clause extraction or multi-document consistency checks.

Evolution of Legal Tech. Historically, early Legal Tech solutions (*Legal Tech 1.0*) primarily offered digital aids such as document indexing or keyword-based search. Over time, the field has adopted machine learning techniques to handle higher-order language analysis (*Legal Tech 2.0*), including advanced text classification and predictive analytics [7, 8]. In its most forward-looking conception (*Legal Tech 3.0*), AI-driven systems could automate entire workflows or facilitate online dispute resolution, representing a more fundamental restructuring of traditional legal service delivery.

Relevance to NLP. As legal practice revolves largely around text-heavy tasks, NLP is a central enabler of Legal Tech's development. Core NLP methods—such as entity recognition, clause identification, or summarization—enable automated systems to interpret and manipulate dense legal documents. The increasing adoption of large language models further broadens these capabilities, allowing, for instance, contract generation or more nuanced risk assessment [6].

Regulatory and Ethical Considerations. Legal Tech operates within ethical and regulatory frameworks specific to the legal profession, which emphasizes confidentiality, accuracy, and professional accountability. Technological applications often require careful oversight to ensure compliance with these standards, particularly where AI is entrusted with sensitive decisions. Consequently, human supervision typically remains integral to the deployment of Legal Tech, ensuring automated outputs adhere to ethical practice and protect client interests.

2.3. Legal AI Radar

The *Legal AI Use Case Radar* is an ongoing research effort launched in early 2023 under the direction of Prof. Dr. Florian Matthes at the Chair of Software Engineering for Business Information Systems (sebis), Technical University of Munich. This interdisciplinary team includes Research Associates Juraj Vladika, Stephen Meisenbacher, and Nektarios Machner, each focusing on different facets of NLP-based innovation in legal contexts. The Radar systematically tracks and classifies AI-driven applications in Germany's legal domain, offering legal professionals and researchers a structured view of software tools, adoption patterns, and methodological trends [9].

Two prior master's theses within this initiative—one by Martina Preis and another by Benedikt Thiess—laid much of the groundwork for the Radar's knowledge base. Preis's work introduced a core taxonomy of NLP use cases and examined the extent to which Ethical, Legal, and Social Aspects (ELSA) receive attention in research and practice, while Thiess refined those use cases by analyzing viewpoints from both "tool providers" and "tool appliers." [10, 11] Building on these foundations, the present thesis expands the Legal AI Use Case Radar through additional literature analyses, expert interviews, and updated categorizations of NLP solutions, contributing to the project's overarching goal of bridging academic insights with real-world legal needs.

3. Related Work

This chapter discusses prior research relevant to this thesis, structured into two sections. First, *Automated Classification Pipelines for Academic Literature* reviews previous scalable approaches to categorizing large sets of academic texts using machine learning. Next, *Adoption Trends and Practitioner Alignment in Legal AI* examines studies exploring how practitioners prioritize legal AI use cases.

3.1. Automated Classification Pipelines for Academic Literature

The task of classifying academic literature — often in a multi-label fashion where a paper may span multiple topics — has seen a clear evolution from manual curation to sophisticated AI-driven pipelines. Generally, labor-intensive manual reviews are considered one of the most popular options. For example, Vladika et al. [12] conducted a systematic manual review of legal AI use-cases and NLP technologies in the German legal domain, identifying seven categories of NLP methods and mapping them to numerous legal use-cases. While thorough, such manual approaches are time-consuming and may not scale well as the volume of publications grows.

To address scalability, researchers have developed automated classification pipelines using classical machine learning techniques. Ortiz and Segarra-Faggioni [13] demonstrated an early pipeline that automatically categorizes research papers using traditional algorithms (specifically a k-Nearest Neighbors classifier combined with Linear Discriminant Analysis). In their study on 596 academic articles, this approach achieved an accuracy of about 88%, showing the feasibility of replacing manual labeling with supervised learning. However, these traditional ML methods often required extensive feature engineering and struggled with capturing the nuanced context of text, especially for multi-label classification tasks where documents belong to multiple categories.

Later, the introduction of transformer-based language models brought a significant leap in text classification performance. Devlin et al.'s BERT model [14] was a milestone that introduced deep bi-directional transformers pre-trained on large text corpora [15]. Building on transformers, recent works leverage large language models (LLMs) to further advance automated classification. A comprehensive survey by Fields et al. finds that LLMs with hundreds of millions or more parameters consistently outperform earlier ML and even base transformer models on diverse text classification tasks [16]. In addition to leveraging these large models, prompt engineering has emerged as an essential technique for guiding LLMs in classification tasks without modifying their core parameters, thereby reducing the need for resource-intensive re-training. By carefully crafting prompts, model outputs

can be steered toward more precise and context-aware classifications, showing significant improvements in performance [17]. Beyond prompt engineering, Rostam and Kertész [18] report that fine-tuning a domain-adapted model can yield even higher accuracy, particularly when classifying research abstracts and keywords. This two-pronged approach—combining prompt engineering with subsequent fine-tuning—demonstrates the evolving state of the art in automating and improving multi-label classification pipelines for academic literature.

3.2. Adoption Trends and Practitioner Alignment in Legal AI

Historically, the legal sector approached AI with caution, devoting years to theoretical debate before implementing any practical solutions [19]. Over the past decade, however, firms have increasingly shifted from concept to practice, deploying AI-driven tools to boost productivity and client service. In recent years, this trend has accelerated markedly: one survey found that UK lawyers' use of AI surged from 11% in 2023 to 41% in 2024 [20], indicating a rapidly growing appetite for legal AI. This surge is widely attributed to technological advances such as generative AI, which have made lawyers more receptive to these innovations than they were just five years ago [21].

Mounting pressure to handle escalating case data and deliver expedient, cost-effective outcomes has led many lawyers to adopt AI for streamlining routine tasks. Surveys confirm that quicker service, improved client results, and a competitive edge rank among the principal drivers for embracing legal AI. Early AI solutions thus focused on high-volume processes like document drafting and legal research, where automation can substantially augment human effort [22]. Nevertheless, the profession's risk aversion and high accuracy requirements have tempered the pace of integration, prompting thorough vetting of each new system's reliability and ethical compliance. Consequently, most implementations preserve human oversight, with AI systems designed to enhance rather than replace legal judgment.

Innovative startups have further stimulated AI uptake by introducing tools that automate repetitive tasks and address inefficiencies in legal workflows, compelling more traditional firms to adapt as these tech-enabled services offer faster, more affordable solutions. Firms that embrace such innovation report notable efficiency gains and strengthened competitive positions. Conversely, those slow to modernize risk losing clientele and becoming less relevant in an increasingly technology-driven marketplace. Taken together, these developments underscore that practitioner demand and startup-led innovation jointly steer AI integration, ensuring emerging solutions remain tightly aligned with the practical needs of legal professionals.

4. Methodology

This chapter outlines the methodological approach employed to systematically address the three core research questions guiding this thesis.

RQ1: How can an automated, scalable pipeline effectively categorize academic literature into predefined legal AI use cases?

RQ2: Which legal AI use cases have received the most attention in academic research, and how have these patterns evolved?

RQ3: Which legal AI use cases do practitioners identify as most relevant to their professional practice, and what factors influence these perceptions?

In the subsequent sections of this chapter, we first define a core taxonomy of Legal NLP use-cases and associated NLP techniques, grounded in both prior literature and empirical insights from legal practitioners. We then present a systematic literature review methodology, encompassing source selection, search query construction and execution, as well as the techniques used to classify and refine our dataset through prompt engineering, manual labeling, and model fine-tuning. Finally, we discuss how structured interviews are employed to gather qualitative perspectives from practitioners, ensuring that our findings are complemented by real-world expert views on the applicability and relevance of identified Legal NLP solutions.

4.1. Defining Use-Cases and NLP Techniques

We adopt the taxonomy proposed by the Sebis research team [12], which was developed through a systematic literature review and extensive interviews with legal practitioners. This framework organizes Legal NLP tasks into seven main *use-case categories*, spanning from Compliance and Risk Management to Legal Research and Information Management, and seven main *NLP technique categories*, ranging from Syntactic Analysis to Text Classification. In later sections, we also examine whether any additional or emerging use-cases and NLP techniques appear in the examined corpus.

Table 4.1.: Legal NLP use-cases (31 subcategories), grouped by higher-level categories

Subcategory	Definition
Compliance and Risk Management	
Automation of Auditing	Automatically flag non-compliant language or discrepancies across legal documents.
GDPR Compliance	Analyze documents for personal data handling and highlight potential GDPR violations.
Risk Assessment	Identify and categorize risk factors by extracting key clauses and obligations from legal texts.
Document Analysis and Management	
File Difference Tracking	Automatically track and explain changes across document versions by generating annotations.
Document Classification	Automatically categorize legal documents by type or domain using classification approaches.
Content Lifecycle Management	Guide legal documents through creation, review, updates, and archiving.
Error Detection	Flag missing critical clauses, contradictory obligations, or non-standard formatting.
Document Generation and Assistance	
Contract Generation	Draft contract templates by leveraging precedent-based text suggestions and clause libraries.
Legal Document Enrichment	Insert references or annotations into existing legal texts using automated insights.
Summarization	Generate concise summaries of legal documents automatically.
Deadline Management	Parse schedules or filings to flag upcoming dates and notify stakeholders.
E-Mail	Suggest relevant legal language or disclaimers in email communications using context analysis.
Class Action Lawsuits	Analyze large volumes of related claims and unify textual evidence for class actions.
Information Processing and Extraction	
Anonymization / Text Scrubbing	Automatically remove or mask sensitive or personal information in legal documents.
Information Extraction	Identify and extract predefined factual entities from legal documents.
Document Retrieval	Use automated search methods to retrieve relevant case law or statutes from large documents.
Transcription	Convert non-textual data (scanned PDFs, audio) into searchable text via OCR or ASR.
Legal Decision Making and Dispute Resolution	
Judge: Decision Making	Analyze case facts and precedents to inform judicial rulings (sentencing, injunctions).
Legal Reasoning	Apply and justify legal rules from statutes and precedents to enable automated reasoning.
Strategy Recommendations	Offer data-driven guidance on litigation or negotiation tactics by analyzing case law.
Dispute Resolution Mechanism	Identify disagreements, provide legal insights, and mediate discussions to streamline resolution.
Legal Information Retrieval and Support	
Chatbot	Automate legal consultations via interactive dialogue systems that explicitly process legal text.
Question Answering	Provide precise legal answers from statutes, regulations, or case libraries.
Ranking of Lawyers	Analyze textual data (reviews, case outcomes) to rank lawyers by expertise or success rates.
Credibility of Witnesses	Assess witness statements for inconsistencies or sentiment cues.
Translation	Provide machine translation for cross-lingual legal documents while preserving terminology.
Legal Research and Information Management	
Changes in Law	Automatically track and highlight legislative updates across large corpora.
Court Rulings Indexing	Index and classify court rulings for efficient retrieval and reference.
Law Systems Divergence	Compare texts from different jurisdictions to identify conflicting or varying provisions.
Research Tool / Research Automation	Automate analysis of large corpora to expedite document review and trial preparation.
e-Discovery	Identify, collect, review, and produce electronically stored information (ESI) as evidence.

Table 4.2.: NLP technique categories (17 subcategories) grouped by higher-level categories

Subcategory	Definition
Syntactic Analysis	
Dependency Parsing	Identifies hierarchical relationships among words (head-modifier links).
Tokenization	Splits text into smaller units (tokens) for downstream processing.
Lexical Normalization	Corrects spelling and standardizes text.
Part-of-Speech Tagging	Assigns grammatical categories (noun, verb, etc.) to each token.
Text Extraction	
Named Entity Recognition	Detects references to real-world entities (people, locations, organizations).
Keyword Extraction	Identifies critical terms or phrases that capture main topics in a document.
Document Analysis	
Entity Linking	Resolves entity mentions in text to canonical knowledge-base entries.
Document Similarity Analysis	Gauges similarity between documents based on semantic or lexical features.
Text Representation	
Word Embedding	Learns vector representations of words from their contexts.
Language Modeling	Estimates probability distributions over word sequences to capture context.
Text Generation	
Text Summarization	Produces concise versions of longer texts while retaining key information.
Machine Translation	Translates text between languages.
Conversational NLP	
Chatbot Development	Builds dialogue systems capable of interactive question-answering.
Question Answering	Retrieves or synthesizes answers to user queries from knowledge sources.
Text Classification	
Topic Modeling	Discovers latent topics in a corpus based on word distributions.
Concept Models	Maps text segments to predefined conceptual or ontological constructs.
Text Classification	Assigns labels or categories (e.g., sentiment, domain) to text.

4.2. Systematic Literature Review

A systematic literature review (SLR) aims to identify, collect, and evaluate the available body of research in response to well-defined research questions [23]. In accordance with these guidelines, our primary objective is to comprehensively survey existing academic work on legal NLP in order to provide a robust foundation for subsequent analyses. Following the established methodology, we formulated the research questions and used them to define a structured search protocol.

4.2.1. Selection of Sources

The first step in our protocol involved identifying the most relevant academic venues and databases. We chose sources that (1) are highly regarded for their wide coverage of peer-reviewed content, (2) have a strong focus on computer science and legal technology, and (3) present diverse perspectives on emerging trends in legal NLP. In selecting these sources, we also aimed to maximize the coverage of publications spanning different disciplinary

boundaries.

- **ACM Digital Library** – Offers comprehensive coverage of computer science and interdisciplinary topics.
- **Scopus** – A large, multidisciplinary database providing extensive indexing of peer-reviewed research.
- **IEEE Xplore** – Well-established platform focusing on computer science, engineering, and related fields.
- **ICAIL** – The International Conference on Artificial Intelligence and Law, chosen specifically for its primary emphasis on Legal AI research.
- **ACL Anthology** – A key venue for natural language processing and computational linguistics, including specialized workshops on legal NLP.

Table 4.3.: Overview of selected databases and conferences

Source	Link
ACM Digital Library	`https://dl.acm.org/`
Scopus	`https://www.scopus.com/`
IEEE Xplore	`https://ieeexplore.ieee.org/`
ICAIL	`https://icailconference.org/`
ACL Anthology	`https://aclanthology.org/`

4.2.2. Constructing the Search Query

To collect a sufficiently broad pool of academic works for our automated review pipeline, we developed a comprehensive search query designed to retrieve papers at the intersection of NLP and legal domains. By allowing for a wide set of potential keywords, we acknowledge that the resulting set of studies may contain a significant proportion of non-relevant items. However, this expansive approach is justified by the subsequent automated filtering pipeline, which can efficiently eliminate irrelevant publications.

Following guidelines set out by Zhang, Babar, and Tell [24], our search query was applied to the title, abstract, and keyword sections of articles in the selected databases (excluding ICAIL). Since the International Conference on Artificial Intelligence and Law (ICAIL) solely features legal AI research, we considered its proceedings inherently relevant and opted not to run an automated keyword-based extraction on them. The final query, presented below, reflects this strategy:

```
(
  ( "NLP" OR "Natural Language Processing" OR "Computational Linguistics"
```

```
    OR "Language Models" OR "LLM" OR "Machine Learning"
    OR "Artificial Intelligence" OR "Deep Learning" )
  AND
  ( "Law" OR "Legal" OR "Jurisprudence" OR "Judicial"
    OR "Attorney" OR "Contracts" OR "Case Law"
    OR "Regulations" OR "Statutes" )
)
OR
(
  ( "LegalTech" OR "Legal Tech" OR "Legal Technology" )
  AND
  ( "Use Case" OR "Application" OR "Case Study" OR "Implementation" )
)
OR
(
  ( "Contract Analysis" OR "Legal Document Processing"
    OR "Legal Information Retrieval" OR "Legal Question Answering" )
)
```

By using this high-recall query, we ensure maximal coverage of potentially relevant studies, thus laying the groundwork for a subsequent, more refined selection process within our automated pipeline.

4.2.3. Execution of the Search Query

After defining the search strategy, we executed the query across all selected databases and venues. Each platform handled exports in distinct formats, thereby necessitating a standardization step. In the case of Scopus and IEEE Xplore, we carried out the search directly on their web interfaces, obtaining the resulting bibliographic data in comma-separated value (CSV) files. By contrast, the ACM Digital Library and ICAIL provided references in `.bib` files, which we parsed programmatically using a custom Python script to homogenize and merge them with the CSV-based records.

ACL Anthology presented a unique challenge: its interface does not permit complex queries. Consequently, we downloaded the complete `.bib` file of the ACL Anthology and programmatically filtered it with the previously defined search terms. Table 4.4 shows the number of initially retrieved papers per source.

All records were integrated into a single dataset, with the following standardized metadata fields: *Title, Authors, Year, Source Title, Volume, Issue, Article No., Page Start, Page End, DOI, Abstract, Keywords, Document Type, Publisher, Source File.* Because some works appeared in multiple venues, duplicate entries were removed based on matching titles, DOIs, and additional heuristic checks. This process yielded a final, deduplicated corpus of 3,578 distinct papers.

With the assembled dataset in place, we proceeded to the subsequent analysis phases,

Table 4.4.: Number of papers extracted per source

Source	Number of Papers
Scopus	2533
ICAIL	632
ACL	407
ACM	119
IEEE	59

namely creating a validation set for manual labeling and developing a comprehensive automated filtering pipeline. These steps ensure that only the most relevant studies (relative to the research questions) are carried forward in our investigation.

Inclusion and Exclusion Criteria

Having identified relevant sources and executing our search query, we next applied a set of inclusion and exclusion criteria to filter out papers that did not meet the scope of this study. These criteria were programmatically enforced through **Prompt A** (see Appendix A.1.1), ensuring that irrelevant or duplicate publications were systematically removed. Specifically, a publication was excluded if it satisfied at least one of the following conditions:

- **No NLP Focus:** The paper does not address Natural Language Processing.
- **No Legal Context:** The paper does not relate to legal applications or legal data.
- **Non-English/German:** The paper is written in a language other than English or German.
- **Non-Research Document Types:** The paper is a book, presentation, conference note, or a survey paper (not an original research article).
- **Outside Publication Range:** The paper was published before January 1980 or after January 2025.
- **No Full-Text Access:** The full text of the paper could not be accessed with the rights granted by the Technical University of Munich.
- **Poor Quality or Invalid Content:** The publication contains severe grammatical or vocabulary deficiencies, making it unsuitable for detailed analysis.
- **General Legal AI or AI Ethics Only:** The paper discusses Legal AI in general or focuses solely on ethical/policy issues without any text-based or NLP methods.
- **Duplicate Entry:** The publication is already included in the selection from another source or was identified in a prior step.

These criteria ensured that only *legal NLP research* articles (explicitly describing text-based or language-based methods) were retained for subsequent use-case and nlp techniques classification.

4.2.4. Prompt Engineering

This section describes the methodology for employing large language models (LLMs) to address a multi-label text classification task encompassing **31 use-case subcategories** and **17 NLP technique subcategories**. We rely on the taxonomy presented in Section 4.1, while the specific prompts are detailed in Appendix A.1.1 (`Prompt A`) and Appendix A.1.2 (`Prompt B`).

Designing the Prompts

Both prompts underwent a shared, iterative refinement process grounded in the taxonomy of Section 4.1 and the inclusion/exclusion criteria of Section 4.2.3. After each round of testing on a subset of the corpus, misclassifications were analyzed and the prompts were adjusted to limit false positives and false negatives. A confidence threshold of 0.8 was established to ensure that classifications rely on explicit textual evidence, and a strict "Zero Implication" rule was introduced to prohibit speculative inferences.

- **Prompt A (Boolean + Use-Case Classification):** This prompt, detailed in Appendix A.1.1, first determines whether a paper discusses legal NLP according to the established taxonomy, referencing the inclusion and exclusion criteria from Section 4.2.3 to filter out purely conceptual or policy-focused work. If the text satisfies these requirements, the system assigns one or more predefined subcategories that are explicitly mentioned, allowing for multi-label classification. Early iterations were prone to misclassifying broader legal AI papers, prompting the integration of a rule-based mechanism that requires direct references to text-processing methods.
- **Prompt B (NLP Techniques):** Described in Appendix A.1.2, this prompt identifies which of the 17 predefined NLP methods are used in each relevant paper. Like Prompt A, it demands unambiguous textual evidence before assigning a label; general AI workflows lacking explicit language-processing details are excluded. Early versions tended to overcount techniques when they were merely mentioned without detailed descriptions. The final iteration stipulates that papers must clearly indicate the deployment or discussion of a technique, thus improving precision in identifying the methods adopted.

Zero-Shot Prompting

Our initial implementation followed a *zero-shot* paradigm, in which the LLM is asked to perform classification without any explicitly labeled examples. This approach is guided by studies demonstrating the generalization capabilities of large language models under minimal supervision [25]. We iteratively refined this prompt setup by reviewing predictions on a small, manually labeled dataset of 50 papers, adjusting both prompt wording and

subcategory definitions to address ambiguities or inconsistencies. To ensure consistent and reproducible outputs, we set the model's sampling `temperature` to `0.0`, thereby removing stochastic variation in the classification results.

Few-Shot Prompting and Reasoning Strategies

After observing some performance gains from prompt refinement, we experimented with a *few-shot* prompting approach [26]. In this method, a limited number of examples with known labels are appended to the prompt, providing the model with more explicit context. However, due to the large number of target labels (31 use-case subcategories and 17 technique categories), few-shot prompting yielded only marginal improvements.

We next explored advanced reasoning-based strategies, including classic Chain-of-Thought (CoT) [27] and Automatic Chain-of-Thought (Auto-CoT) [28], in an effort to enhance model interpretability and classification consistency. Although these techniques can bolster systematic reasoning in some cases, their effect on our multi-label classification tasks proved limited. Consequently, we concluded that prompt engineering alone was insufficient for achieving the desired performance.

In order to further improve results, we opted to *manually label* a selected portion of the dataset to form a higher-quality ground-truth reference, which is discussed in the next section. Building on these manual labels, we subsequently explore a fine-tuning approach to optimize classification accuracy and scalability.

4.2.5. Manual Labeling

Reaching a performance plateau with purely prompt-based classification, we established a *manually labeled dataset* to support a fine-tuning approach. We began by randomly selecting **200** papers from our corpus. Following an initial labeling process, it became evident that certain use-case subcategories and NLP techniques were underrepresented in these papers. Consequently, we added **33** more papers through targeted keyword searches, resulting in **233** manually labeled papers total: **168** of which explicitly address Legal NLP (text-based methods in a legal context) and **65** that do not. Figure 4.1 provides a pie chart illustrating this distribution.

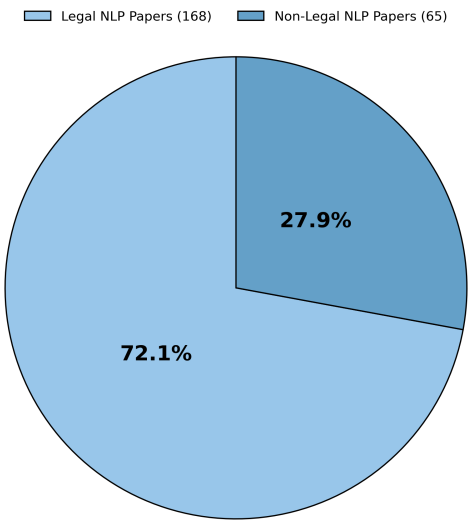


Figure 4.1.: Distribution of the 233 manually labeled papers.

Handling Rare Subcategories

During our initial review of 200 randomly selected papers, we noted the absence or minimal presence of specific *use-case* categories:

- *e-Discovery*
- *Risk Assessment*
- *Ranking of Lawyers*
- *Law Systems Divergence*
- *Database for Court Decisions*
- *E-Mail*
- *Deadline Management*

In addition, certain *NLP techniques* did not appear:

- *Topic Modeling*
- *Part of Speech Tagging*
- *Lexical Normalization*

To ensure adequate coverage, we performed targeted keyword searches aiming to identify papers referencing these missing categories or techniques. Although most were successfully located and manually labeled, we found no relevant papers for *E-Mail* or *Deadline Management*.

Novel Use-Case Identification

To further account for any categories not captured by our known taxonomy, we *designed a separate prompt* dedicated to detecting additional legal AI use cases or NLP techniques. This prompt scanned all 200 initially selected papers for textual references that might imply a *novel or previously undocumented subcategory*. We then manually reviewed each *candidate* paper flagged by the script. Through this process, we confirmed three new subcategories—*Legal Corpus Curation*, *Smart Contract Analysis*, and *Legal Language Interpretation*—within existing main categories, bringing the total number of distinct use-case subcategories to **34**, while finding *no new NLP techniques* (Table 4.5).

Table 4.5.: Newly identified use-case subcategories

Main Category	**Subcategory**	**Definition**
Legal Research and Information Management	Legal Corpus Curation	Build and annotate collections of legal texts for research and knowledge management.
Document Analysis and Management	Smart Contract Analysis	Examine blockchain-based contracts to identify obligations, risks, or compliance gaps.
Document Generation and Assistance	Legal Language Interpretation	Clarify complex provisions, simplify clauses, or explain terminology in legal documents.

Descriptive Statistics of the Final Labeled Dataset

Tables 4.6 and 4.7 present the frequency of each use-case and NLP technique, respectively, for the 168 papers that were confirmed to address Legal NLP.

Table 4.6.: Use-Case subcategory distribution, sorted by total frequency of main categories

Main Category	Subcategory	Count	%
Information Processing and Extraction			
	Information Extraction	68	40.48
	Document Retrieval	23	13.69
	Transcription	9	5.36
	Anonymization / Text Scrubbing	5	2.98
Legal Research and Information Management			
	Research Tool / Research Automation	61	36.31
	Legal Corpus Curation	29	17.26
	Changes in Law	3	1.79
	e-Discovery	3	1.79
	Law Systems Divergence	2	1.19
	Database for Court Decisions	1	0.60
Document Analysis and Management			
	Document Classification	35	20.83
	Error Detection	16	9.52
	Smart Contract Analysis	6	3.57
	File Difference Tracking	6	3.57
	Content Lifecycle Management	5	2.98
Legal Decision Making and Dispute Resolution			
	Judge: Decision Making	22	13.10
	Legal Reasoning	13	7.74
	Dispute Resolution Mechanism	6	3.57
	Strategy Recommendations	3	1.79
Legal Information Retrieval and Support			
	Question Answering	18	10.71
	Chatbot	9	5.36
	Translation	7	4.17
	Credibility of Witnesses	3	1.79
	Ranking of Lawyers	2	1.19
Document Generation and Assistance			
	Legal Language Interpretation	15	8.93
	Summarization	10	5.95
	Contract Generation	4	2.38
	Legal Document Enrichment	4	2.38
	Class Action Lawsuits	3	1.79
	E-Mail	0	0.00
	Deadline Management	0	0.00
Compliance and Risk Management			
	Automation of Auditing	6	3.57
	GDPR Compliance	5	2.98
	Risk Assessment	3	1.79

Table 4.7.: NLP Technique distribution ,sorted by total frequency of main techniques

Main Technique	Subtechnique	Count	%
Text Representation			
	Language Modeling	78	33.48
	Word Embedding	28	12.02
Text Classification			
	Text Classification	70	30.04
	Concept Models	9	3.86
	Topic Modeling	3	1.29
Text Extraction			
	Named Entity Recognition	36	15.45
	Keyword Extraction	12	5.15
Text Generation			
	Text Summarization	35	15.02
	Machine Translation	11	4.72
Document Analysis			
	Document Similarity Analysis	31	13.30
	Entity Linking	11	4.72
Conversational NLP			
	Question Answering	19	8.15
	Chatbot Development	4	1.72
Syntactic Analysis			
	Dependency Parsing	9	3.86
	Tokenization	4	1.72
	Part of Speech Tagging	2	0.86
	Lexical Normalization	1	0.43

By combining random sampling with targeted searches, we achieved a dataset that covers both frequently encountered and rare use-cases and techniques. This labeled corpus provides the basis for the subsequent **fine-tuning** process aimed at boosting classification performance.

4.2.6. Fine-Tuning

Having established a manually labeled dataset of 233 papers, we proceeded to fine-tune our models to improve classification accuracy for Legal NLP use-case identification and NLP technique categorization. Given the primary importance of correctly detecting use-case subcategories, we performed *stratified sampling based on use-case labels* to ensure sufficient representation of each subcategory in both training and test splits. Consequently, the training set comprised 167 papers, while the remaining 76 papers formed the test set. Figure 4.2 presents a stacked bar chart illustrating the distribution of Legal NLP versus non-legal papers across the two sets.

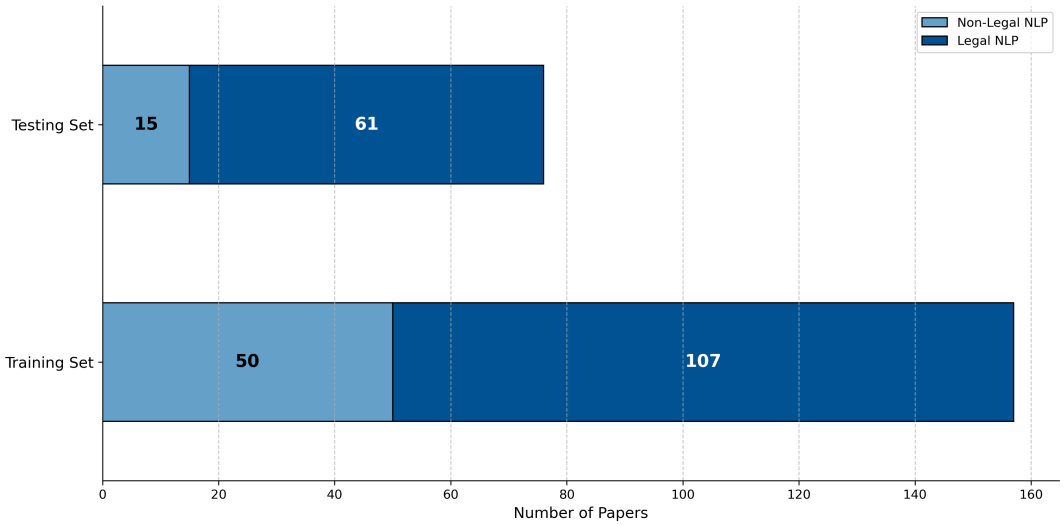


Figure 4.2.: Distribution of legal and non-legal papers in training and test sets

Reflecting our methodological priorities, we developed two independent fine-tuning pipelines:

1. **Use-Case Relevance & Subcategory Classification:** Determines whether a paper addresses Legal NLP and, if so, assigns one or more of the predefined subcategories.

2. **NLP Technique Classification:** Applied exclusively to papers already classified as Legal NLP, identifying relevant techniques discussed or employed.

We employed the *GPT-4o* model in both pipelines, leveraging its capacity for coherent text understanding and multi-label classification. Fine-tuning was performed using standard cross-entropy objectives for multi-label tasks, where each subcategory or technique dimension was treated as an independent binary classification. During training, we systematically monitored loss and performance on a validation subset to mitigate overfitting and ensure stable convergence. This process aimed to further boost classification precision beyond what prompt-based methods alone could achieve.

Performance Metrics and Evaluation

We evaluate model performance at both the **subcategory** and **category** levels to capture fine-grained distinctions and broader domain correctness. In all cases, our evaluation adopts a *multi-label* perspective, recognizing that each paper may belong to multiple subcategories.

Subcategory-Level Accuracy. We consider each (paper, subcategory) pair as a separate binary decision and measure the fraction that the model classifies correctly. This approach directly reflects the model's ability to distinguish among specific, often overlapping subcategories (e.g., *Smart Contract Analysis* and *Error Detection*).

Category-Level Accuracy. Subcategories are grouped under seven main categories (e.g., *Compliance and Risk Management*). We regard a category as correctly predicted for a paper if at least one of its constituent subcategories is correctly identified. This metric evaluates how reliably the model captures broader task domains, even if fine-grained subcategory boundaries are occasionally misclassified.

Precision, Recall, and F_1 Score. To account for imbalances in label frequencies, we calculate:

- **Precision (P)**: Proportion of predicted labels that match ground truth.
- **Recall (R)**: Proportion of ground-truth labels correctly retrieved by the model.
- **F_1 Score**: Harmonic mean of precision and recall, defined as

$$F_1 = 2 \times \frac{P \times R}{P + R}.$$

These metrics are computed on a *multi-label* basis, and we apply **macro-averaging** so that each label—whether frequent or rare—contributes equally to the final scores.

Interpretation. Subcategory-level accuracy and the F_1 score provide insight into the model's granularity in distinguishing similar tasks, while category-level accuracy reveals how robustly the model detects higher-level domains. By examining multiple metrics, we gain a balanced view of classification effectiveness across varying degrees of detail.

4.2.7. Trend Analysis

RQ2 asks: *How have academic trends concerning NLP use cases within the legal domain shifted over time, and which emerging areas have gained increased research interest?* To address this, we analyze publication counts for the 34 use-case subcategories and 17 NLP technique subcategories identified in our taxonomy. Rather than plotting all subcategories in a single chart, we employ two measures—*Compound Annual Growth Rate (CAGR)* and correlation-based *R-values*—alongside corresponding *p-values* to determine statistically significant trends.

Compound Annual Growth Rate (CAGR). For a subcategory with initial publication count N_0 and final count N_t over t years,

$$\text{CAGR} = \left(\frac{N_t}{N_0}\right)^{\frac{1}{t}} - 1.$$

A positive CAGR indicates growth over time; higher values suggest rapidly expanding research interest.

Correlation-Based R-values and p-values. We treat each publication year as x and the corresponding count of papers as y. The Pearson correlation coefficient r gauges linear association:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}.$$

To assess statistical significance, we calculate a **p-value** based on Student's t-distribution:

$$t = r\sqrt{\frac{n-2}{1-r^2}},$$

where n is the number of years under study. A p-value below 0.05 indicates that the observed trend is unlikely due to random fluctuations.

Interpretation.

- **High CAGR, High |r| (p < 0.05)**: Strong, consistent growth or decline—subcategories here are reliably trending upward or downward.
- **High CAGR, Low |r|**: Growth may be concentrated in certain years rather than forming a steady pattern.
- **p $\geq$ 0.05**: No statistically significant trend, even if CAGR or r-values appear non-zero.

By examining these metrics jointly, we isolate subcategories showing meaningful, sustained trajectories from those subject to short-lived spikes. In the results, we highlight which tasks and techniques exhibit statistically significant growth or decline, guiding our understanding of key research directions in Legal NLP.

4.3. Semi-Structured Interviews

To investigate **RQ3**—*Which legal AI use cases do practitioners identify as most relevant to their professional practice, and what factors influence these perceptions?*—we selected semi-structured interviews (SSIs) as our primary data collection method. This choice provides a balance of structure and flexibility, allowing participants to comment on pre-defined legal NLP use-case categories and to explain why certain subcategories motivate their ranking preferences [29].

Rather than reinventing use-case categories established in prior research, we incorporated an *interactive legal NLP use-case ranking exercise* alongside a concise set of background questions (e.g., role, tenure, current use of NLP) to capture both immediate prioritizations and the underlying reasons for them. This interview design—coupled with targeted follow-up questions—helps illuminate which specific subcategories (e.g., contract summarization vs. contract drafting) drive participants' decisions and thereby shapes our understanding of practitioners' needs within the legal tech landscape.

4.3.1. Methodology Design

Following the framework proposed by Kallio, Pietilä, Johnson, and Kangasniemi [30], we developed our semi-structured interview protocol in five iterative phases. Below, we detail how each phase informed the interview design and how the final guide was administered.

Construction of Interview Guide

Our interview guide consists of three introductory questions, followed by an online ranking exercise hosted on PaperForm[1] and related follow-up prompts. The following five phases summarize its evolution:

Phase 1: Identifying Prerequisites for Using Semi-Structured Interviews We first examined whether SSIs would allow us to capture nuanced feedback on an existing set of legal NLP use-case categories. Since these categories are comprehensive, participants might have *specific* reasons for ranking a given category higher or lower. Semi-structured interviews offer the flexibility to delve into such reasons, ensuring that rich qualitative data can be collected [31]. By asking participants to highlight relevant subcategories (e.g., Contract Summarization) within broader categories (e.g., Legal Drafting & Litigation Support), we could more precisely identify the functionalities driving their choices.

Phase 2: Retrieving and Using Previous Knowledge Next, we built on prior interviews and our research group's systematic literature review (SLR) of legal NLP applications. This prior work had already yielded seven broad use-case categories. However, our ongoing studies identified additional emerging applications, such as *Legal Corpus Curation*, *Smart Contract Analysis*, and *Legal Language Interpretation*, which were incorporated into the final definitions (see Table 4.5). By updating our categories in light of these novel subcategories, we ensured that participants' rankings would reflect the most current developments in the field.

Phase 3: Formulating the Preliminary Semi-Structured Interview Guide Using insights from Phases 1 and 2, we created a draft interview protocol. It began with three short, open-ended questions about the participant's (i) role, (ii) tenure in the company, and (iii) current use of NLP tools in legal practice. Immediately thereafter, we introduced the ranking exercise, in which participants were instructed to:

1. Review concise definitions of each of the seven NLP use-case categories.
2. Rank them in order of perceived importance or relevance to their work.
3. Provide verbal comments explaining which specific subcategories informed their rankings.

[1] We use PaperForm for its interactive features and direct embedding of category definitions for reference. The form is accessible at `https://edhr3ivu.paperform.co/`.

This sequence balanced efficiency with depth. Rather than administering numerous background questions, we focused on a core set of three that revealed contextual details, leaving most of the session available for the interactive ranking and subsequent discussion.

Phase 4: Pilot Testing the Guide We next sought feedback from our supervisory team (doctoral researchers), who reviewed the PaperForm interface and the interview flow. They did not complete a full trial interview; rather, they validated the layout, navigational elements, and clarity of the use-case definitions in the form. Their recommendations led to two key refinements: (1) embedding each category's definition directly within the online ranking form, so participants could easily refer back to it; and (2) simplifying the subcategory names for clarity. Additionally, after conducting two initial interviews, we realized that our preliminary design of six introductory questions exceeded our 30-minute limit. We therefore reduced the number of background questions to three (presented in Phase 3) and proceeded with the ranking exercise and follow-up discussion immediately afterward.

Phase 5: Presenting the Complete Semi-Structured Interview Guide In the final design, each interview begins with a brief introduction (allowing 5–7 minutes for the participant to share their role, tenure, and any current NLP usage). We then direct participants to the PaperForm link, allotting approximately 20 minutes for them to rank the seven use-case categories and supply short comments for each. In the concluding 3–5 minutes, we follow up with additional probes if time permits, focusing on clarifications or interesting points raised by their rankings.

This structured-yet-flexible approach ensures that we capture immediate, comparable data on use-case priorities while still affording participants the opportunity to elaborate on specific functionalities driving their decisions. As a result, our study design combines both quantitative ranking and qualitative insights, offering a robust view of how practitioners perceive and prioritize legal NLP use cases in their professional environments.

Interview Analysis

In line with the mixed-methods guidelines advocated by Venkatesh, Brown, and Bala [32], our analysis incorporated both quantitative (rank-order) and qualitative (interview commentary) data. This dual approach enriches our understanding of practitioners' perceived priorities for legal NLP use cases and the underlying reasons that shape those choices.

Data Collection Procedures All interviews were conducted remotely via Zoom and audio-recorded with participants' consent. Recordings were subsequently transcribed using the Otter.ai service, after which we performed a manual review to correct transcription errors and ensure accurate attribution of statements.

Quantitative Analysis of Ranking Data We collected rank-order inputs from 12 participants, each of whom prioritized the seven legal NLP use cases according to their perceived

importance. Given the relatively small sample size, we relied on straightforward descriptive statistics (e.g., mean ranks, frequency of top ranks) to summarize and visualize these results. While more advanced statistical or model-based approaches exist for analyzing rank-order data [33], the present study's exploratory nature and limited number of participants made classic descriptive methods sufficient. However, if this ranking exercise is repeated continuously (e.g., across multiple years or with larger cohorts), more sophisticated analyses such as consensus measures, non-parametric tests, or specialized rank aggregation models [33] would become increasingly relevant for capturing longitudinal or comparative insights.

Qualitative Coding of Subcategory Insights During the interviews, participants often elaborated on specific subcategories or examples motivating their rankings. To capture these qualitative nuances, we selectively applied the thematic analysis principles introduced by Braun and Clarke [34]. Rather than conducting an extensive multi-phase coding process, we grouped comments and examples under each of the seven broad use-case headings and noted salient subtopics (e.g., "contract summarization," "automated review of dispute provisions"). This focused approach helped clarify which functionality features drove higher or lower rankings without requiring a full-scale thematic breakdown.

Integration of Findings Following recommendations by Venkatesh et al. [32], we integrated the descriptive statistics from the ranking exercise with participants' qualitative commentary to develop a more holistic picture of their decision-making. Numerical results provided a clear snapshot of favored use cases, while the associated subcategory discussions illuminated the specific functionalities participants found most compelling or less relevant. By combining quantitative rank-order data with qualitative insights, we offer a robust account of *which* legal NLP use cases professionals prioritize and *why* these particular categories resonate most strongly in their practice.

4.4. Interview Makeup

4.4.1. Identifying Participants

Our initial aim was to recruit legal professionals located in the DACH region (Germany, Austria, Switzerland). However, we broadened our outreach to include individuals working in legal roles at organizations that maintain an operational presence in the DACH region, regardless of the participant's physical location. This expansion ensured coverage of a diverse array of institutional contexts and practice areas.

Channels for Recruiting Interview Participants

To contact prospective interviewees, we employed three main strategies: (1) direct outreach via LinkedIn (using a free trial of LinkedIn Sales Navigator), (2) personal referrals from colleagues and professional networks, and (3) re-engagement of individuals who had previously participated in our research group's studies. In total, we contacted 76 potential participants

across these channels; 12 individuals ultimately agreed to participate, yielding an overall acceptance rate of approximately 15.76%.

Figure 4.3 illustrates the breakdown of these recruitment efforts. We reached out to 50 individuals on LinkedIn, yielding 6 interviews (i.e., a 12% acceptance rate). Referrals generated 10 contacts, of whom 4 participated, representing the highest rate of acceptance (40%). Finally, we approached 16 past participants, 2 of whom agreed to an interview (12.5% acceptance). It was noteworthy that the acceptance rate among previous participants was not substantially higher than that achieved through LinkedIn, suggesting that repeated engagement does not necessarily translate into more favorable responses. Nonetheless, referrals proved the most effective channel in terms of conversion, highlighting the value of professional networks and personal endorsements when recruiting legal practitioners for academic research.

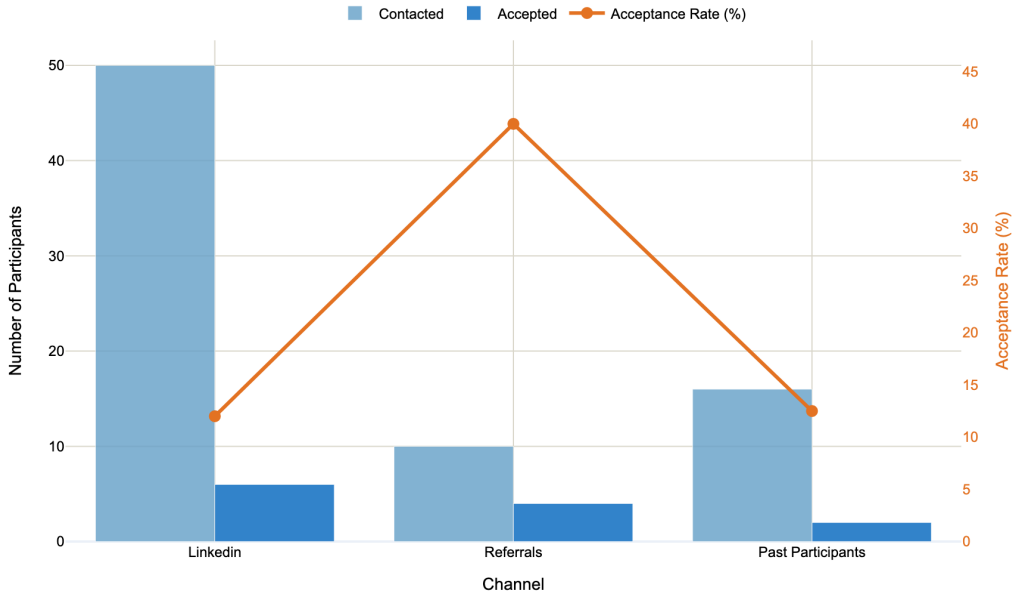


Figure 4.3.: Channel effectiveness in participant recruitment

4.4.2. Demographics of Participants

This section provides an overview of the participants' roles, organizational settings, and gender distribution. Figure 4.4 summarizes the positions held by the interviewees, while Figure 4.5 illustrates their employer sizes (or principal affiliations) along with a gender breakdown.

Position Distribution

As shown in Figure 4.4, the majority of interviewees were *Attorneys* ($n = 10$), supplemented by one *Law Student* ($n = 1$) and one *Prosecutor* ($n = 1$). Within the attorney group, three participants served as *Heads of Legal Departments* at large companies, two were *Entrepreneurs* owning their own law firms, one was an *Advocate*, one was a *Chief Legal Counsel* at a large international organization, and three held various in-house or corporate counsel positions.

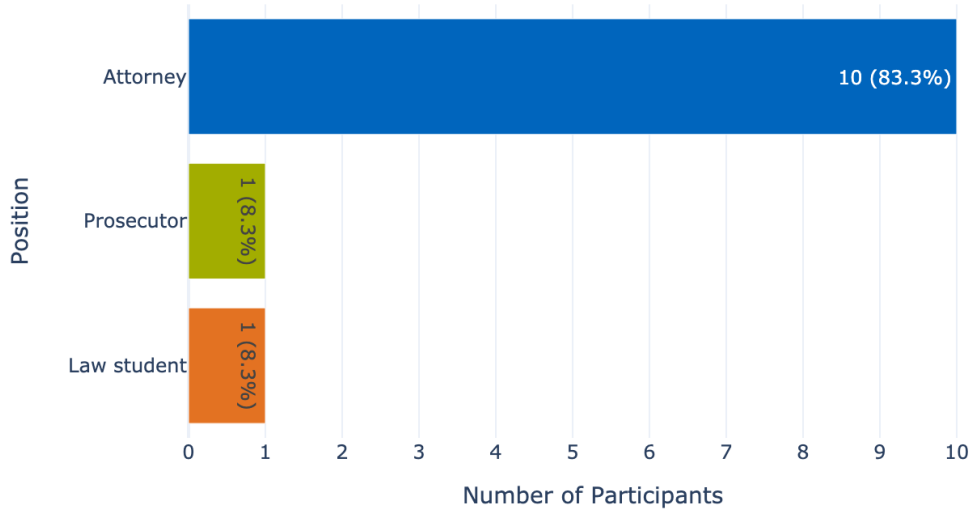


Figure 4.4.: Distribution of participant positions in the legal field.

Company Size and Gender Distribution

Figure 4.5 presents the size of participants' employing entities, referencing the EU recommendation 2003/361 [35] for micro, small, medium, and large enterprises. Additionally, the category *State Institutions* applies to government entities, and the category *Student* covers participants not externally employed. Out of the total sample, 9 participants identified as male and 3 as female, suggesting potential indications of gender imbalance within this subset of legal professionals.

Summary of the Interview Process

Each participant was assigned an anonymized ID (*I-1*, *I-2*, etc.), as summarized in Table 4.8. While the first two interviews slightly exceeded the planned 30-minute limit, subsequent refinements to the interview guide ensured that most interviews thereafter remained closer to 29 minutes on average. Overall, the discussions provided rich insights into how legal

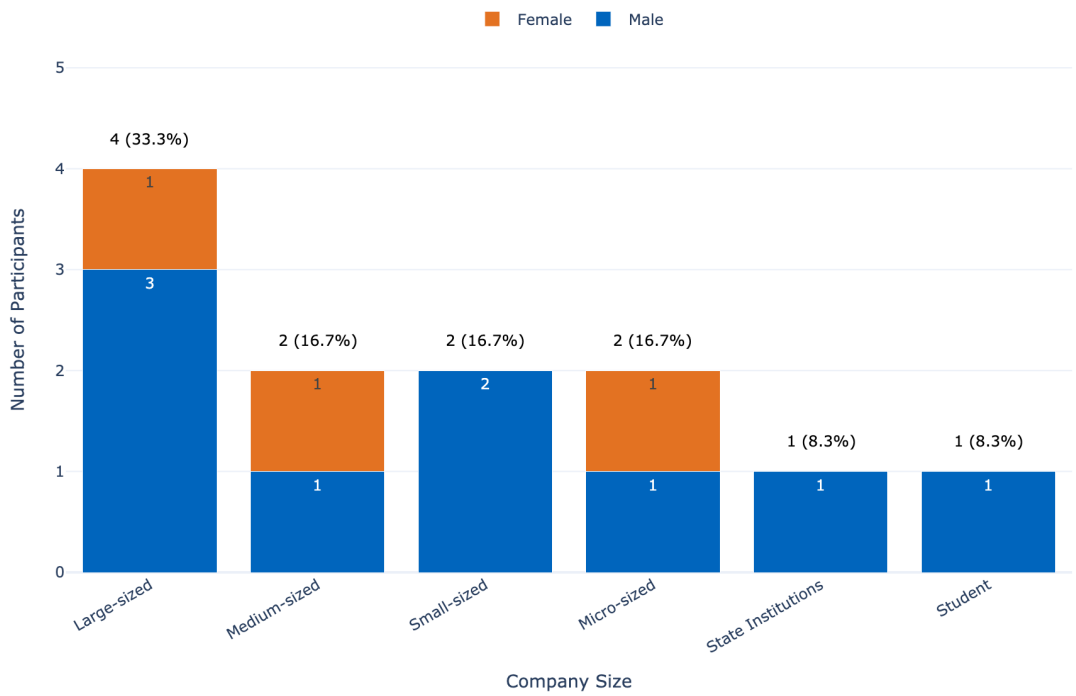


Figure 4.5.: Distribution of participants by employer size category and gender.

practitioners in diverse organizational contexts perceive and prioritize different NLP use cases.

Table 4.8.: Overview of key participant information.

ID	Position	Company Size	Gender	Experience (years)	Duration (mins)
I-1	*Attorney*	Small	Male	20–25	37
I-2	*Attorney*	Large	Male	20–25	35
I-3	*Attorney*	Medium	Male	5–10	28
I-4	*Law student*	Student	Male	0–5	25
I-5	*Prosecutor*	State Institutions	Male	5–10	32
I-6	*Attorney*	Micro	Female	5–10	27
I-7	*Attorney*	Medium	Female	10–15	30
I-8	*Attorney*	Large	Female	15–20	33
I-9	*Attorney*	Large	Male	10–15	25
I-10	*Attorney*	Small	Male	10–15	29
I-11	*Attorney*	Large	Male	20–25	23
I-12	*Attorney*	Micro	Male	20–25	25

5. Results

This chapter presents the primary findings of the thesis, derived from two main sources: the performance evaluation of our fine-tuned NLP models and a large-scale analysis of the classified papers. We begin by examining how accurately the models classify relevant legal NLP publications and assign them to specific use cases or NLP techniques. We then delve into broader patterns observed across the full corpus, highlighting both category-level and subcategory-level distributions. Finally, we summarize insights from semi-structured interviews, showcasing how legal practitioners rank these AI-based applications in their day-to-day work. Together, these results offer a multifaceted perspective on legal NLP's current landscape, providing both quantitative metrics for model performance and a qualitative understanding of real-world priorities.

5.1. Model Performance Results

5.1.1. Legal NLP Relevance Classification

Following the approach outlined earlier, the dataset was split into training and test subsets, comprising a total of 233 papers (168 labeled as "Relevant" and 65 labeled as "Not Relevant"). As illustrated in Figure 4.2, the test set contained 76 documents (15 "Not Relevant" and 61 "Relevant").

A GPT-4o-based language model was fine-tuned on the training data to classify each paper according to its Legal NLP relevance. This fine-tuned model was compared against a zero-shot (base) approach introduced in Section 4.2.4. Table 5.1 presents the confusion matrices for both models, and Table 5.2 shows the performance metrics. The fine-tuned model improves overall accuracy from 0.72 to 0.95, with an F_1-score of 0.97 for the "Relevant" class (compared to 0.80 in the zero-shot scenario).

A closer inspection of misclassifications suggests that the base model's errors predominantly arise when domain-specific legal terminologies are unclear under zero-shot conditions. After exposure to representative training data, the fine-tuned model more accurately associates these domain-specific cues with Legal NLP. Occasional misclassifications by the fine-tuned model often involve interdisciplinary papers with ambiguous contextual cues, indicating that even with fine-tuning, edge cases can present difficulties for automated classification.

5.1.2. Use-Case Classification

This section evaluates the model's performance at both the *category* and *subcategory* levels. We begin by examining category-wide metrics, using Figure 5.1 to illustrate overall gains

Table 5.1.: Confusion matrices for the zero-shot and fine-tuned models.

	Zero-Shot Model		Fine-Tuned Model	
	Pred. Not Rel.	**Pred. Rel.**	**Pred. Not Rel.**	**Pred. Rel.**
Not Rel. (15)	13	2	14	1
Rel. (61)	19	42	3	58

Table 5.2.: Performance metrics comparing the zero-shot and fine-tuned models.

Metric	**Zero-Shot Model**	**Fine-Tuned Model**
Precision (Not Rel.)	0.41	**0.82**
Recall (Not Rel.)	0.87	**0.93**
F_1 (Not Rel.)	0.55	**0.87**
Precision (Rel.)	0.95	**0.98**
Recall (Rel.)	0.69	**0.95**
F_1 (Rel.)	0.80	**0.97**
Accuracy	0.72	**0.95**
Macro Avg. F_1	0.68	**0.92**
Weighted Avg. F_1	0.75	**0.95**

or declines from fine-tuning. Subsequently, we delve into subcategory-specific outcomes, highlighting how task complexity and variations in training data affect classification accuracy. Taken together, these perspectives provide a comprehensive look at how well the model distinguishes among diverse legal NLP use cases.

Category-Level Performance

Figure 5.1 depicts precision, recall, and F_1 metrics for each major category, along with net improvements (in green) or declines (in red) following fine-tuning. Categories with larger subcategory support (*Legal Decision Making & Dispute Resolution*, *Legal Research & Information Management*, and *Information Processing & Extraction*) display particularly pronounced F_1 boosts, ranging from +33.5% to +61.8%. These gains are consistent with the availability of well-represented tasks such as *Judge: Decision Making* (10 documents), *Information Extraction* (24 documents), and *Research Tool / Research Automation* (18 documents).

By contrast, *Compliance & Risk Management* exhibits a -6.1% decrease in F_1, attributable mostly to *GDPR Compliance* (only three training documents), underscoring the model's difficulty in generalizing when subcategories lack sufficient coverage. Nonetheless, most categories profit significantly from fine-tuning. These aggregated results point to the efficacy of additional domain-specific training for broad task classes, while also signaling the importance of robust data representation across subcategories.

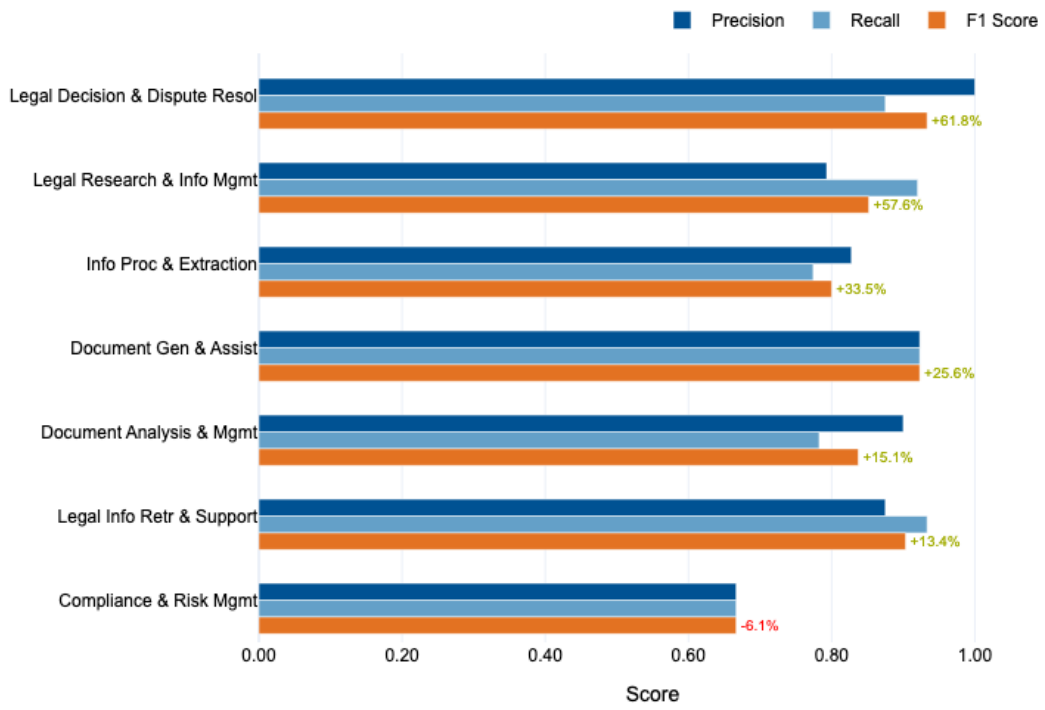


Figure 5.1.: Category-level precision, recall, and F_1 scores (fine-tuned minus zero-shot). Green labels indicate positive gains; red labels represent declines.

Subcategory-Level Performance

Table 5.3 presents a granular breakdown of zero-shot versus fine-tuned F_1 scores for individual subcategories within these categories. Many subcategories register notable improvements, particularly those supported by larger volumes of training data. For instance, *Judge: Decision Making* increases from 0.18 to 0.90, and *Automation of Auditing* rises from 0.50 to 0.86. *Information Extraction* (24 documents) also benefits significantly, improving from 0.40 to 0.74. These results indicate that tasks tied to well-populated subcategories gain most from the additional learning signal offered by fine-tuning.

Nevertheless, certain subcategories with minimal support remain challenging. *GDPR Compliance* falls from 0.80 to 0.57, mirroring the broader decline in *Compliance & Risk Management*. *Anonymization / Text Scrubbing* and *e-Discovery* similarly exhibit modest or no improvement, suggesting that subcategory balance and diversity are critical for maximizing classification robustness. In sum, combining category-level insights with a detailed subcategory breakdown underscores both the strengths and limitations of fine-tuning, emphasizing that data size and task specificity play pivotal roles in model performance.

Table 5.3.: Subcategory-level use-case classification results.

Category	Subcategory	F1 (Zero-Shot)	F1 (Fine-Tuned)	Support
Compliance and Risk Management (n=8)				
	Automation of Auditing	0.50	**0.86**	3
	Risk Assessment	0.50	**0.67**	2
	GDPR Compliance	**0.80**	0.57	3
Document Analysis and Management (n=26)				
	Document Classification	0.71	**0.80**	11
	File Difference Tracking	1.00	1.00	3
	Content Lifecycle Management	0.00	**0.80**	3
	Error Detection	0.00	**0.73**	6
	Smart Contract Analysis	1.00	1.00	3
Document Generation and Assistance (n=16)				
	Legal Language Interpretation	0.67	**0.91**	6
	Legal Document Enrichment	0.00	**0.50**	2
	Contract Generation	0.67	**0.80**	2
	Class Action Lawsuits	0.67	**1.00**	2
	Summarization	0.75	**1.00**	4
Information Processing and Extraction (n=35)				
	Document Retrieval	0.44	**0.60**	5
	Information Extraction	0.40	**0.74**	24
	Transcription	0.67	**0.86**	4
	Anonymization / Text Scrubbing	0.50	0.50	2
Legal Decision Making and Dispute Resolution (n=18)				
	Legal Reasoning	0.00	**0.50**	3
	Strategy Recommendations	0.00	**0.67**	2
	Judge: Decision Making	0.18	**0.90**	10
	Dispute Resolution Mechanism	0.50	**0.80**	3
Legal Information Retrieval and Support (n=18)				
	Question Answering	0.86	**0.93**	7
	Ranking of Lawyers	1.00	1.00	1
	Translation	0.57	**0.89**	5
	Credibility of Witnesses	0.67	**1.00**	2
	Chatbot	0.00	**0.67**	3
Legal Research and Information Management (n=32)				
	Law Systems Divergence	1.00	1.00	1
	Research Tool / Research Automation	0.20	**0.79**	18
	Changes in Law	0.00	**0.67**	2
	Legal Corpus Curation	0.00	**0.75**	9
	e-Discovery	0.67	0.67	2

5.1.3. NLP Technique Classification

This subsection applies the same two-tier analytical approach used in Section 5.1.2, but focuses on NLP techniques rather than legal use cases. The model was evaluated on 61 test papers determined to be legally relevant. Subsection 5.1.3 discusses category-level results, referring to Figure 5.2, while Subsection 5.1.3 examines subcategory-level performance. Where pertinent, parallels to the use-case classification outcomes are highlighted.

Category-Level Performance

Figure 5.2 illustrates precision, recall, and F_1 gains (green) or limited improvements (orange) for seven main NLP technique categories. *Document Analysis* achieves the largest F_1 increase (+43%), driven by enhancements in tasks like *Document Similarity Analysis* and *Entity Linking*. Similar to the trends observed in the use-case classification (Section 5.1.2), categories with robust data representation (e.g., *Text Representation* with 41 labeled instances) often exhibit more substantial performance boosts. Notably, *Text Representation* improves by approximately +29.7%, thanks largely to subcategories such as *Language Modeling*.

By contrast, *Text Generation* registers a more modest F_1 gain (+7.4%). This mirrors the challenges seen in certain use-case subcategories that require complex reasoning and rich contextual information (e.g., *Summarization*). Likewise, *Conversational NLP* and *Syntactic Analysis* benefit from fine-tuning (increases of +9.6% and +27.3%, respectively), although small support in specific tasks (e.g., *Chatbot Development*, *Tokenization*) seems to limit further improvements. Overall, these category-level results parallel the pattern observed in the use-case classification, underscoring that domains with ample, well-distributed training examples tend to yield more pronounced gains from fine-tuning.

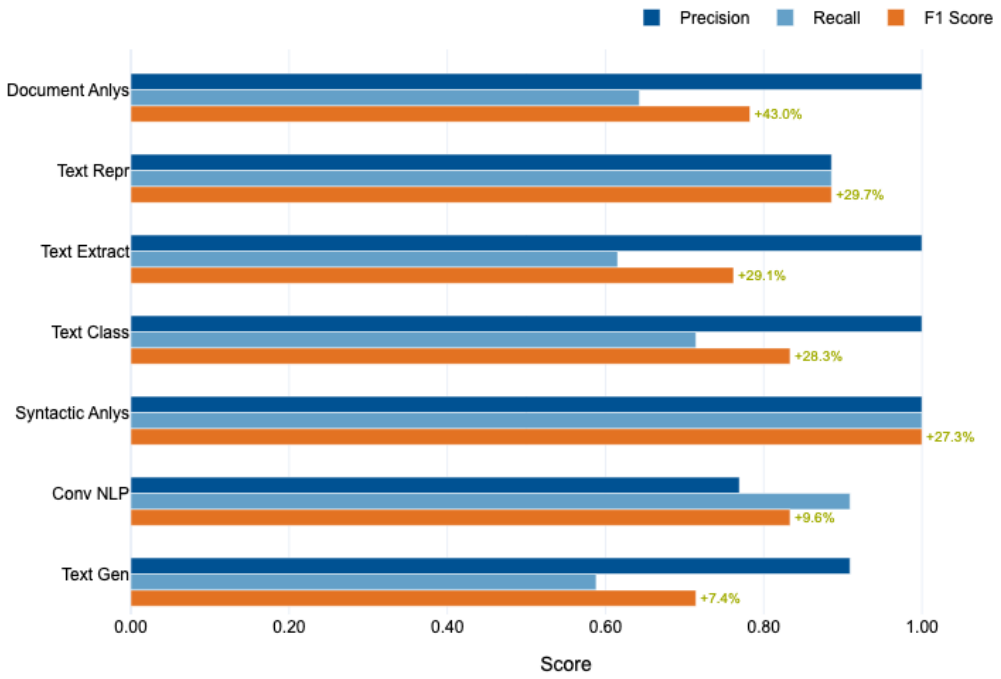


Figure 5.2.: Category-level performance for NLP techniques in 61 legally relevant test papers, showing precision, recall, F_1 scores, and net improvements after fine-tuning.

Subcategory-Level Performance

Table 5.4 provides a more granular breakdown, contrasting zero-shot (base) and fine-tuned (FT) F_1 scores across 17 subcategories. In line with the category-level findings, tasks bolstered by substantial data (e.g., *Language Modeling*, 30 documents) show marked progress, rising from an F_1 of 0.55 to 0.90. Subcategories involving *Entity Linking* and *Lexical Normalization* similarly record sizable jumps, from 0.00 to 0.86 and 0.00 to 1.00, respectively, reflecting fine-tuning's efficacy for specialized linguistic tasks once sufficient representative examples are available.

Certain techniques remain more challenging. *Chatbot Development* decreases from 1.00 to 0.80, likely influenced by a mere two test samples, illustrating how sparse data can lead to variable outcomes. Likewise, *Text Summarization* sees a moderate boost (0.50 to 0.64), hinting that complex generation tasks—similar to certain advanced use cases—may require more comprehensive datasets to fully exploit the advantages of fine-tuning. Overall, these subcategory-level observations reinforce the central conclusion that improvements depend heavily on both the complexity of the task and the size and diversity of the training corpus.

Table 5.4.: Subcategory-level NLP technique classification results.

Category	Subcategory	F1 (Zero-Shot)	F1 (Fine-Tuned)	Support
Conversational NLP (n=11)				
	Chatbot Development	**1.00**	0.80	2
	Question Answering	0.67	**0.80**	9
Document Analysis (n=14)				
	Document Similarity Analysis	0.43	**0.71**	11
	Entity Linking	0.00	**0.86**	3
Syntactic Analysis (n=9)				
	Dependency Parsing	0.89	0.89	5
	Tokenization	0.67	**0.80**	2
	Lexical Normalization	0.00	**1.00**	1
	Part of Speech Tagging	0.00	**1.00**	1
Text Classification (n=29)				
	Concept Models	0.00	**0.86**	4
	Topic Modeling	0.67	**1.00**	2
	Text Classification	0.59	**0.82**	23
Text Extraction (n=15)				
	Keyword Extraction	0.67	**0.86**	4
	Named Entity Recognition	0.31	**0.78**	11
Text Generation (n=22)				
	Machine Translation	0.83	**0.86**	7
	Text Summarization	0.50	**0.64**	15
Text Representation (n=41)				
	Language Modeling	0.55	**0.90**	30
	Word Embedding	0.43	**0.71**	11

5.2. Large-Scale Analysis of the Fine-Tuned Models

This section details the application of two fine-tuned classification models (see Section 4.2.6) to a corpus of 3,578 publications assembled according to the methodology in Section 4.2.3. The first model determines whether a paper discusses at least one of the legal NLP use-cases defined in Table 4.1 and extracts those use-cases; the second model identifies which specific NLP techniques are present in papers that meet the first criterion.

5.2.1. Overall Distribution of Papers

Figure 5.3 summarizes the outcomes of the first-stage classifier on the entire dataset of 3,578 papers. Approximately 27.6% (988 papers) were found to address one or more legal NLP use-cases, whereas 72.4% (2,590 papers) showed no explicit reference to legal NLP tasks. Although the initial query aimed to capture predominantly legal content, these results confirm that a substantial fraction of retrieved papers focus on more general NLP applications or tangential legal discussions.

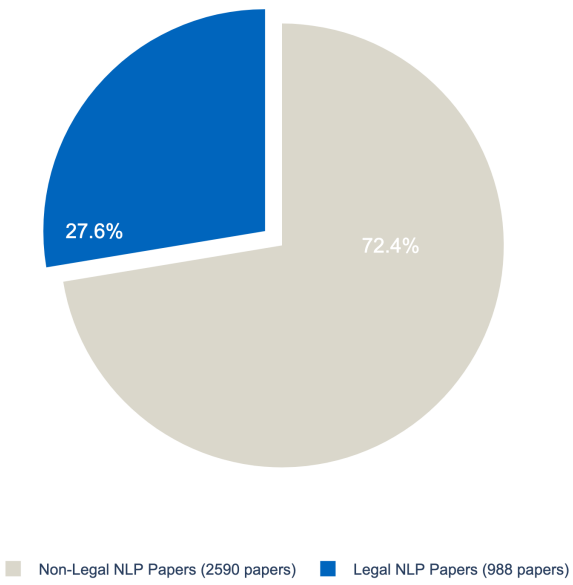


Figure 5.3.: Distribution of 3,578 publications

In addition to these overall proportions, Figure 5.4 tracks the yearly publication trend for the 988 papers containing at least one legal NLP use-case. The steep increase after 2019–2020 suggests heightened interest in legal-domain applications of NLP, potentially reflecting advances in large language model (LLM) architectures, the availability of more extensive legal datasets, or emergent regulatory demands. A detailed exploration of subcategory-level growth drivers will be presented later in this *Results* chapter, offering insights into which specific use-case may be fueling this recent surge.

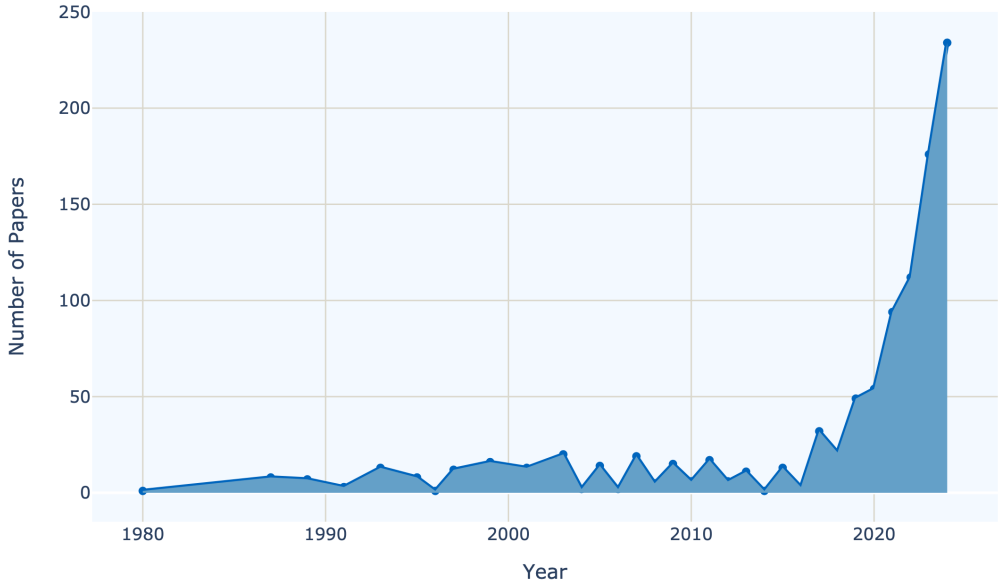


Figure 5.4.: Yearly publication frequency for the relevant papers.

5.2.2. Use-Case Category-Level Distribution

Figure 5.5 reports the total number of *category label assignments* across the 988 papers identified as containing at least one legal NLP use-case. Unlike a single-label scenario, each paper can receive multiple labels within the same category if it has multiple subcategory assignments (see Section 5.2.3). Thus, the total count of category labels may exceed the unique paper count.

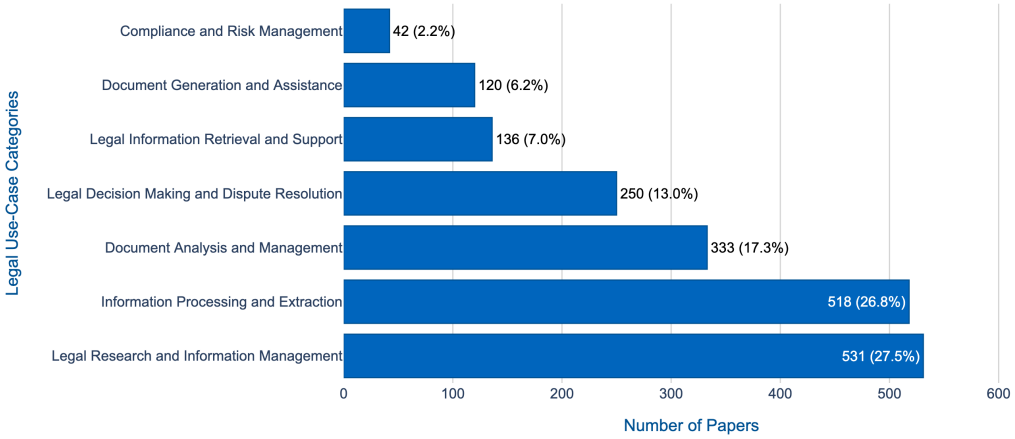


Figure 5.5.: Frequency of high-level use-case category labels across the 988 legally relevant papers.

As displayed in Figure 5.5, *Legal Research & Information Management* accounts for 531 label assignments (**27.5%** of the total), followed by *Information Processing & Extraction* with 518 assignments (**26.8%**). Together, these two categories represent over half of all category labels, indicating a significant focus on tasks such as legal-information retrieval, the structuring or enrichment of large legal corpora, and automated research tools.

Document Analysis & Management comprises 333 assignments (**17.3%**), reflecting ongoing research into classifying and organizing legal documents, while *Legal Decision Making & Dispute Resolution* appears in 250 label assignments (**13.0%**), demonstrating continued interest in predictive modeling for court outcomes or AI-driven dispute resolution protocols. *Legal Information Retrieval & Support* (136 assignments, **7.0%**) and *Document Generation & Assistance* (120 assignments, **6.2%**) add further insight into the rising importance of interactive legal systems and automated text-generation strategies (e.g., summarization, drafting, translation). Finally, *Compliance & Risk Management* claims the smallest share of label assignments, at 42 (**2.2%**), which may reflect limited data availability or a more nascent stage of academic focus in areas like auditing automation or GDPR compliance.

Overall, these category-level label assignments confirm that the bulk of legal NLP work addresses the foundational challenges of large-scale information processing, retrieval, and document management, while also highlighting a substantial—though comparatively smaller—emphasis on decision support, compliance, and automated text handling. The next subsection (Section 5.2.3) explores how these high-level categories break down into more specific subcategories, providing a finer-grained perspective on current research priorities and emerging topics in legal NLP.

5.2.3. Use-Case Subcategory Distributions

Figure 5.6 shows how 37 subcategories are assigned to the 988 papers identified as discussing legal NLP use-cases. The vertical bars reflect each subcategory's proportion within its parent

category, whereas the percentages reported below indicate the fraction of the *entire 988-paper corpus* to which each subcategory applies. The text highlights five highly prevalent subcategories, five rare ones, two newly introduced subcategories, and comments on two subcategories not found at all.

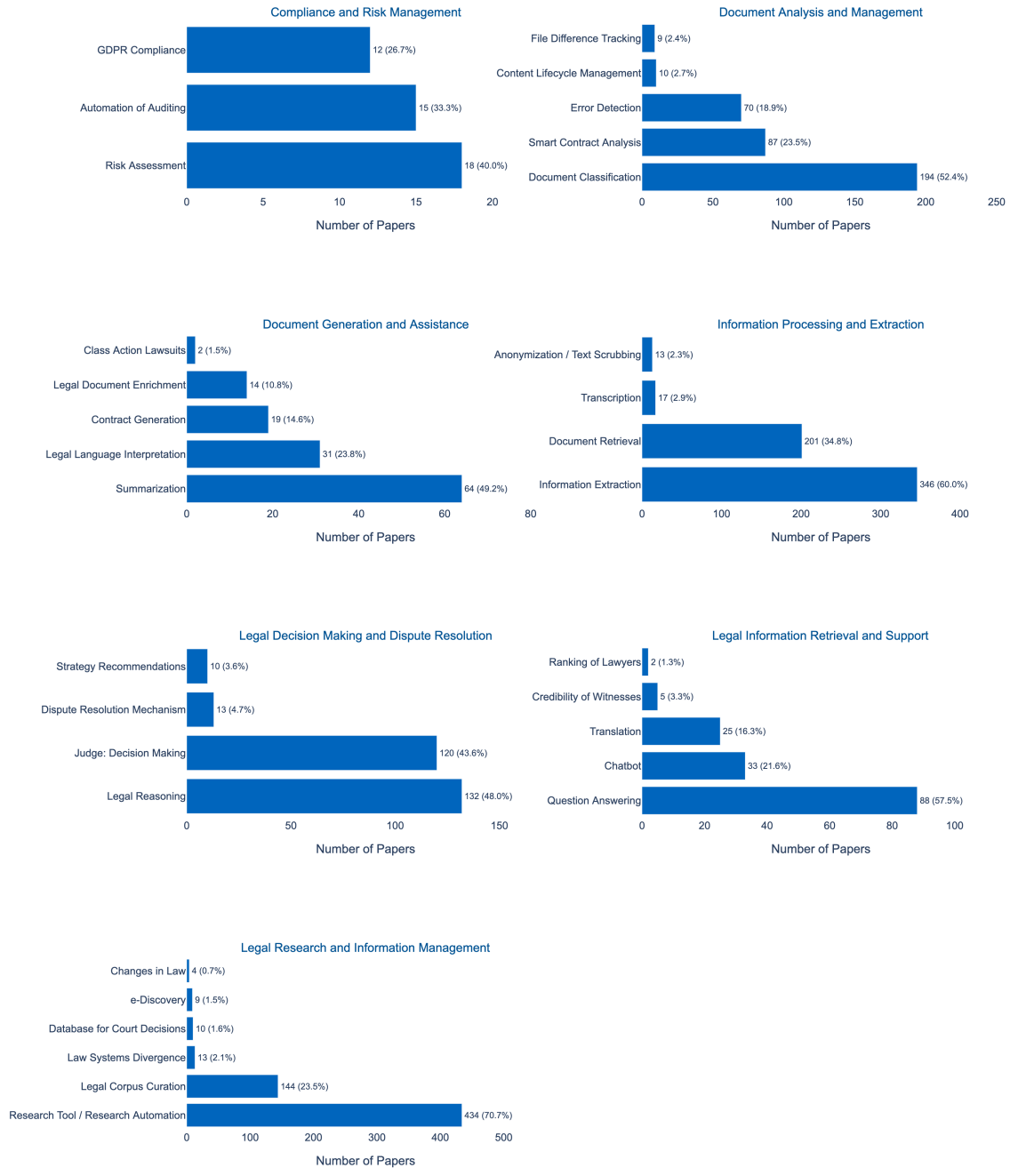


Figure 5.6.: Distribution of subcategories within the seven high-level legal use-case categories.

Top Five Most Frequent Subcategories.

- **Research Tool / Research Automation (434 papers, 43.9%).** Although initially perceived as an industry-driven need, many papers in this subcategory pursue purely academic

goals, such as creating benchmarks, testing new algorithms on complex legal texts, or demonstrating novel retrieval and classification methods. Consequently, while real-world uptake remains somewhat limited, this line of work is especially popular among researchers aiming to push technical boundaries or establish new state-of-the-art results in legal NLP.

- **Information Extraction (346 papers, 35.0%).** Systematic extraction of entities, facts, and relationships from legal texts is a cornerstone of data-driven legal analytics. Methods typically involve transformer-based models (e.g., BERT, RoBERTa) fine-tuned on domain-specific corpora, with emphasis on robust annotation protocols given the high stakes of misidentifying critical information.
- **Document Retrieval (201 papers, 20.3%).** Legal retrieval engines are increasingly incorporating neural ranking models and retrieval-augmented generation systems. These papers underscore the importance of accurate, domain-specific indexing and often combine traditional methods (like BM25) with cutting-edge embedding-based approaches, aiming to handle the nuanced semantics inherent to statutes, case law, and regulatory texts.
- **Document Classification (194 papers, 19.6%).** Tasks range from standard topic classification to more granular labeling (e.g., identifying document types, procedural stages, or jurisdictions). Many studies adopt advanced language models, training on large annotated sets to handle the complex taxonomy of legal documentation. Ensemble and transfer learning approaches also appear, reflecting the diversity of classification objectives within legal corpora.
- **Legal Corpus Curation (144 papers, 14.6%).** ***(Novel)*** This new subcategory spans data collection, preprocessing, annotation, and maintenance. Efforts include filtering noisy or duplicate records, linking texts to external knowledge bases, and constructing splits for downstream tasks such as classification or summarization. Researchers increasingly recognize that high-quality, domain-specific corpora are critical for achieving credible results in legal NLP benchmarks.

Five Least Frequent Subcategories.

- **Ranking of Lawyers (2 papers, 0.2%).** Automated assessments of attorney performance remain nearly unexplored, likely due to data sensitivity and the difficulty of codifying "quality" metrics for legal counsel.
- **Credibility of Witnesses (5 papers, 0.5%).** Although crucial in trials, witness credibility garners limited study. The few existing approaches use linguistic cues or external validation metrics, pointing to substantial methodological challenges in operationalizing subjective credibility factors.
- **Changes in Law (4 papers, 0.4%).** Automatically detecting newly enacted or amended legislation entails monitoring multiple overlapping sources. The scarcity of robust

multilingual or multi-jurisdictional datasets may explain its low presence, despite clear importance in compliance and legal research.

- **Strategy Recommendations (10 papers, 1.0%).** Offering tactical guidance (e.g., settlement strategies, negotiation stances) is a multifaceted task. Researchers have proposed models blending historical outcome data with heuristic-driven recommendations, yet the complexity and variability of legal strategy hinder a cohesive research focus.
- **File Difference Tracking (9 papers, 0.9%).** This subcategory targets automated detection of textual changes (e.g., contract revisions), but relatively few papers tackle systematic diff algorithms tailored to legal formats. Most rely on generic text comparison methods lacking fine-grained domain knowledge.

Two Additional Novel Subcategories.

- **Smart Contract Analysis (87 papers, 8.8%).** Falling under *Document Analysis & Management*, these papers examine blockchain-based agreements for vulnerabilities, compliance checks, or interpretability. Proposed frameworks often integrate symbolic reasoning with NLP-based semantic parsing to ensure robust, verifiable smart contracts.
- **Legal Language Interpretation (31 papers, 3.1%).** Work in this subcategory aims to disambiguate legal terminology, parse complex contractual clauses, or align cross-lingual domains. Techniques frequently employ specialized embeddings or dictionary-based expansions of transformer models, underscoring the unique lexical and semantic demands of legal language.

Subcategories Not Present. Neither *E-Mail* nor *Deadline Management* appears in the dataset, highlighting a divergence between certain industry-driven tasks and the focus of academic research. Despite practitioners' emphasis on scheduling and communication workflows, no papers in the corpus explicitly address these issues. This discrepancy underscores the potential for more applied or collaborative research bridging academic novelty and real-world legal workflow requirements.

5.2.4. NLP Technique Identification

Having established which papers discuss at least one legal NLP use-case (Section 5.2.3), we now apply the second fine-tuned model (see Section 4.2.6) to those 988 documents to identify the NLP techniques employed. Figure 5.7 shows that 70.2% (694 papers) explicitly mention at least one NLP technique in their abstracts, whereas 29.8% (294 papers) do not. Although some of these non-identifications may reflect papers focusing on conceptual or theoretical legal issues, it also raises the possibility that certain legal NLP approaches are not clearly described at the abstract level.

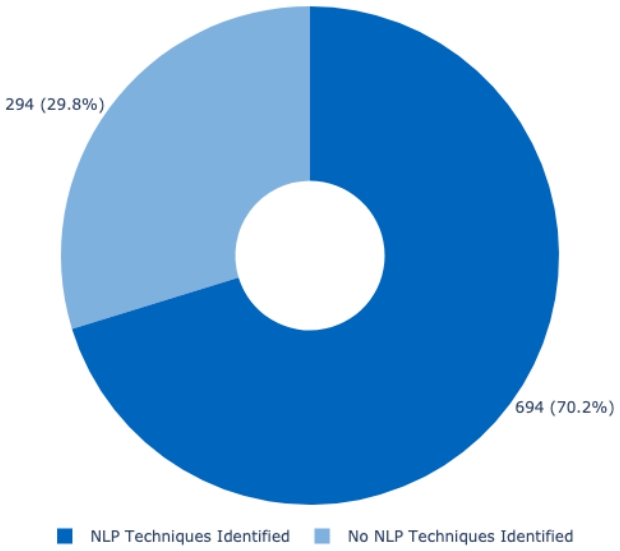


Figure 5.7.: Proportion of the 988 legal NLP papers whose abstracts include an identifiable NLP technique.

5.2.5. NLP Technique Category-Level Distribution

Figure 5.8 illustrates the overall frequency of each NLP technique category across all label assignments in our set of 988 legally relevant papers. Since a single paper can receive multiple labels (even from the same technique category), the counts and percentages in the figure refer to the total number of label assignments, rather than the unique number of papers. For example, if two distinct subcategories under "Text Representation" appear in the same paper, that paper contributes *two* label assignments toward the total for that category.

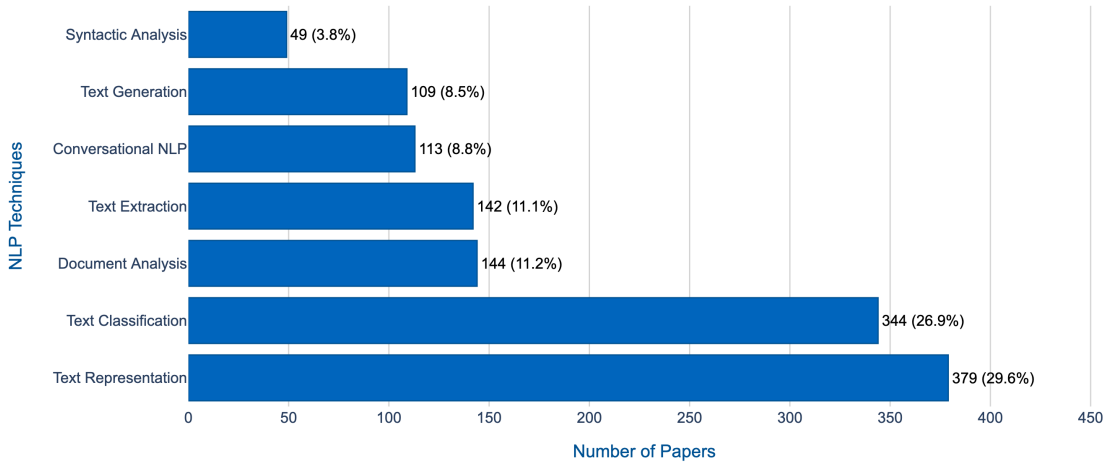


Figure 5.8.: Frequency of NLP technique category labels.

As shown in Figure 5.8, *Text Representation* (379 label assignments) accounts for **29.6%** of all assigned technique labels, making it the largest single category. *Text Classification* follows with 344 label assignments (**26.9%**), while *Document Analysis* (144, **11.2%**) and *Text Extraction* (142, **11.1%**) each represent roughly one-tenth of the total. *Conversational NLP* (113, **8.8%**) and *Text Generation* (109, **8.5%**) point to steady interest in dialogue, summarization, and translation, although at lower frequencies than classification or representation. Finally, *Syntactic Analysis* (49, **3.8%**) emerges as the smallest category, which may indicate that tasks like parsing or part-of-speech tagging are frequently embedded within broader approaches rather than emphasized as primary research focus.

5.2.6. NLP Technique Subcategory Distribution

Figure 5.9 breaks down the 988 legally relevant papers according to seven core NLP technique categories—*Syntactic Analysis*, *Text Extraction*, *Document Analysis*, *Text Representation*, *Text Generation*, *Conversational NLP*, and *Text Classification*—and their corresponding subcategories.[1] Unlike the legal use-case taxonomy, no subcategory here is entirely absent or newly introduced; however, notable differences emerge in popularity and technical approaches.

[1] All percentages refer to the proportion of the entire 988-paper corpus.

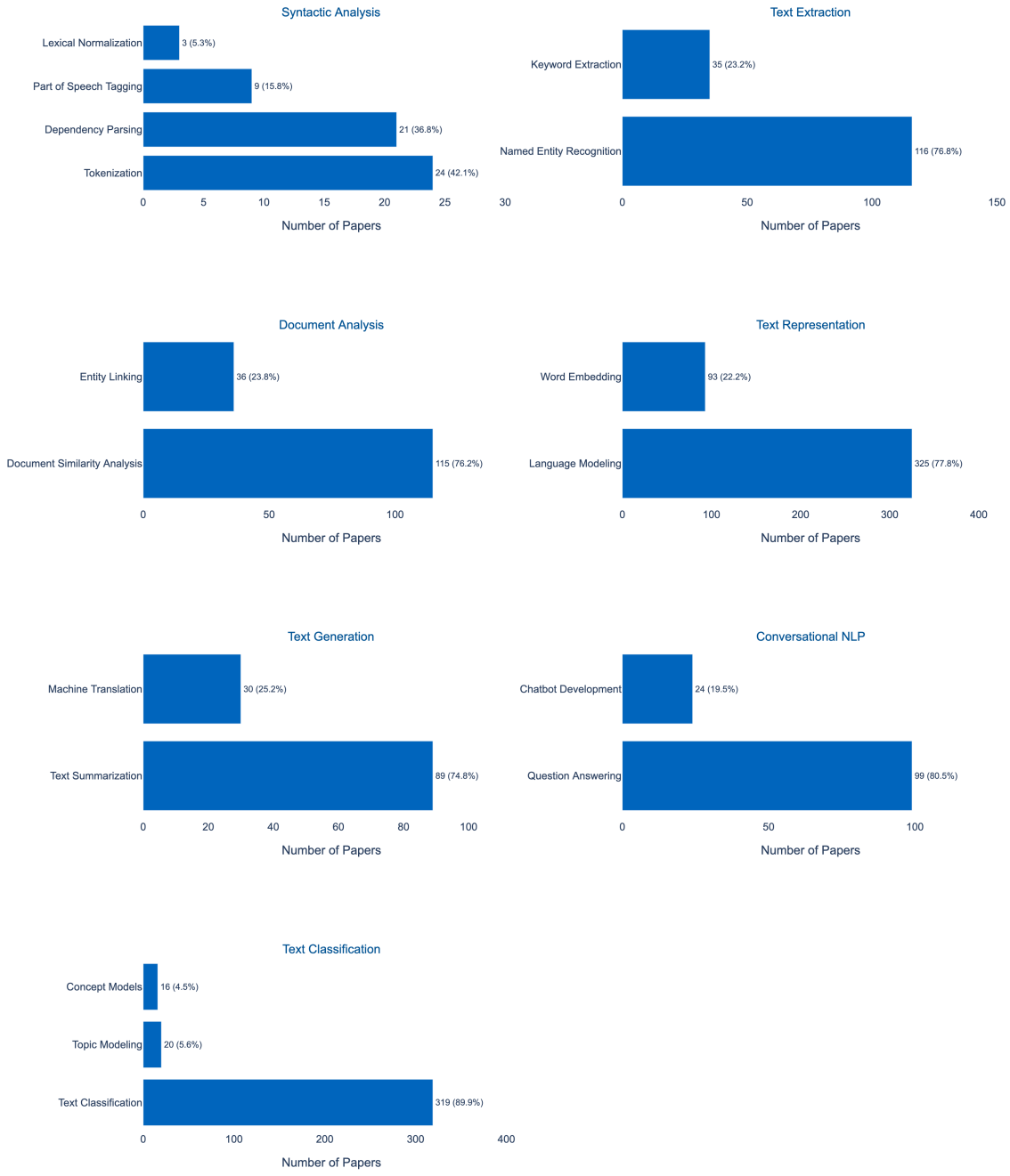


Figure 5.9.: Distribution of NLP technique subcategories within each of the seven major categories.

Syntactic Analysis.

- **Lexical Normalization (3 papers, 0.3%)** is the least common subcategory overall. Early

approaches employ rule-based or dictionary-based corrections for archaic legal terms, while recent papers use BERT-like contextual encoders to capture variant spellings. Data scarcity and a reliance on robust tokenization methods likely contribute to its low popularity.

- **Part of Speech Tagging (9 papers, 0.9%)** typically adapts standard tools (e.g., spaCy, Stanford CoreNLP) to handle legal-specific tagsets. Some studies report improved accuracy by fine-tuning domain-specific BERT models (e.g., "LegalBERT") for token-level predictions. Challenges include capturing archaic or unusually structured legal phrases.
- **Dependency Parsing (21 papers, 2.1%)** features a mix of biaffine neural parsers and transformer-based seq2seq architectures tailored to legal syntax, which can deviate from general-domain norms. Researchers integrate parsing results into tasks like argument mining or contract clause structuring, leveraging domain-adapted embeddings to boost performance.
- **Tokenization (24 papers, 2.4%)** addresses specialized segmentation rules for statutes, citations, or exhibit references. While conventional subword algorithms (e.g., WordPiece) remain common, several papers implement custom heuristics or expansions to handle nested references or archaic terms that general-domain tokenizers frequently mis-split.

Text Extraction.

- **Named Entity Recognition (116 papers, 11.7%)** ranks among the top three subcategories. Many studies fine-tune transformer-based models (e.g., BERT, RoBERTa) on legal corpora with tailored label sets (e.g., parties, statutes, court names). Some investigate multi-task training (e.g., entity linking plus NER), while others focus on domain-specific embedding initialization (LegalBERT variants).
- **Keyword Extraction (35 papers, 3.5%)** uses both statistical (TF-IDF, RAKE) and neural (deep keyphrase generation) methods. Recent works incorporate contextual embeddings to filter domain-specific terms (e.g., "subpoena," "amicus brief"), aiming for more nuanced coverage than generic keyword extraction tools.

Document Analysis.

- **Document Similarity Analysis (115 papers, 11.6%)** often applies dense embeddings (e.g., Sentence-BERT, SBERT variants) for measuring semantic overlap across contracts, case law, or statutory provisions. Common applications include precedent retrieval, near-duplicate detection, and improved e-discovery pipelines.
- **Entity Linking (36 papers, 3.6%)** extends the extraction task by mapping entities (e.g., statutory references) to official IDs or knowledge-base entries. Transformer-based mention-encoding strategies are frequent, sometimes augmented with graph-

based disambiguation or specialized legal databases (e.g., referencing official statute repositories).

Text Representation.

- **Word Embedding (93 papers, 9.4%)** covers both static (Word2Vec, GloVe) and contextual (ELMo) embeddings re-trained or fine-tuned on legal corpora. Several studies compare domain-specific embeddings with generic counterparts, reporting gains in tasks such as classification or named entity recognition once specialized lexicons are included.
- **Language Modeling (325 papers, 32.9%)** stands out as the single most frequent subcategory overall. Many papers construct or adapt large pretrained models (e.g., BERT, GPT-2) to legal text, often leading to specialized variants ("LegalBERT," "LawGPT"). Primary emphases include interpretability, domain vocabulary expansions, and empirical benchmarking on tasks like summarization or QA.

Text Generation.

- **Text Summarization (89 papers, 9.0%)** typically focuses on advanced transformer-based seq2seq models (BART, T5) for distilling lengthy court rulings or statutes. Evaluation often involves ROUGE or BERTScore, with additional qualitative checks by legal experts to capture domain nuances, e.g., legislative references or case precedents.
- **Machine Translation (30 papers, 3.0%)** applies mainstream neural translation frameworks (Marian, transformer-based seq2seq) to legal contexts. Custom dictionaries or external knowledge modules aim to preserve domain-specific terminology. Projects usually target cross-border litigation or multilingual legislative frameworks where precision is critical.

Conversational NLP.

- **Chatbot Development (24 papers, 2.4%)** describes interactive systems offering basic client assistance or preliminary legal advice. Techniques vary from rule-based dialogue flows to retrieval-augmented generation pipelines. Owing to the legal domain's complexity, many chatbots employ tight domain constraints or fallback mechanisms to minimize liability from misinterpretation.
- **Question Answering (99 papers, 10.0%)** leverages reading comprehension models or knowledge-augmented frameworks that parse statutes, contractual passages, or case law to produce succinct, authoritative answers. Studies often explore specialized passage-ranking or context windowing, refining general-purpose QA approaches to suit the verbose, formal structure of legal text.

Text Classification.

- **Concept Models (16 papers, 1.6%)** is among the least popular subcategories. Approaches typically construct conceptual ontologies or hierarchical taxonomies (e.g., liability vs. negligence) and then attempt to label text segments accordingly. Challenges stem from ambiguous or overlapping legal concepts that require intricate domain knowledge.
- **Topic Modeling (20 papers, 2.0%)** adapts unsupervised or semi-supervised methods (LDA variants, neural topic modeling) for large legal corpora. Researchers often note suboptimal fit due to polysemy and overlapping regulatory themes; hybrid models combining supervised signals or domain dictionaries sometimes address these issues.
- **Text Classification (319 papers, 32.3%)** is the second most prevalent subcategory, eclipsed only by Language Modeling. Many papers fine-tune BERT or GPT-based encoders for multi-label tasks, reflecting the multifaceted nature of legal documents (topic, jurisdiction, procedural stage). Few-shot learning and data augmentation strategies are common, tackling label imbalance and domain variation.

Collectively, these distributions underscore a core academic emphasis on large language models and classification strategies, complemented by ongoing exploration of entity-level extraction and advanced semantic techniques like document similarity or generative summarization. At the same time, specialized tasks—such as lexical normalization, concept models, or part of speech tagging—remain less frequently addressed, potentially due to robust downstream pipeline components and the comparative complexity of building or evaluating domain-targeted syntactic resources in legal NLP.

5.2.7. Use-Case Trend Analysis

Figure 5.4 reveals a significant uptick in legal NLP publications after 2019, but it does not clarify which subcategories are fueling that growth. To identify the most influential drivers, we modeled time-series data for each subcategory and retained only those with *statistically robust upward trends*, as determined by correlation-based tests with p-values below 0.5 (see Section 4.2.7 for methodological details). Subcategories that did not meet this threshold were removed to avoid misinterpreting short-lived or inconsistent fluctuations as genuine trends. Figure 5.10 depicts the normalized growth curves for eight subcategories; none exhibits a decline. In particular, three subcategories show *strong correlation coefficients* (r-values) with publication year, suggesting that they account for a large portion of the post-2019 surge:

Legal Corpus Curation. This subcategory has the largest growth rate and one of the highest r-values, indicating a consistently strong linear increase over time. Its prominence can be attributed to the recognition that high-performing domain-specific NLP systems rely on large, curated datasets that capture the complexities of legal language. Recent advancements in **Large Language Models (LLMs)** further emphasize the need for substantial, high-quality

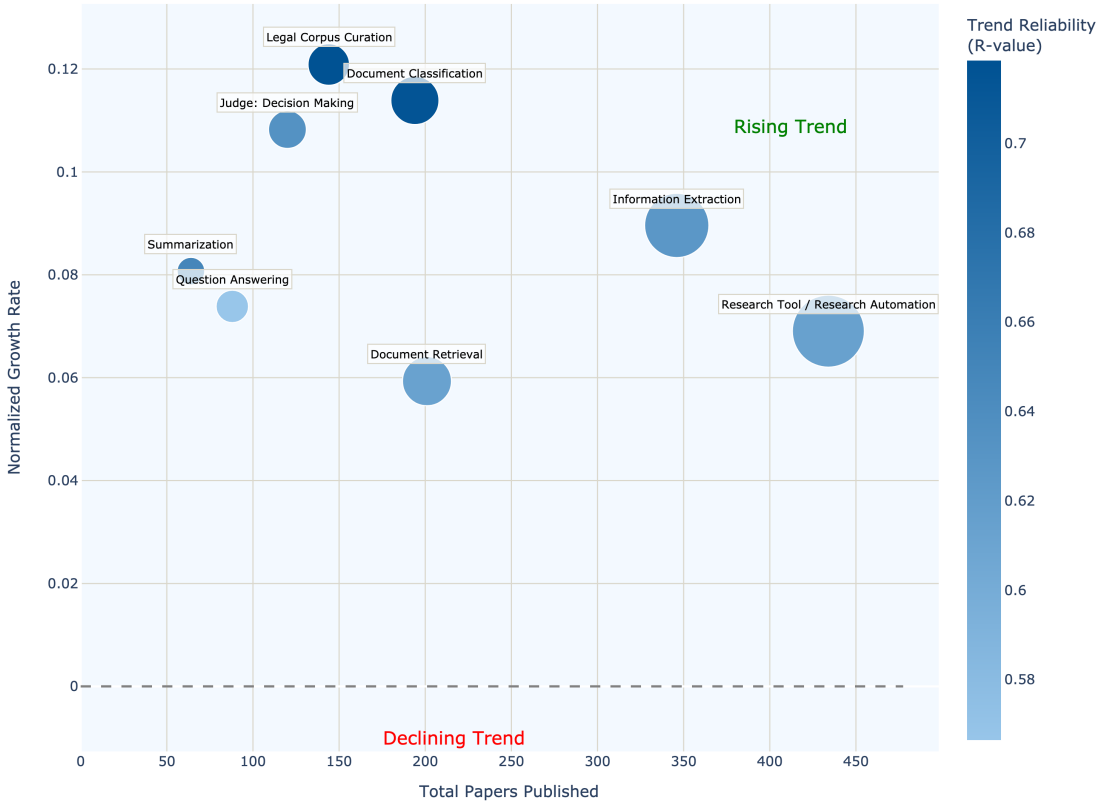


Figure 5.10.: Time-series growth trajectories for use-case subcategories with statistically significant upward trends.

corpora. Researchers therefore invest in assembling, annotating, and maintaining domain-focused data repositories, which can later be leveraged to develop robust models for tasks such as contract analysis, summarization, and question answering.

Document Classification. Long regarded as a foundational task in legal NLP, *Document Classification* has also maintained a notably high correlation with publication year, indicating renewed interest and rapid expansion. Many papers adopt **pretrained encoders** (e.g., BERT, RoBERTa) or **genuine LLMs** that support multi-label legal taxonomies. Researchers focus on improving classification accuracy, scalability, and interpretability, often introducing new benchmarks or specialized annotation schemes for legislative, case-law, or contract datasets. The combination of advanced model architectures and more extensive training resources—partly provided by corpus curation efforts—fuels this category's continued growth.

Judge: Decision Making. This subcategory similarly exhibits a robust linear trend over recent years. The underlying studies often employ **transformer-based models** for predicting or analyzing judicial rulings. Efforts to embed deeper semantic or argumentative structures have broadened the scope from outcome prediction alone to more nuanced tasks, such as extracting legal rationales or modeling appellate relationships. The consistent expansion of accessible court verdict datasets (many curated within broader *Legal Corpus Curation* initiatives) has facilitated systematic experimentation with advanced architectures, including *hierarchical transformers* designed to process lengthy rulings, as well as *graph neural networks* that capture inter-case citations.

Other Significantly Growing Subcategories. Five additional subcategories—*Information Extraction*, *Summarization*, *Question Answering*, *Research Tool / Research Automation*, and *Document Retrieval*—also manifest statistically significant growth, albeit with slightly lower r-values:

- *Information Extraction* benefits from specialized embeddings or adapted LLMs to handle intricate domain-specific entity typing and relational structures.
- *Summarization* gains traction as transformer-based seq2seq frameworks prove more adept at condensing dense legal text.
- *Question Answering* leverages retrieval-augmented pipelines for increasingly complex queries, a demand reflected in legal practice.
- *Research Tool / Research Automation* aligns with growing academic interest in building platform-like infrastructures and benchmark datasets for evaluating legal NLP tasks.
- *Document Retrieval* remains a cornerstone, with neural ranking methods continuously improved by domain adaptation and multi-stage retrieval strategies.

Crucially, no subcategory exhibits a negative slope. This universal upward trend suggests that legal NLP research is broadening—rather than consolidating—its scope. The synergy between **large-scale data curation**, **improved neural architectures**, and **LLMs** appears to be a primary catalyst, enabling research communities to tackle both established tasks (e.g., classification, retrieval) and newly emerging frontiers (e.g., advanced judicial outcome modeling). These findings set the stage for examining analogous trajectories in *NLP technique categories*, providing a complementary perspective on how methodological innovations parallel domain-specific advancements.

5.2.8. NLP Technique Trend Analysis

Figure 5.11 depicts the ascending publication trends for selected NLP technique subcategories, normalized by year. Two methods, *Text Classification* and *Language Modeling*, clearly emerge as the most dominant approaches, reflecting widespread adoption of pretrained transformers (e.g., BERT) and Large Language Models (e.g., GPT-3) in tasks ranging from document categorization to advanced sequence-to-sequence generation. Meanwhile, other subcategories,

such as *Named Entity Recognition* and *Text Summarization*, show more moderate but still consistent growth, mirroring the increasing need for automated extraction of legal entities and concise representations of extensive legal documents.

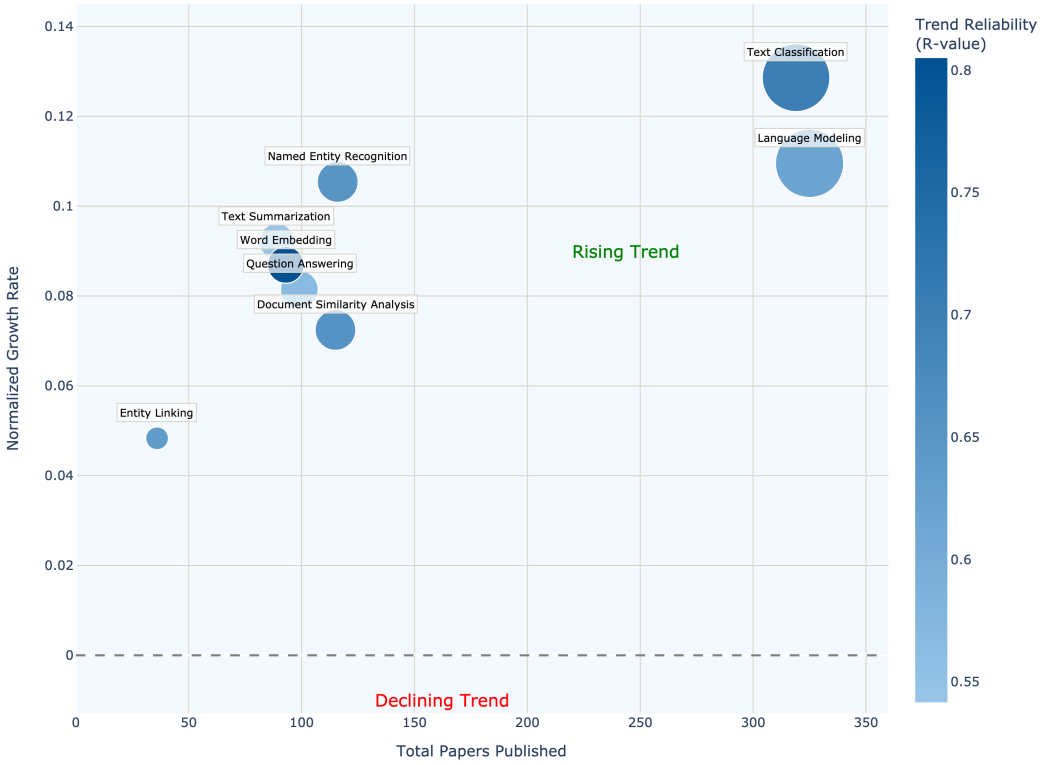


Figure 5.11.: Time-series growth trajectories for NLP technique subcategories with statistically significant upward trends.

These patterns align well with evolving legal NLP demands. For instance, *Named Entity Recognition* underpins many entity-centric use cases, while *Text Summarization* responds to practitioners' need to condense voluminous rulings or statutes. The broad adoption of *Text Classification* resonates with the prevalence of classification-centric tasks in legal text management, and the rise of *Language Modeling* reflects the shift toward domain-adapted GPT-style architectures capable of handling increasingly sophisticated generation and reasoning. Overall, the synergy between advanced model architectures and targeted legal applications appears to be driving a robust, multi-faceted expansion of NLP techniques in this domain.

5.3. Ranking Outcomes from Semi-Structured Interviews

In addition to categorizing legal NLP use-cases in the academic literature, we gathered insights from twelve legal professionals, each of whom ranked seven high-level use-case categories. Figure 5.12 presents these rankings as boxplots. A lower numeric score indicates a higher priority, and each box reflects the interquartile range (IQR), while whiskers capture the full range.

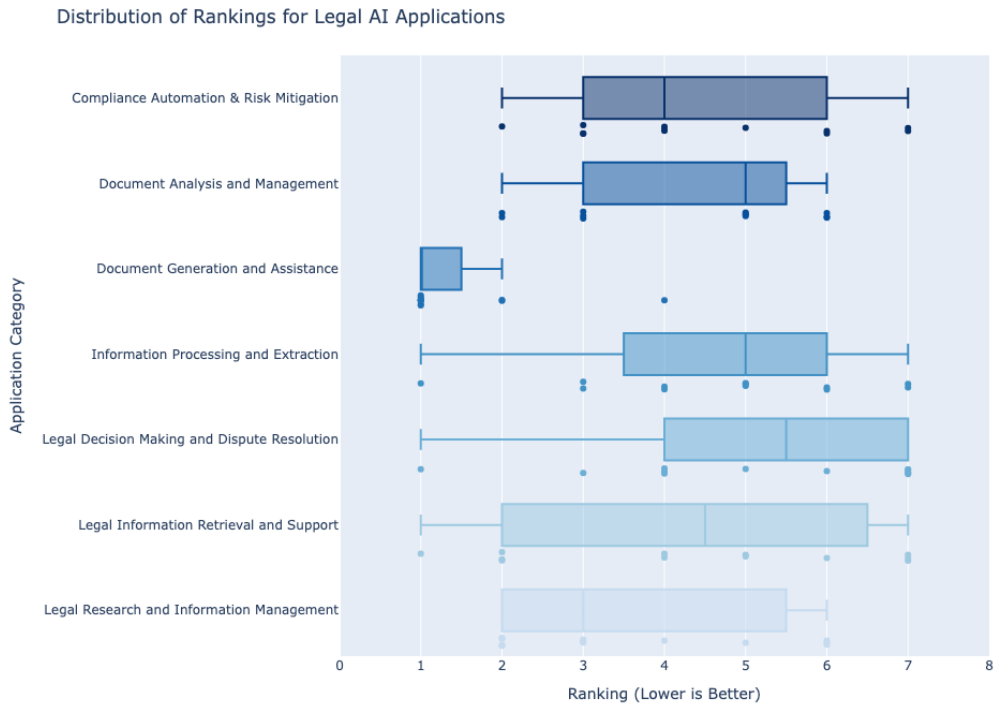


Figure 5.12.: Distribution of participant rankings for seven legal NLP categories ($n = 12$). A lower rank implies higher perceived priority.

From the boxplot, it is evident that *Document Generation and Assistance* stands out as a clear winner (located near rank 1 for most participants), with a notably tight IQR reflecting strong consensus. In second place, *Legal Research and Information Management* tends to cluster at a lower numeric rank than the remaining categories, though its IQR is broader, indicating more diverse perspectives. The other five categories group into a middle-to-lower priority cluster, each showing varying degrees of overlap and relatively higher numeric ranks. Below, we discuss the qualitative feedback for each category; direct quotes or paraphrased comments reference anonymized participant IDs defined in Table 4.8.

5.3.1. Document Generation and Assistance (Avg. Rank: 1.42)

This category appears as the top priority for the majority of participants, who consistently cited potential time savings, error reduction, and ease of deployment. Nine out of twelve participants expanded on sub-use-cases related to contract generation, automated summarization, and specialized drafting tools.

Interview Insights.

- *I-10*, an attorney at a small firm, remarked on the "significant friction" that repetitive contract drafting causes, emphasizing that "a generative solution integrated into [our] contract workflow would reduce overhead by at least 30%."
- *I-11*, head of a legal department at a large enterprise, underscored advanced summarization: "Quick overviews of complex regulations allow us to respond to internal queries faster."
- Two other participants found *language interpretation* for archaic or foreign-language clauses "promising" for bridging cross-jurisdictional gaps.

Overall, this category's narrow boxplot range reinforces a near-unanimous view that document generation yields immediate, tangible returns.

5.3.2. Legal Research and Information Management (Avg. Rank: 3.67)

Despite occupying a lower numeric rank than Document Generation, the distribution indicates moderate interest in research automation and data organization. Interviews highlight how specific tasks—*Changes in Law*, *Law Systems Divergence*, and *e-Discovery*—can drastically differ in importance depending on participants' practice areas.

Interview Insights.

- Two attorneys from large multinational settings (*I-2* and *I-8*) repeatedly stressed the importance of monitoring *Law Systems Divergence*, noting that "aligning strategies across multiple jurisdictions is the crux of global compliance."
- *Changes in Law* automation drew interest from three participants who maintain that "missing new legislation or amendments is a severe risk," though they also flagged that robust regulatory updates require "well-structured data from official sources."
- *e-Discovery* appeared essential to three in-house counsels who handle large-scale litigation, one (*I-9*) stating, "the ability to rapidly sift documents is indispensable in major disputes." Others with smaller caseloads found e-Discovery "less relevant."

5.3.3. Document Analysis and Management (Avg. Rank: 4.25)

Ranked in the mid-lower cluster, Document Analysis and Management often garnered attention for *Document Classification*. Seven participants recognized classification-based workflows as an important—though not always mission-critical—approach.

Interview Insights.

- *I-7*, an attorney in a large-sized law firm, praised classification models for "tagging thousands of corporate filings with minimal human review," speeding up internal search.
- Two participants suggested classification as a "stepping stone" to advanced analytics, such as predictive modeling or specialized retrieval.
- Others voiced caution, noting that existing content management systems already incorporate rule-based categorization; one participant (*I-1*) questioned whether "the marginal gains justify major overhauls or AI investments."

5.3.4. Legal Information Retrieval and Support (Avg. Rank: 4.33)

Professionals were split on this category's usefulness. Five participants explicitly valued advanced search or interactive support tools, but the remainder found them nonessential or prone to reliability gaps.

Interview Insights.

- *I-9*, working in a multilingual legal environment, highlighted cross-lingual retrieval: "We need to handle regulations in at least three languages, so domain-tuned search is invaluable."
- Two in-house lawyers found "chatbot or QA-based systems potentially helpful," but questioned their "domain adaptation" and "trustworthiness" if not regularly updated.

5.3.5. Compliance Automation & Risk Mitigation (Avg. Rank: 4.50)

Participants (*I-8* and *I-11*) who placed this category higher typically operate or operated before in heavily regulated domains, such as healthcare and finance. Their comments focused on the potential to preempt costly violations.

Interview Insights.

- *I-8* described *GDPR compliance checks* as "an invaluable safeguard" but acknowledged the need for specialized knowledge to interpret evolving rules.

- One counsel from a financial institution (*I-11*) praised "risk detection modules that scan large volumes of contracts for red flags," yet also noted that "tailoring them to our regulatory specifics is non-trivial."

For most others, compliance tasks appeared secondary unless they regularly navigated strict oversight.

5.3.6. Information Processing and Extraction (Avg. Rank: 4.67)

Sitting near the bottom of the chart, Information Processing and Extraction drew enthusiastic support from only a minority of participants, reflecting more specialized or bulk-processing needs.

Interview Insights.

- *I-3* singled out advanced *Named Entity Recognition* as "game-changing for large contract reviews," but also commented that "fine-tuning models to handle subtle textual variations is no small feat."
- Several participants (e.g., *I-12*, *I-4*) rely on manual extraction, citing "costly errors" or "uncertain reliability" when dealing with nuanced legal phrasing.

5.3.7. Legal Decision Making & Dispute Resolution (Avg. Rank: 5.17)

Ranked last overall, this category shows the widest spread in the boxplot, with a single outlier awarding it the top spot.

Interview Insights.

- That #1 rank came from *I-5*, a prosecutor in a state institution, who pointed to "predictive models for early resource allocation," allowing them to prioritize or decline certain cases.
- The remaining participants predominantly questioned AI's interpretive limitations, especially in ethically fraught decisions. *I-1*, a corporate counsel, remarked, "We can't outsource complex judicial reasoning to machines," reflecting a general sentiment that advanced adjudication support remains too opaque and legally sensitive for widespread adoption.

In a forthcoming Discussion chapter, these practitioner rankings will be compared against the academic trends identified in earlier sections (Sections 5.2.2–5.2.5), providing a holistic perspective on whether industry priorities align—or diverge—from prevailing research directions.

6. Discussion

6.1. Alignment and Gaps between Academia and Industry

Figure 6.1 compares the academic focus on seven legal NLP use-case categories (as identified in the SLR) with an inverted representation of practitioners' ranked priorities. In this figure, a higher line value corresponds to higher importance among professionals, while taller bars reflect more academic publications. Note that we reversed the ranking scale solely for visual consistency—no changes were made to the interview data itself.

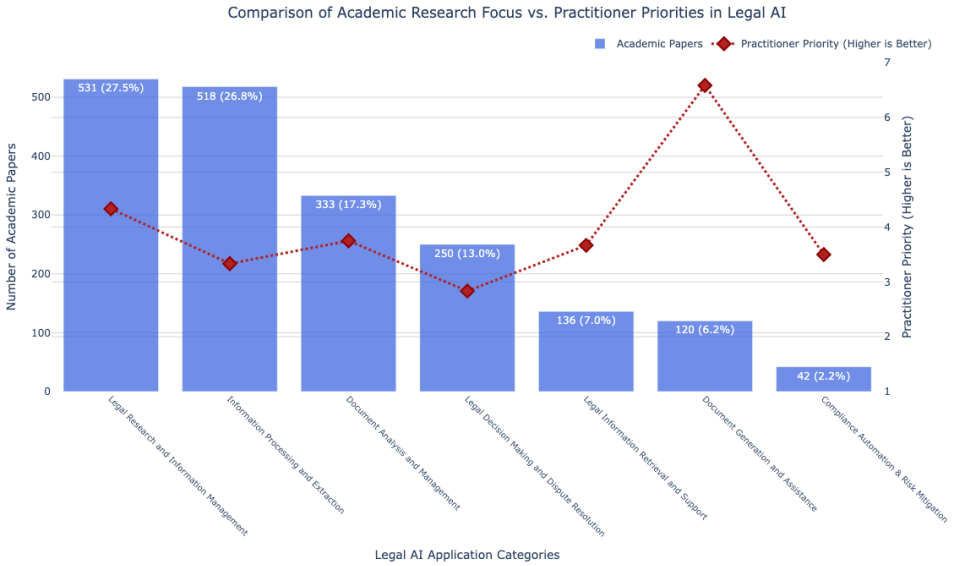


Figure 6.1.: Comparison of academic attention (*bars*) and practitioner relevance (*line*) across seven use-case categories. Higher line values indicate stronger priority among professionals, whereas taller bars represent a greater number of publications.

6.1.1. Areas of Strong Alignment

Document Analysis and Management. Academically, 333 papers fall under this category, of which more than half address *Document Classification*. Practitioners also acknowledge classification as a useful tool for organizing large volumes of legal documents—roughly half the interviewees rated classification-based workflows as beneficial for daily operations (e.g., corporate filings, docket management). Although the category as a whole ranked in the

mid-lower tier among interview participants, both groups appear to agree on classification's foundational value, suggesting further refinement in multi-label or hierarchical models can produce direct workplace benefits.

Legal Research and Information Management. Out of 531 papers in this category, a substantial majority focus on generic *Research Tool / Research Automation*. Practitioners, who collectively ranked the category second, likewise appreciate research-driven solutions for knowledge retrieval and large-scale data handling. However, they often highlighted narrower tasks—such as *Law Systems Divergence*, *Changes in Law*, and *e-Discovery*—as more critical for their day-to-day challenges. While the broad notion of "research automation" might offer a technical foundation, real-world workflows may demand specialized applications that track legislative amendments or navigate cross-jurisdictional differences.

6.1.2. Discrepancies and Potential Gaps

Over-Researched vs. Under-Prioritized?

Information Processing and Extraction. Academically, this area (518 papers) is the second-largest after *Legal Research*. Significant attention goes to entity recognition, anonymization, and document retrieval. By contrast, fewer than a third of interview participants regard such tasks as immediate priorities for their own practices; some view these tools as "not mature enough" for sensitive or nuanced legal language, while others face insufficient volumes of data to justify advanced pipelines. This mismatch suggests that while foundational research is robust, more tailored implementation strategies—aimed at domain adaptation, interpretability, and reliability—are needed to gain traction in smaller or more specialized legal contexts.

Legal Decision Making & Dispute Resolution. A total of 250 publications delve into predictive modeling, argument mining, or automated dispute settlement, yet this category ranks last among practitioners. Although a small minority of interviewees (fewer than one-fifth) considered predictive triage useful (e.g., for screening large caseloads), most remain skeptical due to interpretability, ethical risks, and dataset biases. This implies that even high-level modeling advances often fail to align with day-to-day legal practice, where liability and accountability concerns remain paramount.

Practitioners' Rising Needs vs. Scarce Academic Coverage

Document Generation and Assistance. Despite emerging as practitioners' top priority, fewer than 130 papers address this category in an explicit, use-case-focused manner. Moreover, within that subset, *Summarization* features more prominently than the other sub-use-cases (e.g., *Contract Generation*, *Legal Language Interpretation*). Interview participants consistently cited automated drafting as a game-changing application, pointing to immediate productivity gains and faster response to client queries. Yet, the academic literature tends to cluster around

more generic text-generation frameworks—few studies delve deeply into specialized drafting, interpretive guidance, or domain-based template assembly.

Compliance Automation & Risk Mitigation. Roughly 40 papers address compliance or risk-focused tasks, such as internal auditing and GDPR checks, despite a notable subset of practitioners (approximately one-third) asserting that automated risk assessment would significantly streamline regulatory workflows. Limited dataset availability and the need for in-depth domain expertise may explain academia's lower coverage. In heavily regulated industries, professionals see clear value in "red-flag detection," yet they rarely find off-the-shelf solutions that combine accuracy, adaptability, and legal interpretability—indicating a prime opportunity for new research and targeted, real-world experimentation.

6.1.3. Key Observations and Opportunities

- **Converging Foundations vs. Diverging Specializations.** While broad agreement exists on classification and research tools as foundational elements, subcategories that demand detailed domain knowledge (e.g., compliance checks, automated document drafting) show a greater industry demand than academic output.
- **Adoption Barriers for Over-Researched Topics.** Sizable academic focus on *Information Extraction* and *Decision-Making Models* has not translated into consistent industry uptake. Many professionals doubt the robustness of extraction pipelines for complex or sensitive legal texts and consider the ethical and interpretive concerns around dispute-resolution AI insurmountable for the moment.
- **Potential Future Directions.** Addressing these gaps may hinge on deeper researcher–practitioner collaboration. Researchers could channel text-generation innovations into specialized drafting and interpretation frameworks, while law firms or corporate legal departments can share real-world compliance datasets to spur more accurate, context-aware automation solutions.

Overall, the combined picture underscores a still-evolving landscape. Researchers and practitioners do converge on certain foundational tasks (like *Document Classification* and *Research Automation*), yet significant mismatches remain, especially in drafting workflows, compliance automation, and AI-driven dispute resolution. Bridging these discrepancies will likely require robust interpretability standards, domain-specific corpora, and user-centered design approaches that align advanced computational possibilities with the nuanced realities of legal practice.

6.2. Interpreting the Rising Trend of Legal NLP Publications (2010–2024)

In the Results section (Figure 5.4), we presented the general trajectory of legal NLP research spanning from 1980 to 2024. To elucidate the marked spike in recent years, Figure 6.2 below

zooms in on publications between 2010 and 2024. Notably, there is a sharp increase after 2017, with vertical dashed lines highlighting two seminal transformer-based advancements: *BERT* (2018) [14] and *GPT-3* (2020) [26]. While these milestones alone do not explain the entire surge, they coincide with a renewed academic focus on domain-specific text analytics, incentivizing researchers to tackle more complex legal tasks.

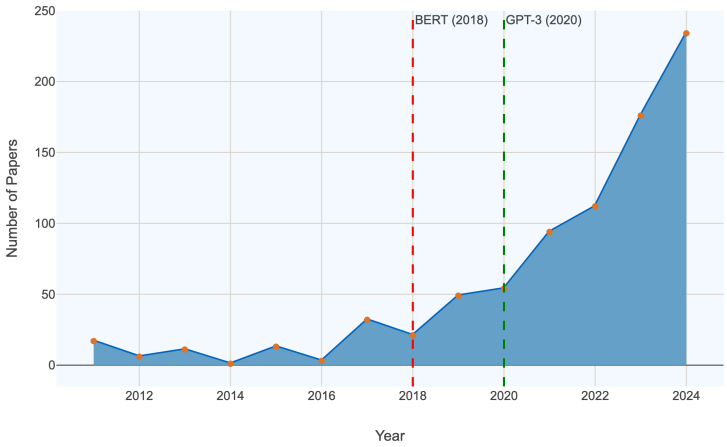


Figure 6.2.: Yearly count of Legal NLP papers (2010–2024).

Key Drivers Fueling Post-2017 Expansion

Technological Breakthroughs. A principal driver of this growth is the shift to attention-centric architectures. *BERT* introduced in 2018 [14] and *GPT-3* released in 2020 [26] elevated the state of the art in context modeling and generative NLP, respectively. As highlighted by our Trend Analysis (Section 5.2.7), domain-specific corpus creation efforts surged alongside these advances, particularly in *Legal Corpus Curation*, the fastest-growing subcategory. By assembling and refining large, specialized corpora, researchers can adapt these powerful frameworks more precisely to legal language. Examples include *LegalBERT* [36] and *Lawyer GPT* [37], each trained to handle statutory references, contract clauses, and jurisdictional nuances beyond the reach of general-purpose models. Such targeted adaptations demonstrate measurable gains in tasks like case classification, compliance checks, and contract analysis, creating a feedback loop that fuels both academic momentum and practical deployment in legal NLP.

Growing Market Visibility. Figure 6.3 demonstrates a sustained increase in Google search volumes for "Legal AI," suggesting heightened awareness among practitioners, entrepreneurs, and the broader public. The pronounced upswing from 2018 onward parallels the expansion of advanced language modeling techniques and newly emerging NLP solutions, reinforcing the

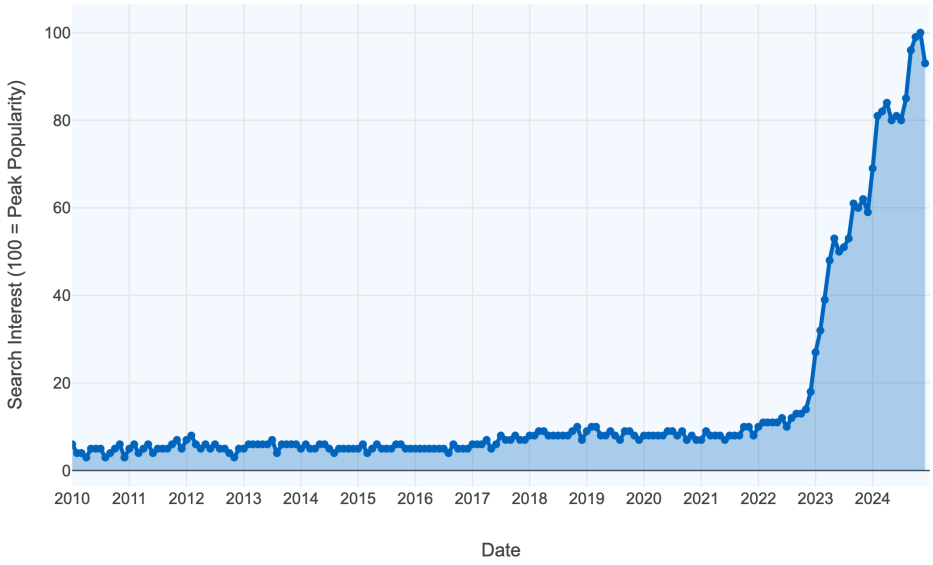


Figure 6.3.: Global search trend for the keyword "Legal AI".

notion that technological breakthroughs tend to flourish under favorable market conditions. As more stakeholders investigate potential legal automation and analytics, the spotlight on AI-driven innovation continues to intensify, encouraging greater academic and commercial efforts in legal NLP.

Accelerated Investment and Record Funding. An equally pivotal factor is the influx of capital into legal tech startups and established ventures (Figure 6.4). After moderate growth in the mid-2010s, investments surged post-2018, culminating in over $5 billion of global funding in 2024 [38]. This infusion of resources supports real-world pilot deployments of contract-generation tools, compliance checkers, and domain-adapted NLP pipelines, all of which stimulate new academic inquiry and data-sharing opportunities. As legal tech companies scale their offerings, they generate use-case feedback and curated corpora that further refine model architectures, thus reinforcing the observed correlation between increased funding levels and heightened legal NLP publication rates.

6.3. Representative Papers by Use-Case Category

In this section, we highlight a selection of particularly notable or popular papers—identified via our automated pipeline—that exemplify each use-case category in practice. By focusing on these high-impact works, we illustrate how diverse NLP methodologies are applied across sub-use-cases.

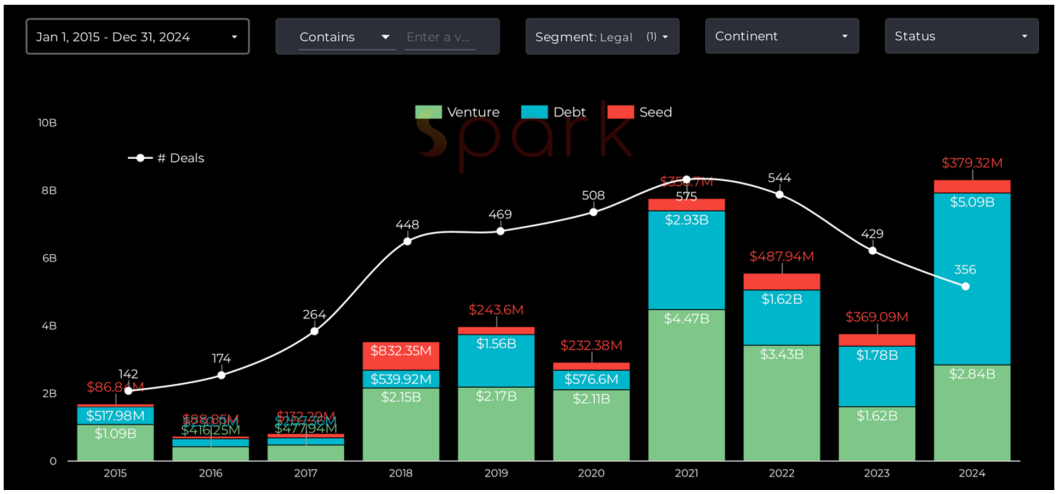


Figure 6.4.: Data: Legal Complex 2025 [38]. Trend of global legal tech funding.

Compliance and Risk Management. This category addresses tasks such as automating financial audits, verifying regulatory obligations, and detecting potential liabilities within legal documents. In *automation of auditing*, Sifa et al. [39] develop a machine learning-driven recommender tool to align sections of financial statements with relevant legal statutes, reducing reliance on manual checks. For *GDPR compliance*, Cejas et al. [40] propose an NLP-based approach that compares phrasal-level representations of Data Processing Agreements (DPAs) against extracted "shall" requirements, improving baseline performance by about 20 percentage points. Meanwhile, Chakrabarti et al. [41] emphasize *risk assessment*, introducing "risk-o-meter," a paragraph-vector-based system that automatically flags high-risk sections in lengthy documents, promising greater efficiency in contract analysis and liability identification.

Document Analysis and Management. In this category, practitioners frequently rely on automated classification, contract analysis, and error detection systems to handle large document repositories. For *document classification*, Bambroo and Awasthi [42] propose an extended attention version of DistilBERT to accommodate the lengthier content typical of legal documents, while Chalkidis et al. [43] introduce a multilingual and multi-label dataset (MultiEURLEX) specifically designed for zero-shot cross-lingual transfer. Turning to *smart contract analysis*, Ahmed et al. [44] combine NLP methods with blockchain technology, demonstrating a prototype that generates code from legislative provisions and reports 96% accuracy in named entity recognition. Lastly, *error detection* is addressed by Bernsohn et al. [45], who devise an LLM-based pipeline to identify legal violations and link them to affected parties, underscoring how large models can reveal hidden inconsistencies in voluminous, unstructured case data.

Document Generation and Assistance. This category encompasses a range of systems designed to draft, summarize, or interpret legal texts more efficiently. For *Summarization*, Sheik and Nirmala [46] explore several deep neural architectures that compress legal documents into concise overviews, significantly reducing the burden on human reviewers. In the realm of *legal language interpretation*, Serediuk [47] describes thematic modeling and semantic analysis methods that help dissect dense clauses and clarify ambiguous passages, ultimately speeding up comprehension. Meanwhile, Semo et al. [48] introduce a new resource focused on *class action lawsuits*, specifically analyzing *authentic* complaints for legal judgment prediction. By working directly with unfiltered case filings rather than curated facts, their dataset and experiments reveal how automated language models can support attorneys' tasks in assessing or responding to large-scale civil suits.

Information Processing and Extraction. This category spans tasks such as structured data extraction, anonymization of sensitive content, and document retrieval. To exemplify *information extraction*, Bommarito et al. [49] introduce *LexNLP*, an open source Python package offering named entity recognition, text segmentation, and both unsupervised and supervised learning pipelines tailored to legal and regulatory text. Another study by Lison et al. [50] addresses *anonymization / text scrubbing*, reviewing existing approaches for mitigating disclosure risks while preserving data utility and outlining future directions in privacy-preserving natural language processing. Finally, Sansone and Sperlí [51] focus on *document retrieval*, surveying the state-of-the-art in Legal Information Retrieval systems and identifying open challenges such as efficiently mining large unstructured repositories for parallel statutes, related cases, or critical precedents.

Legal Decision Making and Dispute Resolution. This category focuses on predicting judgments, modeling legal reasoning, and facilitating dispute resolution. In the realm of *judge decision making and legal reasoning*, Wang et al. [52] introduce "LegalReasoner," a multi-stage framework that infuses domain knowledge into large language models (LLMs) for tasks such as multi-hop reasoning and case-law retrieval. Another study by Ma et al. [53] targets *judge decision making* in a real court setting, leveraging raw inputs (e.g., plaintiff's claims, court debate data) to create a multi-task architecture that learns factual logic for more accurate legal judgments. Meanwhile, Van Der Haegen [54] addresses the application of AI as a *dispute resolution mechanism*, discussing both the potential benefits of quicker, data-driven adjudication and the procedural challenges posed by delegating high-stakes decisions to algorithmic systems.

Legal Information Retrieval and Support. This category covers automated chatbots, domain-specific question answering, and language translation tools aimed at reducing barriers in legal comprehension and communication. For *chatbot* solutions, Kandula et al. [55] propose an AI-based legal assistance system that uses NLP and machine learning to retrieve and rank pertinent laws, reporting an accuracy of over 80%. In *question answering*, Louis et al. [56] focus on interpretability, employing a retrieval-augmented approach to generate detailed,

long-form answers for statutory queries—thereby bridging the gap between terse replies and more nuanced legal clarifications. Lastly, *translation* receives attention from Greńczuk et al. [57], who compare LLM-based translators (e.g., DeepL, Google Translate) for legal texts in less popular languages, underscoring the continuing necessity for accurate, unambiguous language support in international legal contexts.

Legal Research and Information Management. In this category, [58] illustrate *law systems divergence* by applying transformer-based models to cluster thematically related lawsuits in the Brazilian judicial system, capturing cross-regional variations in legal texts. Another study, [59], highlights *legal corpus curation* through constructing a large-scale jurisprudence database, even as part of their work intersects with anonymization (an aspect of *Information Processing and Extraction*). Finally, [60] introduce CUAD, a specialized contract-review dataset containing over 13,000 expert-labeled segments, underscoring the central role of curated resources in enhancing performance on domain-focused NLP tasks such as contract analysis and case retrieval.

6.4. Limitations

Although this study provides an overarching view of the current state of Legal NLP research and practitioner priorities, several constraints limit the scope and depth of its conclusions. Below, we structure these limitations according to the two key methodological pillars of the thesis: the *Systematic Literature Review* and the *Semi-Structured Interviews*.

6.4.1. Systematic Literature Review

Search Keywords. The SLR hinges on keyword-based queries, and while these terms were selected to capture a broad range of legal NLP topics, the inevitable risk is that certain relevant studies will remain undiscovered. For instance, older or highly specialized works may use nonstandard terminology or focus on narrower subdomains (e.g., e-Discovery within a single jurisdiction) that did not match our predefined search strings. Consequently, our final corpus might underrepresent niche or emergent research areas, thereby skewing the overall analysis toward more mainstream trends. To address this limitation in future work, researchers could employ more iterative or adaptive keyword refinement—possibly supplementing manual browsing with citation-based snowball sampling—to ensure that less prominent but potentially important papers are included.

Reliance on Abstract-Level Classification. We relied primarily on abstracts to categorize each publication. However, not all abstracts contain detailed methodological overviews or explicit statements of the authors' objectives and findings. In practice, some legal NLP tasks or datasets might only be described fully in the body of the text. Our automated pipeline thus risks misclassifying or overlooking subtler research contributions when authors either omit relevant keywords from their abstracts or dedicate minimal space to methodological

specifics. This limitation can especially affect categories where the distinction between sub-use-cases (e.g., compliance vs. dispute resolution) is nuanced. Improving our pipeline might entail additional steps such as introduction scanning or selected full-text parsing, but these approaches also increase computational overhead and complexity.

Training Data Limitations. Our fine-tuned classification models were trained on 233 labeled abstracts spanning multiple legal NLP categories. Despite encompassing numerous sub-use-cases, certain tasks inevitably remain undersampled. Rare or emerging topics (e.g., Law systems divergence or Smart Contract Analysis) may have only a handful of representative papers, limiting the model's capacity to recognize them in unlabeled data. Furthermore, if the labeled training set exhibits biases—such as overemphasizing certain jurisdictions or publication venues—the model might learn skewed decision boundaries. Consequently, our pipeline's accuracy can drop when applied to less common or region-specific research areas. More extensive and balanced training sets, perhaps derived from collaborative multi-institution labeling efforts, would help mitigate these issues.

6.4.2. Semi-Structured Interviews

Sample Size. We performed 12 interviews, an appropriate scale for explorative or pilot investigations but insufficient for robust quantitative generalizations. While these discussions yielded valuable insights, a larger cohort would likely uncover more diverse perspectives and identify additional or conflicting priorities. Moreover, the limited sample size inflates the influence of each participant's idiosyncratic experiences—particularly relevant if one interviewee works in a highly specialized domain (e.g., GDPR compliance in healthcare). Future expansions could combine larger participant pools with more systematic sampling to ensure the results better reflect the heterogeneity of legal practice.

Professional Diversity. The interviewed group consists of 10 attorneys, one law student, and one prosecutor. Crucially missing are roles like judges, paralegals, in-house legal engineers, compliance officers, and legal knowledge managers who may approach technology adoption and legal NLP use-cases quite differently. For example, a paralegal might prioritize document summarization differently from a litigator, or a policy-focused judge could value advanced argument mining over contract drafting tools. Hence, the use-case rankings gleaned from our interviews could underrepresent critical workflows or functional needs faced by these alternate roles. Future research can target a more balanced range of legal professionals to capture a fuller spectrum of technical requirements and readiness.

Scalability of Semi-Structured Interviews. Semi-structured interviews offer rich qualitative input but require intensive time and expertise to conduct, transcribe, and interpret. Expanding beyond a small participant set multiplies the burden of thematic coding and opens the door to inconsistencies in how findings are categorized. Additionally, researcher bias can seep in during open-ended questioning or in the selective integration of participant feedback. While

structured survey instruments might reach broader audiences with less effort, they cannot provide the same level of nuanced feedback. Balancing the depth of semi-structured dialogues with the need for larger, more heterogeneous samples remains a challenge for future projects aiming to map practitioner needs at scale.

7. Conclusion

7.1. Summary

This thesis set out to investigate emerging trends and priorities in the application of natural language processing (NLP) to the legal domain, with a specific focus on the DACH region. Building on a scalable and automated literature review pipeline, we analyzed 3,578 papers, of which 988 were identified as explicitly discussing at least one legal AI use-case. By examining the distribution of legal NLP tasks and their underlying techniques, we established a detailed view of academic directions in areas such as document classification, compliance automation, legal reasoning, and beyond. To complement these findings with practitioner insights, we conducted 12 semi-structured interviews with industry experts, primarily from DACH-based legal practices, thereby highlighting possible alignments and discrepancies between academic research and real-world needs.

RQ1: How can an automated, scalable pipeline effectively categorize academic literature into predefined legal AI use cases? We developed and fine-tuned large language models to classify relevant papers into a structured taxonomy of legal NLP use-cases. As demonstrated in our categorization results (Table 5.3), the pipeline achieved strong accuracy across multiple subcategories (e.g., compliance, document generation, legal decision making), benefiting from domain-specific data and systematic labeling. These findings confirm that combining domain-tuned language models with well-structured taxonomy definitions enables reproducible, large-scale analysis of legal research output. Such a pipeline also shows potential for continuous updates, allowing future work to track evolving directions in legal NLP without the overhead of manual scanning.

RQ2: How have academic trends concerning NLP use cases within the legal domain shifted over time, and which emerging areas have gained increased research interest? A temporal breakdown of the 988 identified papers revealed a pronounced uptick in legal NLP studies over the past five years, which we examined in detail in Chapter 6.2. Much of this escalation correlates with wider adoption of transformer-based architectures, alongside fresh domain adaptations (e.g., LegalBERT) and the development of comprehensive corpora. Within this expansion, select subcategories have demonstrated especially high growth rates, including *Legal Corpus Curation*, *Document Classification*, *Judge: Decision Making*, and *Information Extraction*. Together, these areas reflect a shift toward more sophisticated tasks requiring robust data infrastructures and complex reasoning—thus underscoring academia's drive to push legal NLP beyond basic text processing into multi-stage or multi-hop reasoning pipelines.

RQ3: Which legal AI use cases do practitioners identify as most relevant to their professional practice, and what factors influence these perceptions? Figure 5.12 aggregates the feedback from 12 industry experts, revealing a strong preference for *Document Generation and Assistance*—in particular, contract drafting and automated summarization. Many participants stressed the time savings and error reduction offered by text generation tools, while also recognizing the need for domain adaptation to ensure legal validity. *Legal Research and Information Management* followed closely, reflecting the ongoing necessity of navigating through massive law jurisdictions and updating knowledge in line with new legislation. *Document Analysis and Management* completed the upper tier of practitioner rankings, highlighting the value of advanced classification or contract analytics in day-to-day workflows. These preferences appear driven by a combination of immediate efficiency gains, interpretability requirements, and the trustworthiness of AI outputs in high-stakes environments.

7.2. Future Outlook

The findings in this thesis provide a clear snapshot of Legal NLP research and practitioner preferences in the DACH region, yet several developments could further expand, refine, and validate these insights. Building on the pipeline results and interview feedback, four potential directions for future work are outlined below.

1) Enlarging and Balancing the Dataset of Labeled Abstracts. Our fine-tuning pipeline relies on a dataset in which some use-cases, such as *GDPR Compliance* or *Anonymization*, are notably underrepresented. As a result, the model can struggle to detect these categories reliably, limiting its overall effectiveness. To address this, future work could focus on systematically expanding the corpus of labeled abstracts and titles, particularly in the least populated use-cases. One strategy would be to refine and iterate on keyword-based queries, tailoring search terms to niche legal subdomains—such as contract redaction under specific regulations or privacy-related tasks in healthcare law—to capture papers omitted by the current filtering. Another approach involves monitoring newly published works (e.g., from preprint servers or specialized conferences) to ensure the dataset remains current as novel tasks and applications emerge in the legal field. By populating these undersampled categories with additional representative samples, not only does the model's classification accuracy improve, but it also gains a broader view of domain-specific terminology and context. Over time, repeated iterations of data collection and labeling could further reinforce the pipeline's capacity to categorize new research accurately and keep pace with evolving trends in legal NLP.

2) Expanding Industry Ranking Exercises. The legal use-case rankings in this thesis emerged from a limited pool of DACH-based practitioners, most of whom had specific areas of expertise and organizational contexts. Although these interviews provided high-quality, in-depth insights, the sample size restricts the ability to generalize trends across diverse legal markets or practice settings. Future studies could extend this ranking process to

larger, more varied groups—both within and beyond the DACH region—to yield broader perspectives on emerging industry demands. In particular, including practitioners from multiple jurisdictions, firm sizes, and specialization areas (e.g., corporate law, intellectual property) would uncover whether widely cited use cases, such as advanced contract drafting and generation, retain top priority under different regulatory or economic conditions. Collecting and analyzing these expanded rankings on an ongoing basis—perhaps biannually—would also allow researchers to track shifting priorities as new tools enter the market, legal requirements evolve, and AI maturity increases. This iterative approach can ultimately help calibrate academic focus toward those tasks that align most closely with real-world challenges and practitioner readiness for advanced NLP solutions.

3) Maintaining and Evolving Use-Case Taxonomies. Although this thesis refines an existing taxonomy of legal NLP use-cases, new requirements and applications regularly surface, reflecting the rapid advancements in AI and shifts within legal practice. During this study, for instance, we identified three additional use-cases that were not part of the original classification from 2023. Such findings underscore the value of regularly updating the taxonomy—perhaps on an annual or biennial cycle—to encompass newly emergent tasks as well as sub-use-cases that gain traction. In doing so, the taxonomy remains an active tool for guiding research and practical development, rather than a static snapshot of legal AI possibilities.

In light of these perspectives, the research points to a continuing need for data expansion, broader practitioner engagement, and systematic updates to the legal AI taxonomy. By enlarging and balancing labeled abstracts, future developers can achieve more reliable classification of emerging tasks. Extending the ranking exercise to a larger, more diverse practitioner base would enable clearer insights into the evolving needs of legal professionals. Likewise, maintaining and evolving the use-case taxonomy will help capture newly emergent areas and ensure that research agendas keep pace with practical demands. Collectively, these directions underscore the dynamic nature of legal NLP and the ongoing dialogue required between academia and industry to foster meaningful, scalable innovation in the sector.

A. General Addenda

If there are several additions you want to add, but they do not fit into the thesis itself, they belong here.

A.1. Prompt Definitions

A.1.1. Use-Case Extraction Prompt

```
usecase_prompt = f"""
You are an advanced classifier focusing EXCLUSIVELY on text, language, or NLP applications
within the legal domain.
USE CASE CATEGORIES TO EXPLICITLY SEARCH:
{legal_use_case_text}

TASK:
- The classification is MULTI-LABEL: the paper can match multiple sub-categories.
- Use EXACT sub-category names from the known set of legal NLP use-cases only.
- Do NOT create or use new sub-categories.
- If the text appears to be a literature survey, workshop summary, a book or book chapter
  do NOT classify it.
- Only proceed with classification if the text is a research paper explicitly describing
  NLP/text-based methods in legal contexts.
- If there is NO explicit mention or clear description of a text-based or language-based
  or knowledge-based method in a legal context, output FALSE.
- Immediately exclude papers which discuss AI/automation in legal contexts **without
  explicit text/language-based methods** (e.g., general AI ethics, governance, or policy
      debates).

IMPORTANT NOTE:
- NLP in this context includes any form of text analytics, text mining, text classification,
  information extraction, summarization, or other language-based techniques specifically
  applied to legal documents or legal data.
- Even if the terms "NLP" or "natural language processing" are not used, you should treat
  references to analyzing, processing, comparing, extracting, curation or transforming text
  within a legal context as an NLP use-caseprovided the paper explicitly discusses
  these text-based methods.
- "AI assisting judges"  NLP unless it describes text analysis (e.g., extracting case facts,
  summarizing precedents).
- "Chatbots"  NLP unless they process legal text (e.g., parsing statutes, answering legal
  questions from text corpora).
- Exclude papers that only mention automation, algorithms, or AI without **text-based
    workflows**.
```

```
CRITICAL INSTRUCTIONS:
- ONLY classify use cases DIRECTLY MENTIONED in the text. ZERO tolerance for inference.
- **Exclude** classifications with confidence < 0.8 from the final output
- ONLY classify use cases and concerns DIRECTLY MENTIONED or clearly described in the text
  (e.g., mention of analyzing legal documents, extracting clauses, comparing legal text
      versions,
   summarizing case files, etc.).
- If NO clear mention of text-based or language-based analysis for a legal purpose,
  do NOT speculateoutput FALSE.
- ZERO tolerance for inference without textual proof.
- No classification without explicit text/language + legal connection.
- Immediately EXCLUDE papers which mention terms like "black box," "transparency," or "
    fairness"
  **without tying them to NLP techniques**.
- Immediately EXCLUDE papers which focus on societal/ethical implications of AI in law
  rather than NLP applications.
- **ZERO IMPLICATION RULE**: NEVER use phrases like "can be interpreted as," "implies," "
    suggests,"
  or "indirectly supports."
- **DIRECT EVIDENCE ONLY**: A sub-category is valid ONLY if the text **explicitly describes
    the task**
  (e.g., "detect missing clauses," "analyze blockchain contracts").
- **Reject** classifications where the reasoning relies on assumptions or extrapolation.

CLASSIFICATION CRITERIA (confidence scores):
- 1.0 : Verbatim, direct use-case description in the text.
- 0.80.99 : Clear, specific application discussion.
- 0.60.79 : Implied but substantive discussion of a known use-case.
- 0.40.59 : Vague or peripheral mention (not well-defined).
- 0.00.39 : No meaningful mention or connection.

CRITICAL FINAL CHECK:
- No classification if no mention of text analytics or language-based processing in a
  legal context.
- No speculation or invention of sub-categories.
- Maximum academic rigor and textual evidence.
- Preserve the exact sub-category names.

OUTPUT FORMAT (STRICT JSON):
A single string of comma-separated sub-categories (ONLY sub-categories with confidence >= 0.8).

If no valid classifications, output empty string.
Example: "Automation of Auditing, Risk Assessment"
If the paper does not discuss legal NLP use cases:
  Return None
"""
```

A.1.2. NLP Techniques Prompt

```
prompt = f"""
```

```
You are an advanced classifier focusing EXCLUSIVELY on identifying specific Natural Language
    Processing (NLP) techniques used in research papers.

NLP TECHNIQUE CATEGORIES TO EXPLICITLY SEARCH:
{nlp_technique_text}

TASK:
- The classification is MULTI-LABEL: the paper can use multiple NLP techniques.
- Use EXACT technique names from the known set of NLP techniques only.
- Do NOT create or use new technique names.
- Only identify techniques that are EXPLICITLY mentioned or clearly demonstrated in the paper.
- Identify ONLY techniques that appear in the provided categories list above.
- If a technique is not explicitly mentioned or described, do NOT include it.

IMPORTANT GUIDELINES:
- Identify ONLY techniques that have DIRECT EVIDENCE in the text.
- A technique is valid ONLY if the text explicitly mentions or describes its use.
- Even if the exact name of the technique isn't used, include it if the description clearly
    matches.
- Exclude techniques with confidence < 0.8 from the final output.
- Maintain the original technique names exactly as listed above.
- Do NOT speculate on techniques that might be implied but not explicitly described.

CONFIDENCE CRITERIA:
- 1.0: Explicit mention by name with implementation details
- 0.8-0.99: Clear description of the technique even if not named explicitly
- 0.6-0.79: Strong indication of the technique's use
- 0.4-0.59: Possible but unclear reference to the technique
- 0.0-0.39: No substantive evidence of the technique

CRITICAL FINAL CHECK:
- Only include techniques with high confidence (>= 0.8)
- No speculation or invention of techniques
- Use exact technique names
- Maximum academic rigor and textual evidence

OUTPUT FORMAT:
A single string of comma-separated technique names (ONLY techniques with confidence >= 0.8).
If no techniques are identified with sufficient confidence, output an empty string.
Example: "Named Entity Recognition, Text Classification, Word Embedding"
If no valid techniques are identified:
Return an empty string ""
"""
```

List of Figures

List of Tables

Bibliography

[1] IBM. *What is natural language processing?* IBM Think. URL: https://www.ibm.com/think/topics/natural-language-processing.

[2] Chair of Software Engineering for Business Information Systems (sebis), TUM. *Legal AI Use Case Radar*. 2024. URL: https://legal-ai-radar.de/.

[3] D. H. Hagos, R. Battle, and D. B. Rawat. *Recent Advances in Generative AI and Large Language Models: Current Status, Challenges, and Perspectives*. 2024. arXiv: 2407.14962 [cs.CL]. URL: https://arxiv.org/abs/2407.14962.

[4] InfoQ. *Large Language Models (LLMs) and Prompting*. URL: https://www.infoq.com/articles/large-language-models-llms-prompting/.

[5] V. B. Parthasarathy, A. Zafar, A. Khan, and A. Shahid. *The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities*. 2024. arXiv: 2408.13296 [cs.LG]. URL: https://arxiv.org/abs/2408.13296.

[6] R. Whalen. "Defining legal technology and its implications". en. In: *Int. J. Law Inf. Technol.* 30.1 (Apr. 2022), pp. 47–67.

[7] O. R. Goodenough. "Legal Technology 3.0". In: *HuffPost* (2015). URL: https://www.huffpost.com/entry/legal-technology-30_b_6603658.

[8] J. Webb. "Legal technology: The great disruption?" en. In: *SSRN Electron. J.* (2020).

[9] S. Meisenbacher, N. Machner, J. Vladika, and F. Matthes. *Legal AI Use Case Radar 2024 Report*. en. Tech. rep. Technical University of Munich, July 2024. URL: https://mediatum.ub.tum.de/1748412.

[10] M. Preis. "Application of Natural Language Processing in the Legal Domain: Identification and Categorization of Use Cases". https://wwwmatthes.in.tum.de/pages/14o7u6va8n8zd/Master-s-Thesis-Martina-Preis. Master's Thesis. Technical University of Munich, 2021.

[11] B. Thiess. *Automated Classification of Legal AI Use Cases: A Machine Learning Approach*. Bachelor's Thesis. https://wwwmatthes.in.tum.de/pages/aqqwjky2k48w/Bachelor-s-Thesis-Benedikt-Thiess. 2022.

[12] J. Vladika, S. Meisenbacher, M. Preis, A. Klymenko, and F. Matthes. *Towards A Structured Overview of Use Cases for Natural Language Processing in the Legal Domain: A German Perspective*. 2024. arXiv: 2404.18759 [cs.CL]. URL: https://arxiv.org/abs/2404.18759.

[13] V. Segarra-Faggioni. "Automatic Classification of Research Papers Using Machine Learning Approaches and Natural Language Processing". In: Springer Nature, Jan. 2021, pp. 80–87.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

[15] E. C. Garrido-Merchan, R. Gozalo-Brizuela, and S. Gonzalez-Carvajal. "Comparing BERT Against Traditional Machine Learning Models in Text Classification". In: *Journal of Computational and Cognitive Engineering* 2.4 (Apr. 2023), pp. 352–356. ISSN: 2810-9570. DOI: 10.47852/bonviewjcce3202838. URL: http://dx.doi.org/10.47852/bonviewJCCE3202838.

[16] J. Fields, K. Chovanec, and P. Madiraju. "A Survey of Text Classification With Transformers: How Wide? How Large? How Long? How Accurate? How Expensive? How Safe?" In: *IEEE Access* 12 (2024), pp. 6518–6531. DOI: 10.1109/ACCESS.2024.3349952.

[17] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha. *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications*. 2025. arXiv: 2402.07927 [cs.AI]. URL: https://arxiv.org/abs/2402.07927.

[18] Z. R. K. Rostam and G. Kertész. *Fine-Tuning Large Language Models for Scientific Text Classification: A Comparative Study*. 2024. arXiv: 2412.00098 [cs.CL]. URL: https://arxiv.org/abs/2412.00098.

[19] T. R. (UK). *Legal Solutions (UK) Blog*. Accessed: 2025-03-24. 2023. URL: https://legalsolutions.thomsonreuters.co.uk.

[20] LegalTechnology.com. *Legal Technology Trends and Insights*. Accessed: 2025-03-24. 2024. URL: https://legaltechnology.com.

[21] T. R. (Global). *Thomson Reuters Legal Insights*. Accessed: 2025-03-24. 2023. URL: https://legal.thomsonreuters.com.

[22] A. B. A. (ABA). *Law Practice Resources: Technology Today*. Accessed: 2025-03-24. 2023. URL: https://www.americanbar.org/groups/law_practice/resources/law-technology-today/.

[23] B. Kitchenham and S. Charters. "Guidelines for performing Systematic Literature Reviews in Software Engineering". In: 2 (Jan. 2007).

[24] H. Zhang, M. A. Babar, and P. Tell. "Identifying relevant studies in software engineering". In: *Information and Software Technology* 53.6 (2011). Special Section: Best papers from the APSEC, pp. 625–637. ISSN: 0950-5849. DOI: https://doi.org/10.1016/j.infsof.2010.12.010. URL: https://www.sciencedirect.com/science/article/pii/S0950584910002260.

[25] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. "Language Models are Unsupervised Multitask Learners". In: 2019. URL: https://api.semanticscholar.org/CorpusID:160025533.

[26] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL]. URL: https://arxiv.org/abs/2005.14165.

[27] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. 2023. arXiv: 2201.11903 [cs.CL]. URL: https://arxiv.org/abs/2201.11903.

[28] Z. Zhang, A. Zhang, M. Li, and A. Smola. *Automatic Chain of Thought Prompting in Large Language Models*. 2022. arXiv: 2210.03493 [cs.CL]. URL: https://arxiv.org/abs/2210.03493.

[29] P. Gill, K. Stewart, E. Treasure, and B. Chadwick. "Methods of data collection in qualitative research: interviews and focus groups". en. In: *Br. Dent. J.* 204.6 (Mar. 2008), pp. 291–295.

[30] H. Kallio, A.-M. Pietilä, M. Johnson, and M. Kangasniemi. "Systematic methodological review: developing a framework for a qualitative semi-structured interview guide". en. In: *J. Adv. Nurs.* 72.12 (Dec. 2016), pp. 2954–2965.

[31] J. Horton, R. Macve, and G. Struyven. "Qualitative research: Experiences in using semi-structured interviews". In: *The Real Life Guide to Accounting Research*. Elsevier, 2004, pp. 339–357.

[32] V. Venkatesh, University of Arkansas, S. A. Brown, H. Bala, University of Arizona, and Indiana University. "Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems". In: *MIS Q* 37.1 (Jan. 2013), pp. 21–54.

[33] P. L. H. Yu, J. Gu, and H. Xu. "Analysis of ranking data". en. In: *Wiley Interdiscip. Rev. Comput. Stat.* 11.6 (Nov. 2019).

[34] V. Braun and V. Clarke. "Using thematic analysis in psychology". en. In: *Qual. Res. Psychol.* 3.2 (Jan. 2006), pp. 77–101.

[35] Commission of the European Communities. *Commission Recommendation of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises (notified under document number C(2003) 1422)*. 2003. URL: https://eur-lex.europa.eu/eli/reco/2003/361/oj/eng.

[36] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. *LEGAL-BERT: The Muppets straight out of Law School*. 2020. arXiv: 2010.02559 [cs.CL]. URL: https://arxiv.org/abs/2010.02559.

[37] S. Yao, Q. Ke, Q. Wang, K. Li, and J. Hu. "Lawyer GPT: A legal large language model with enhanced domain knowledge and reasoning capabilities". In: *Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering*. Vol. 2023. Singapore Singapore: ACM, July 2024, pp. 108–112.

[38] R. Blyd. *Legalcomplex*. Accessed: 2025-04-12. 2025. URL: https://www.legalcomplex.com/.

[39] R. Sifa, A. Ladi, M. Pielka, R. Ramamurthy, L. Hillebrand, B. Kirsch, D. Biesner, R. Stenzel, T. Bell, M. Lübbering, U. Nütten, C. Bauckhage, U. Warning, B. Fürst, T. D. Khameneh, D. Thom, I. Huseynov, R. Kahlert, J. Schlums, H. Ismail, B. Kliem, and R. Loitz. "Towards Automated Auditing with Machine Learning". In: *Proceedings of the ACM Symposium on Document Engineering 2019*. DocEng '19. ACM, Sept. 2019. DOI: 10.1145/3342558.3345421. URL: http://dx.doi.org/10.1145/3342558.3345421.

[40] O. A. Cejas, M. I. Azeem, S. Abualhaija, and L. C. Briand. "NLP-Based Automated Compliance Checking of Data Processing Agreements Against GDPR". In: *IEEE Transactions on Software Engineering* 49.9 (2023), pp. 4282–4303. DOI: 10.1109/TSE.2023.3288901.

[41] D. Chakrabarti, N. Patodia, U. Bhattacharya, I. Mitra, S. Roy, J. Mandi, N. Roy, and P. Nandy. "Use of Artificial Intelligence to Analyse Risk in Legal Documents for a Better Decision Support". In: *TENCON 2018 - 2018 IEEE Region 10 Conference*. 2018, pp. 0683–0688. DOI: 10.1109/TENCON.2018.8650382.

[42] P. Bambroo and A. Awasthi. "LegalDB: Long DistilBERT for Legal Document Classification". In: *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*. 2021, pp. 1–4. DOI: 10.1109/ICAECT49130.2021.9392558.

[43] I. Chalkidis, M. Fergadiotis, and I. Androutsopoulos. *MultiEURLEX – A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer*. 2021. arXiv: 2109.00904 [cs.CL]. URL: https://arxiv.org/abs/2109.00904.

[44] S. U. Ahmed, A. Danish, N. Ahmad, and T. Ahmad. "Smart Contract Generation through NLP and Blockchain for Legal Documents". In: *Procedia Computer Science* 235 (2024), pp. 2529–2537. ISSN: 1877-0509. DOI: 10.1016/j.procs.2024.04.238. URL: http://dx.doi.org/10.1016/j.procs.2024.04.238.

[45] D. Bernsohn, G. Semo, Y. Vazana, G. Hayat, B. Hagag, J. Niklaus, R. Saha, and K. Truskovskyi. *LegalLens: Leveraging LLMs for Legal Violation Identification in Unstructured Text*. 2024. arXiv: 2402.04335 [cs.CL]. URL: https://arxiv.org/abs/2402.04335.

[46] R. Sheik and S. J. Nirmala. "Deep Learning Techniques for Legal Text Summarization". In: *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*. 2021, pp. 1–5. DOI: 10.1109/UPCON52273.2021.9667640.

[47] V. Serediuk. "Possibilities of using artificial intelligence and natural language processing to analyse legal norms and interpret them". In: *Social Legal Studios* 7.2 (Apr. 2024), pp. 191–200.

[48] G. Semo, D. Bernsohn, B. Hagag, G. Hayat, and J. Niklaus. "ClassActionPrediction: A Challenging Benchmark for Legal Judgment Prediction of Class Action Cases in the US". In: *Proceedings of the Natural Legal Language Processing Workshop 2022*. Ed. by N. Aletras, I. Chalkidis, L. Barrett, C. Goanță, and D. Preoțiuc-Pietro. Abu Dhabi, United Arab

Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 31–46. DOI: 10.18653/v1/2022.nllp-1.3. URL: https://aclanthology.org/2022.nllp-1.3/.

[49] M. J. Bommarito II, D. M. Katz, and E. M. Detterman. "LexNLP: Natural language processing and information extraction for legal and regulatory texts". In: *Research Handbook on Big Data Law*. Edward Elgar Publishing, May 2021, pp. 216–227.

[50] P. Lison, I. Pilán, D. Sanchez, M. Batet, and L. Øvrelid. "Anonymisation Models for Text Data: State of the art, Challenges and Future Directions". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by C. Zong, F. Xia, W. Li, and R. Navigli. Online: Association for Computational Linguistics, Aug. 2021, pp. 4188–4203. DOI: 10.18653/v1/2021.acl-long.323. URL: https://aclanthology.org/2021.acl-long.323/.

[51] C. Sansone and G. Sperlí. "Legal Information Retrieval systems: State-of-the-art and open issues". en. In: *Inf. Syst.* 106.101967 (May 2022), p. 101967.

[52] X. Wang, X. Zhang, V. Hoo, Z. Shao, and X. Zhang. "LegalReasoner: A Multi-Stage Framework for Legal Judgment Prediction via Large Language Models and Knowledge Integration". In: *IEEE Access* 12 (2024), pp. 166843–166854. DOI: 10.1109/ACCESS.2024.3496666.

[53] L. Ma, Y. Zhang, T. Wang, X. Liu, W. Ye, C. Sun, and S. Zhang. "Legal judgment prediction with multi-stage case representation learning in the real court setting". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Virtual Event Canada: ACM, July 2021.

[54] M. Van Der Haegen. "Quantitative legal prediction as a dispute resolution mechanism". en. In: *Eur. Rev. Priv. Law* 31.2/3 (Sept. 2023), pp. 299–328.

[55] A. R. Kandula, M. Tadiparthi, P. Yakkala, S. Pasupuleti, P. Pagolu, and S. M. Chandrika Potharlanka. "Design and Implementation of a Chatbot for Automated Legal Assistance using Natural Language Processing and Machine Learning". In: *2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS)*. 2023, pp. 1–6. DOI: 10.1109/AICERA/ICIS59538.2023.10420298.

[56] A. Louis, G. Van Dijck, and G. Spanakis. "Interpretable long-form Legal Question Answering with retrieval-augmented large language models". In: *Proc. Conf. AAAI Artif. Intell.* 38.20 (Mar. 2024), pp. 22266–22275.

[57] A. Greńczuk, I. Chomiak-Orsa, and K. Tryczyńska. "AI-supported translation tools for legal texts: A comparative analysis". en. In: *Procedia Comput. Sci.* 246 (2024), pp. 5545–5554.

[58] R. S. d. Oliveira and E. G. Sperandio Nascimento. "Analysing similarities between legal court documents using natural language processing approaches based on transformers". en. In: *PLoS One* 20.4 (Apr. 2025), e0320244.

[59] D. Garat and D. Wonsever. "Automatic curation of court documents: Anonymizing personal data". en. In: *Information (Basel)* 13.1 (Jan. 2022), p. 27.

[60] D. Hendrycks, C. Burns, A. Chen, and S. Ball. *CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review*. 2021. arXiv: 2103.06268 [cs.CL]. URL: https://arxiv.org/abs/2103.06268.