Team: LexLead.AI

By

Edwin Sanchez

Vaishnavi Kokadwar

Zhanye Luo

Supervisor: Dr. Nick Kadochnikov

A Capstone Project

Submitted to the University of Chicago in partial fulfillment
of the requirements for the degree of

Master of Science in Applied Data Science

Division of Physical Sciences

December 2025

Abstract

Roughly 55 million Americans face more than 260 million legal problems each year, and about 120 million go unresolved, an enduring “justice gap” driven by complex language, fragmented processes, and high costs. This capstone, in partnership with Lexlead.AI, develops a jurisdiction-aware legal language model focused on the Texas Family Code. The project combines targeted text preprocessing, fine-tuning of open-source models (LLaMA, Gemma, DeepSeek), and a retrieval-augmented generation (RAG) layer over eighty statute PDFs to improve factual accuracy and reduce hallucinations.

On a held-out set, fine-tuning raised average judge scores by +9.8% (Gemma-4B), +16.4% (Deepseek-7B) and +28.1% (LLaMA-3.1-8B) versus base; production gating requires RAGAS Faithfulness ≥ 0.85, strong Answer Relevancy, and an LLM-as-judge (1–4) average ≥ 3.0 with a lower “Score-1” error rate. Where fine-tuning gains fall short, the system defaults to RAG-only to ensure verifiable, citation-anchored answers. The expected outcome is a scalable, compliant, and auditable framework that delivers clear, jurisdiction-specific legal information while respecting ethical and regulatory boundaries.

Keywords: legal tech, legal language model, text processing, fine-tuning, retrieval-augmented generation, LLM-as-judge, texas family law, open-source LLMs

Executive Summary

Each year, millions of people confront legal issues without access to affordable guidance, leaving an estimated 120 million problems unresolved in the U.S. The gap persists because legal texts are hard to understand, procedures are fragmented, and help is expensive. Lexlead.AI targets this need with a jurisdiction-aware platform that provides clear, compliant legal information at scale.

This capstone builds a domain-specific legal LLM starting with the Texas Family Code. The system integrates curation and preprocessing of statutory text, fine-tuning of open models (LLaMA, Gemma, DeepSeek) for legal phrasing and definitions, and RAG over eighty statute PDFs that grounds every answer in section-anchored passages.

The evaluation framework applied an LLM-as-judge rubric (1–4 scale) for clarity and responsiveness alongside RAGAS for retrieval quality. On held-out data, fine-tuning improved average scores (+9.8% Gemma-4B; +16.4% DeepSeek-7B; +28.1% LLaMA-3.1-8B). Production thresholds required RAGAS Faithfulness ≥ 0.85, strong Answer Relevancy and Context Precision, and LLM-as-judge ≥ 3.0 with reduced Score-1 errors. As fine-tuning gains did not meet these criteria, the system adopted a RAG-only production approach, using fine-tuning insights solely for stylistic refinement while retaining retrieval as the source of truth.

The result is a transparent, auditable, and maintainable path to legal accessibility: rapid updates via re-indexing rather than retraining, standardized citations for trust, and guardrails that avoid the unauthorized practice of law. This Texas-first foundation is scalable to additional civil domains and jurisdictions, advancing practical, compliant AI support for the justice gap.

Table of Contents

List of Figures

List of Tables

Introduction

Annual data indicated that approximately 55 million Americans experienced over 260 million legal issues, yet a significant proportion—120 million—remained unresolved or resulted in perceived unfair outcomes. This disparity, termed the "justice gap," was fundamentally driven by several structural barriers: the proliferation of complex legal terminology, fragmented procedural workflows, and the prohibitive expense associated with conventional legal services. For numerous litigants, particularly those from low-income demographics and small business owners, the initial hurdle lay in comprehending the precise nature of their legal predicament, thereby precluding the execution of appropriate subsequent actions. Consequently, the lack of readily accessible guidance often led individuals to postpone intervention, abandon the pursuit of resolution, or make ill-informed choices that ultimately exacerbated their circumstances.

Despite the demonstrable magnitude of unmet legal requirements, the legal sector historically exhibited a deficit of scalable, technology-driven solutions capable of effectively closing this gap. The prevailing service models were characterized by reliance on manual workflows, resource-intensive individual consultations, and localized, jurisdiction-specific expertise. This operational framework inherently impeded the efficient and large-scale delivery of legal support. The resultant absence of accessible tools and automation left a considerable portion of the population without meaningful mechanisms for support. However, the subsequent emergence of Large Language Models (LLMs) presented a pivotal and timely opportunity to fundamentally transform the provision of legal assistance, thereby facilitating processes that were significantly faster, clearer, and more inclusive. In short, modern LLMs

provide a scalable approach to translating complex, jurisdiction-specific law into plain-language guidance at a low marginal cost.

Lexlead.AI positioned itself at the vanguard of this technological transformation. The organization's core objective was to systematically address the three principal impediments to justice accessibility: the inadequacy of unambiguous guidance, the complexity inherent in jurisdiction-specific legal frameworks, and the elevated cost of services. By strategically leveraging the capabilities of LLMs, Lexlead.AI targeted the $50 billion underserved legal market with the specific aim of resolving a substantial fraction of the estimated 120 million annual unmet legal needs within the U.S. This initiative was predicated on the development of an AI-native, compliant, and highly scalable platform.

Problem Statement

Millions of individuals and small businesses confront significant legal challenges but are often unable to access effective resources to resolve them. The confluence of high legal service costs, opaque legal terminology, and complex procedural requirements leads many to forgo seeking legal counsel entirely. This contributes to a profound "justice gap," resulting in over 120 million legal issues remaining unresolved annually in the United States, thereby exacerbating societal inequity.

However, existing general-purpose Large Language Models (LLMs) are demonstrably inadequate for legal contexts. They frequently suffer from hallucination, lack the requisite domain-specific precision, and fail to integrate jurisdictional differences, which are critical for legal accuracy and compliance. This exposes a crucial technological and domain-specific gap that must be bridged to enable effective AI-driven legal assistance.

Our capstone project, conducted in collaboration with Lexlead.AI, was designed to address this dual challenge: a pressing societal problem and a significant, scalable business opportunity. Our approach involves fusing advanced language modeling techniques with deep legal domain expertise to architect AI systems that demonstrably improve legal access, enhance interpretability, and build user trust. This initiative unlocks value within the $50 billion underserved legal market while simultaneously tackling real-world barriers to justice.

The project was dedicated solely to Domain-Specific Legal Language Model Development, with the overarching goal of constructing a specialized language model meticulously tailored to the legal domain. The core objective was to significantly enhance the accuracy, interpretability, and trust associated with AI-driven legal support tools.

Table 1. *SMART Framework Summary*

SMART	Plan (What we'll do)	How (Methods & Data)	Evaluation Criteria / Metrics	Notes
Specific	Build a domain-specific legal LLM to improve accuracy, interpretability, and accessibility.	Fine-tune open models (LLaMA, Gemma, DeepSeek) on curated legal QA; integrate a light RAG fallback over 80 Texas Family Law PDFs.	Primary: legal/factual accuracy; interpretability (plain language); jurisdictional alignment.	If fine-tuning underperforms, switch to RAG mode to produce answers.
Measurable	Make performance objectively trackable.	Use legal-specific benchmarks and an LLM-agent evaluator.	LLM-as-judge 1–4 (4 = correct, 1 = incorrect); reasoning quality; step-by-step logical consistency; answer faithfulness; RAGAS Faithfulness ≥ 0.85 (for RAG).	Report Outcomes+A1:E5; compare to general-model baselines.
Achievable	Keep scope practical with proven tools.	Public legal datasets; open-source LLMs; lightweight RAG; Hugging Face + PEFT/LoRA workflows.	Feasibility checkpoints pass (data readiness, training stability, eval pipeline functioning).	Resource-aware: fits within available compute and team bandwidth.
Relevant	Directly address the justice gap and align with legal-tech trends.	Focus on Texas Family Law to ensure jurisdiction-aware outputs and real user utility.	Stakeholder validation (clarity for lay users); compliance guardrails respected.	Supports a scalable, compliant product direction.
Time-bound	Deliver results in 20 weeks.	Weeks 1–10 (Spring): data collection, feasibility, architecture design. Weeks 11–20 (Fall): fine-tuning, RAG and iterative evaluations.	Milestone completion on schedule; endline metrics vs. baseline reported.	Final Presentation

Specific: The team developed a domain-specific legal LLM, fine-tuned on curated legal Question-Answering (QA) data and utilizing open-source models such as LLaMA, Gemma, and DeepSeek. The aim was to optimize legal accuracy, interpretability, and accessibility.

To ensure verifiable answers when fine-tuning alone proved insufficient, the team integrated a lightweight Retrieval-Augmented Generation (RAG) fallback that drew from eighty PDFs from The Texas Family Law. When the fine-tuned model did not yield reliable results, the system switched to RAG mode to produce answers.

Measurable: Model efficacy was rigorously assessed using legally specific benchmarks, with accuracy, factual consistency, and jurisdictional alignment prioritized. The team also developed an LLM-agent-based evaluation framework to measure complex metrics such as reasoning quality (on a 1–4 scale, where 4 indicated correctness and 1 incorrectness), step-by-step logical consistency, and answer faithfulness in response to sophisticated legal queries. RAG success criteria were defined as RAGAS Faithfulness ≥ 0.85.

Achievable: The project was grounded in practicality, leveraging publicly available legal datasets, established open-source LLMs (e.g., LLaMA, Gemma, DeepSeek), and lightweight RAG. The team utilized the established, state-of-the-art frameworks such as Hugging Face and PEFT to facilitate development.

Relevant: This research directly addressed the widely acknowledged justice gap and strategically aligned with the legal-tech industry's ongoing evolution toward AI-enabled legal services.

Time-bound: All project milestones were clearly delineated within a 20-week timeline. The initial 10 weeks (Spring quarter) were dedicated to data collection, feasibility validation, and the design of the core model architecture. The subsequent 10 weeks (Fall quarter) focused on fine-tuning the specialized model and building RAG framework and executing iterative evaluations to confirm its alignment with the stated development goals.

Analysis Goals

Project Goal

The capstone project was designed to investigate the efficacy of large language models (LLMs) within the legal domain. Specifically, the research aimed to identify methods for effectively applying LLMs to simplify complex legal information, thereby enhancing its understandability and accessibility for non-specialist users.

Track 2: Domain-Specific Legal LLM Development

This component focused on the construction of a domain-specific legal large language model (LLM), specifically engineered to address the fundamental challenges associated with legal accessibility, interpretability, and trust. By fine-tuning open-source LLMs using a meticulously curated corpus of legal data, the research successfully established the foundational groundwork for sustained innovation within the legal technology sector. The ultimate design objective was to develop a computational system that demonstrated deep alignment with legal reasoning methodologies, ensured high factual accuracy, and rigorously adhered to jurisdiction-specific legal requirements.

Key Objectives and Scope

The project's primary objectives were defined to ensure the development of a legally sound and technically advanced system. Specifically, the team curated and constructed a high-quality legal Question-Answering (QA) dataset, which was systematically organized by specific legal domains (e.g., housing, employment, immigration) and meticulously tailored to reflect crucial jurisdictional variations across relevant U.S. states and localities. Concurrently, the team fine-tuned open-source Large Language Models (LLMs), including LLaMA, Gemma, and

DeepSeek, employing advanced techniques such as Low-Rank Adaptation (LoRA), quantization, and Direct Preference Optimization (DPO). In parallel, the team architected a Retrieval-Augmented Generation (RAG) layer grounded in a curated corpus of eighty (80) Texas Family Law PDFs, pre-processed into versioned, jurisdiction. This technical process was essential for aligning the models' generative behavior with established legal reasoning methodologies and strict compliance norms. To rigorously validate the model's performance, the project necessitated the development of specialized evaluation metrics. These metrics were based on LLM-agent frameworks, measuring criteria such as chain-of-thought consistency, retrieval faithfulness, and legal reasoning steps, to comprehensively assess response accuracy, completeness, and alignment with jurisdiction-specific legal standards; RAG-specific measures included Context Relevance, and RAGAS Faithfulness. Furthermore, the team planned to implement auditing tools to guarantee explainability and compliance within regulated legal settings. The Initial Minimum Viable Product (MVP) was strategically focused on the Texas Family Code, providing a robust foundation for the subsequent expansion of the model to encompass other legal domains and jurisdictions over a projected timeline.

Background

Literature Review

Understanding UPL - Unauthorized Practice of Law

UPL refers to the provision of legal services by unlicensed individuals, those who are disbarred, or those lacking bar admission. Each state defines and regulates UPL independently. For example, in Washington State, statutes RCW 2.48.180 and RPC 5.5 restrict legal practice to licensed attorneys and prohibit non-lawyers from giving legal advice, representing clients, or

drafting legal documents. Certified Limited Practice Officers (LPOs) and Limited License Legal Technicians (LLLTs) have limited exceptions. Some jurisdictions, such as the District Court of Ohio, have permitted the use of authorized legal research engines like Westlaw and LexisNexis for research purposes.

Violations can result in injunctions, fines, criminal charges, and civil litigation, all of which are designed to protect consumers from unqualified or fraudulent legal services. Some AI startups centered around legal tech, like DoNotPay or Augrented, have been warned by the court against UPL, after which they pivoted their business model.

AI and UPL: Risks and Regulatory Challenges

While the integration of artificial intelligence (AI) presents significant benefits for legal professionals—primarily by enhancing efficiency and improving access to information—it simultaneously introduces critical concerns regarding the unauthorized practice of law (UPL). Specifically, AI tools must not provide personalized legal advice; interpreting the law or applying legal principles to specific case facts remains strictly reserved for licensed attorneys.

Furthermore, the use of unsupervised AI poses a consumer protection risk, as inaccurate or harmful advice could be dispensed, potentially leading to client detriment. Complicating the professional landscape is the issue of accountability and liability, as AI systems cannot be held responsible under existing professional conduct rules when errors occur. This uncertainty is amplified by the current regulatory environment, where courts and regulatory bodies are actively developing policies; for instance, while some jurisdictions prohibit AI-generated legal filings, others permit the use of AI for legal research databases.

Table 2. *Authorized vs. Unauthorized AI Legal Activities*

Type of Legal Question	Authorized for AI?	Notes
General legal information (statutes, concepts)	Yes	Must not apply the law to specific facts
Self-help resources and public forms	Yes	No customization or completion for the user
Summaries of law or legal news	Yes	No tailored advice
Hypothetical/educational questions	Yes	Must not be presented as legal advice
Personalized legal advice	No	Applying law to the user's facts is UPL
Drafting/completing legal documents	No	Unless supervised by a licensed attorney
Representation in legal proceedings	No	Only licensed attorneys may represent clients

The potential for Unauthorized Practice of Law (UPL) constituted a critical risk within the legal technology domain, as violations carried the potential for severe consequences, including substantial financial penalties, litigation, or the forced cessation of business operations. To effectively mitigate this regulatory exposure, it was deemed essential to align the developed AI system with UPL compliance guidelines through meticulous fine-tuning and instruction tuning procedures.

The model was therefore trained to adopt specific compliant behaviors: it was mandated to either refrain from generating responses to personalized legal inquiries or, alternatively, to present its output accompanied by clearly stated, mandatory disclaimers. To facilitate this compliant training, the development team curated a specialized set of example

question-answer pairs that explicitly demonstrated the required compliant behavior. These pairs served as targeted training data used to appropriately guide and shape the model's outputs.

GenAI Legal Solutions - Current Competitors

As legal professionals increasingly explore the use of Artificial Intelligence (AI) to enhance research, drafting, and analysis workflows, several specialized Generative AI (GenAI) platforms emerge to meet this demand. This analysis provides a summary evaluation of three leading solutions—Westlaw CoCounsel, Lexis+ AI, and Harvey AI—based on their core capabilities, data architectures, and reported model performance, referencing data from a recent Stanford Study.

Westlaw CoCounsel is characterized by its tight integration with the proprietary Westlaw legal database, leveraging a robust Retrieval-Augmented Generation (RAG) approach coupled with rigorous manual evaluations to ensure quality. However, the system exhibits a high hallucination rate (33%), achieving only 42% accuracy in responses when drawing from the Westlaw Precision knowledge base. Conversely, Lexis+ AI utilizes the Shepard's Knowledge Graph to provide enhanced citation tracking and relationship analysis, and notably demonstrates the highest accuracy (65%) among the evaluated systems, along with a lower hallucination rate (17%). Despite its superior performance metrics, Lexis+ AI explicitly acknowledges key limitations inherent to GenAI, such as deficiencies in emotional intelligence and the completeness of source materials. Finally, Harvey AI, operating in partnership with OpenAI, provides a distinct model that offers custom training for individual firms, enabling targeted contract analysis, research, and litigation support. Harvey AI distinguishes itself by offering multilingual capabilities and firm-specific customization, making it highly suitable for

both boutique and large law firms, although it lacks independent performance benchmarks comparable to those available for Westlaw and Lexis.

Table 3. *Comparison Table: GenAI Legal Tools*

Feature/Criteria	Westlaw CoCounsel	Lexis+ AI	Harvey AI
Primary Use	Legal Professionals	Legal Professionals	Law Firms, Solo Practitioners
GenAI Core	OpenAI GPT-4	Multimodel: GPT-4, Claude 2 and 3, Mistral 7B	GPT (OpenAI) + Legal + Custom Firm Data
Data Source	Westlaw proprietary legal database	LexisNexis proprietary + Shepard's KG	GPT base, legal docs, firm-specific data
Training	30K expert QA pairs, manual fine-tuning	Relationship-based modeling (KG nodes/edges)	Foundation model + tailored firm training
RAG Implementation	Yes	Yes	Yes
Document Features	Review, Summarization, Comparison	Summarization, Drafting, Citation	Contract Analysis, Summarization, Research
Litigation Support	Limited	No	Yes
Multi-lingual Support	No	No	Yes
Pricing	$330+ upfront	$200/month or $99/use	$500/yr (Standard), $1200/yr (Premium)

Hallucination Rate (Stanford Study)	33% (Westlaw Precision), 17% (Ask AI)	17%	Not publicly available
Accuracy (Stanford Study)	42% (Precision), 20% (Ask AI)	65%	Not publicly available
Responsiveness Issues	25–62% incomplete answers	18% incomplete	Not specified
Error Acknowledgement	Yes – factual, outdated, bias, emotion gap	Yes – hallucinations, bias, citation gaps	Not specified
Peer/SME Review	Yes – manual + SME + peer reviews	Not emphasized	Not specified

AI accountability or model interpretability

When people use an AI system for legal information, they need two things above all: to trust that the tool is safe, and to understand what it's telling them. Around the world, laws and policies are moving in the same direction to make that possible. In Europe, a new AI law outlines clear rules for higher-risk systems, including managing risk, maintaining accurate records, and informing users about the system's capabilities and limitations. In the United States, there isn't one big AI law, but regulators like the Federal Trade Commission are already enforcing basic truth-in-advertising: don't exaggerate what your AI can do, be honest about how it works, and back up your claims. Taken together, these efforts encourage builders to create AI that leaves a trail we can audit, allowing us to see where information originated, who verified it, and whether it aligns with the correct jurisdiction.

Beyond laws, there are practical playbooks that help teams run AI responsibly on a day-to-day basis. The NIST AI Risk Management Framework, for example, provides a simple, repeatable loop (Govern, Map, Measure, Manage) that enables organizations to identify risks, test what matters, and improve over time. An international standard called ISO/IEC 42001 goes a step further, laying out how to build a management system for AI (policies, roles, controls, and continuous improvement). And global principles from the OECD encourage practices such as transparency and traceability throughout the AI lifecycle. These aren't just buzzwords: in a legal tool, they translate into concrete features, such as searchable logs, versioned sources, and "time-aware" answers that match the law as it existed on the date the user is concerned with.

Just as important as safety is understanding. People don't want a black box; they want explanations in clear language. Research on interpretability argues that explanations should be systematic and testable, rather than merely post-hoc justifications. In practice, this means publishing simple "nutrition labels" for models (Model Cards) and datasets (Datasheets for Datasets) that clearly state their intended uses, known limitations, and how performance was measured. For legal information tools, we can take it a step further: design the system to demonstrate its effectiveness. With retrieval-augmented generation (RAG), every answer can point to the exact section of law it relied on, note the version and effective date, and flag the jurisdiction. Those receipts make answers easier to trust, easier to contest, and easier to fix if something changes.

Data

The development of Legal Large Language Models (LLMs) required datasets that contained general legal information without crossing the boundary into unauthorized legal advice. However, training such models on sensitive or unverified data could raise ethical and legal risks, particularly regarding the unauthorized practice of law. To mitigate these issues, it was crucial to use only datasets that provided general legal information, not personalized advice.

Developing a civil law-focused legal chatbot (built on LLaMA 2, Gemma, and DeepSeek integration) required a robust and reliable corpus of legal texts. As a starting point, the project selected the Texas statutes as the primary data source. Texas statutes were readily accessible in digital form and organized in a well-structured manner, making them ideal for training the initial version of the chatbot. This essay outlines the rationale behind choosing Texas law, described the focus on the Texas Family Code in the initial phase, provides the complete list of Texas statutory codes, presents a sample excerpt of statute text, and discusses potential data gaps or limitations in using this source.

Data Sources

Focus on the Texas Civil Code

Texas was selected as the starting jurisdiction for several reasons, primarily due to its accessibility and data structure. The Texas Legislature made its statutes freely available online in a centralized location, meaning there were no licensing barriers or fees for using the law text. In fact, the official website provided the full text of all Texas codes, current through recent

legislative sessions (as of the 88th Legislature in 2023). This ensured the data was up-to-date and could be obtained directly from an authoritative source.

Additionally, Texas statutes were published in a consistent, well-organized format, with laws codified by subject area (e.g., Family Code, Property Code), broken down into titles, chapters, and sections, all using a uniform numbering system. Such a structured arrangement was advantageous for machine parsing and for the language model to learn the hierarchy and context of provisions. The statutes came as complete documents (available in PDF or HTML) that included clear headings and section labels, which simplified preprocessing.

Moreover, Texas had a comprehensive set of civil law statutes covering a wide range of topics (family law, contracts, property, etc.), providing breadth for the chatbot's knowledge. By starting with one state's laws, the team maintained a consistent legal framework for the model to learn from, reducing complexity that would have arisen from mixing multiple jurisdictions initially. Overall, Texas offered a combination of easy access, public-domain data, and logical organization that made it an excellent foundation for a legal AI's training corpus.

Focus on Texas Family Law

Within the Texas statutory corpus, the Family Code was chosen as the focus of the chatbot's initial training phase. The Family Code governs civil matters, including marriage, divorce, child custody, and related family law issues, making it a highly relevant domain for a civil law-oriented assistant. By concentrating on a single code first, the team could develop the chatbot's ability to interpret and answer questions within a narrower legal context before expanding to other areas.

The Family Code is a self-contained body of law, consisting of 81 pdf documents, with its own definitions and structure, providing a manageable starting dataset for fine-tuning the LLaMA 2, Gemma, and DeepSeek models. In practical terms, this meant the chatbot was trained on the full text of the Texas Family Code, learning the language and structure of the statutes (titles, chapters, sections, etc.) specific to family law. This approach enabled the development team to assess the chatbot's performance in a specific domain (e.g., answering questions about divorce procedures or custody rules under Texas law) and identify and resolve any issues related to interpretation or accuracy. An excerpt from the Texas Marriage Relationship Code is provided within Appendix B.

Descriptive Analysis

Gaps and limitations of the current data

While using the Texas statutes as a training corpus offered many advantages, it was important to acknowledge certain data gaps or limitations inherent in that approach. One significant limitation was ambiguity or the lack of context in isolated provisions. Statutory language proved difficult to interpret in isolation, as many sections assumed context from related provisions or definitions elsewhere in the code. For example, a section might refer to terms defined in a different chapter. If the chatbot was asked about a single section without that context, it might have struggled to fully understand the intent. Training on the text alone meant the model was required to infer context, which could lead to ambiguous interpretations if not carefully handled. A second limitation involved updates and amendments. The static files downloaded from the legislature's site represented the law as of the last update (current through the 2023 sessions in this case), meaning there could be a time lag between when new

legislation was enacted and when the compiled statutes were updated online. This meant the chatbot's knowledge could become outdated until the data was refreshed, necessitating a plan for periodic data updates to ensure the model's information remained current.

Additionally, the dataset provided only the text of the laws themselves, lacking any annotations, commentary, or case law interpretations. This was a significant limitation because legal understanding often requires more than just the black-letter law; context, such as legislative history or judicial interpretations, could be crucial. The chatbot, having been trained only on statutory text, might not have captured nuances derived from case law or might have been unable to resolve ambiguities that courts had previously addressed. In other words, the model might have known what the statute said, but not how it was typically applied or interpreted in real scenarios.

Finally, the data presented challenges related to formatting inconsistencies and metadata noise. Although the codes were generally well-structured, some formatting quirks required management during the preprocessing stage. For instance, a few collections, like the "Insurance Code – Not Codified" and Vernon's Civil Statutes, used an older style of numbering (organized by article numbers instead of the section format). This meant the model would encounter a slightly different organization for those texts. Furthermore, the PDF files included pagination headers and footers (e.g., lines indicating the rendering date and page number) that were not part of the statute text. Such extraneous metadata needed to be stripped out during data preprocessing; otherwise, it could have introduced noise into the training corpus. It was essential that the parsing script cleanly removed page breaks, footers, and any formatting

characters so that the model could be trained on pure statutory text. Minor inconsistencies (like spacing or numbering formats) were anticipated and normalized to maintain data uniformity.

Additionally, a comprehensive setup for custom dataset preparation was required to make sure the documents were converted to a Question and Answer format that could support fine-tuning (described in Methodology section). Despite these challenges, the Texas statutes remained a strong starting point due to their comprehensive coverage and structured presentation. Awareness of these gaps ensured they were handled proactively during the construction and refinement of the civil law chatbot.

Methodology

Custom Dataset Preparation

The foundational corpus for this project consisted of the Texas Family Law statutes, which were obtained in the form of 81 PDF documents. A primary project objective was to fine-tune an open-source LLM on these Texas Family Law statutes; however, publicly available datasets lacked the required specificity for Texas family law. It was therefore necessary to generate a custom corpus of question-answer (QA) pairs that mirrored the project's specific business problem: a legal assistant chatbot specific to Texas Family law.

This generation process was designed to yield two distinct datasets: one for fine-tuning and one for evaluation. A core requirement for the fine-tuning data was that the questions reflected common public inquiries, and the corresponding answers were formulated in simple, straightforward language accessible to a layperson. To this end, two techniques were employed.

First, a manual QA generation process was undertaken. The team systematically analyzed the Family Code PDFs and manually authored QA pairs. This process prioritized

readability, relatability, and the exclusion of direct legal citations or complex cross-references. To enhance the model's robustness and its ability to recognize semantic equivalence, each question was paired with three to five distinct, correctly-phrased answers. While syntactically different, each answer variant contained all the necessary factual details. A total of 2209 QA pairs were generated manually, out of which 100 were set aside as the evaluation set.

Second, this manual effort was augmented by an LLM-powered generation process, which utilized the GPT-5 model to scale dataset creation. This model was developed using guidelines similar to those established for the manual process, with a focus on providing simplified and accessible answers. This procedure was systematically applied to all 81 documents comprising the corpus, with a target of five unique questions generated per document. For each of these questions, the model produced four to five semantically equivalent correct answers for the fine-tuning set. 2045 QA pairs were generated using this technique to ensure a balance between manual and automated generation.

Combining the two techniques provided a total of 4154 QA pairs for fine-tuning and 100 QA pairs for evaluation. One example of this QA pair is provided in Table 4.

Table 4. *Sample Question-Answer Dataset*

ID	*question*	*answer*	*source*	*mode of generation*
1	In Texas marriage law, what is the default policy about the validity of marriages, and when will	The policy is to preserve and uphold each marriage unless clear grounds exist to make it void or voidable.	fa.1.pdf	manual

	a marriage be treated as void or voidable?			

A distinct dataset was then generated specifically for the development of an LLM-based evaluator. For this purpose, a custom 4-point rating scale was introduced which represented the level of correctness of a candidate answer compared to the ground truth. Higher value of the rating meant a more complete answer, 4 being the most perfect and complete response.

Table 5. *Evaluation Rating Scale*

Rating	*Reasoning*
4	Perfectly correct and complete, faithful to the text
3	Largely correct but with minor omissions or weak phrasing; still answers the question
2	Factually correct yet non-responsive (or mixes correct with irrelevant details so the conclusion is unreliable).
1	Completely incorrect (clearly contradicts the PDF or invents rules).

The LLM was prompted to generate answers corresponding to each of these four quality ratings. This sub-dataset of graded responses was created specifically to validate the accuracy of the LLM-based evaluator, which is discussed in a later section of this report. This entire dual-methodology for custom dataset generation was designed to ensure that the resulting

fine-tuning and evaluation corpora were maximally aligned with the project's objectives and to ensure redundancy.

LLM-Based Evaluation

To assess the quality of responses from the fine-tuned open-source models, a custom LLM-based evaluator was designed and implemented. This "LLM-as-judge" methodology was chosen because conventional lexical-overlap metrics, such as BLEU and ROUGE, were deemed insufficient. Such metrics, which rely on n-gram comparison against a ground truth, cannot adequately validate the factual accuracy, legal soundness, or semantic equivalence of answers written in accessible, non-legal language.

The custom evaluator was designed to provide a multi-dimensional assessment tailored to the specific context of legal aid. The primary evaluation criteria included: (1) Legal and Factual Soundness (the accuracy of the response when compared to the source legal text), (2) Clarity and Accessibility (the ease with which a non-legal layperson could understand the response), and (3) Contextual Relevance (the degree to which the answer directly and fully addressed the user's query).

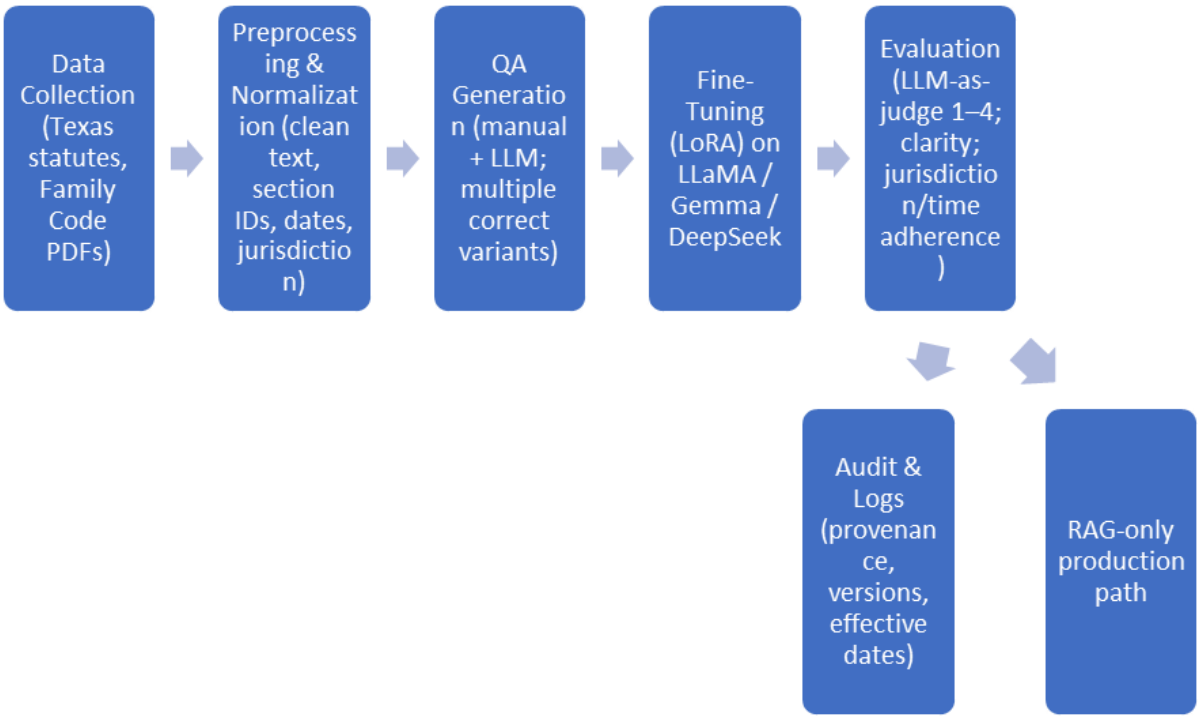
The "Gemini 2.5-flash" model was selected to serve as the evaluator. This was operationalized using an In-Context Learning (ICL) strategy. The prompt was engineered as a "few-shot" prompt, containing eight distinct examples to demonstrate the evaluation task. Each example included a question, the ground-truth answer, a candidate answer (the response to be judged), a rating (on a scale of 1-4 defined in Table 4), and a detailed rationale explaining why the candidate answer received that specific rating.

The inclusion of an explicit rationale in each example was a critical component, designed to anchor the evaluator's judgments in the defined quality criteria and guide their reasoning process. The evaluator was comprehenisvely tested using the custom built dataset, as referenced in the previous section, by generating a predicted rating and then comparing it with the true rating. During inference, the evaluator was provided with the question, the ground truth, and the candidate answer generated by the model under test. It was then instructed to output both a numerical rating (1-4) and a written rationale for its decision, mirroring the structure of the in-context examples.

Fine-Tuning

For this research project, the core technical approach involved the fine-tuning of open-source Large Language Models (LLMs). Specifically, a tripartite selection of models was utilized: Gemma (Google's open-source variant of the Gemini architecture), Meta's LLaMA, and DeepSeek. This selection was based on a comprehensive assessment of several critical factors, including licensing costs, computational resource requirements (both software and hardware), flexibility in fine-tuning, and overall baseline performance in natural language processing tasks.

Figure 1. *Fine-tuning process flow*



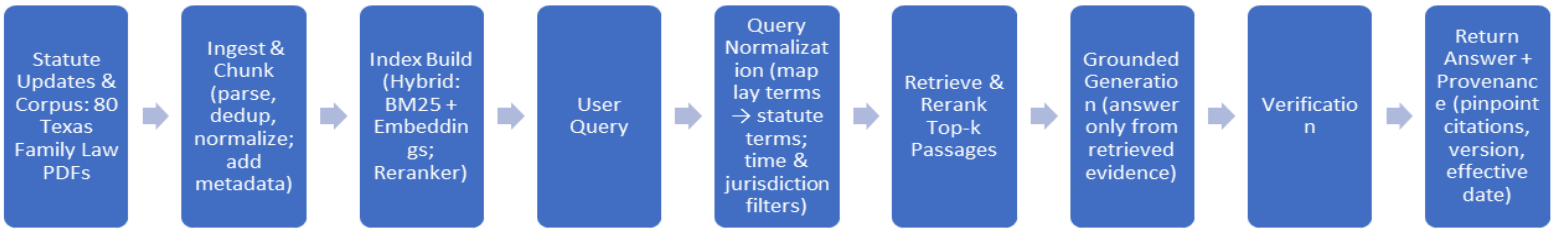
The fine-tuning procedure employed the Low-Rank Adaptation (LoRA) technique. This method was specifically chosen to customize the generalized LLMs for the specialized requirements of the legal domain. LoRA offered a highly optimized alternative to full model pre-training by selectively modifying only a small fraction of the model's parameters, thereby significantly reducing the computational overhead and resource intensity while achieving the desired use-case-specific adaptation.

Fine-Tuning Contingency Policy

If future fine-tuning trials fail to demonstrate statistically significant gains over the RAG baseline (for example, +80% improvement in the average 1–4 judge rating), the system remains RAG-only.

Retrieval-Augmented Generation (RAG)

Figure 2. *RAG Process Flow*



The production system adopts a RAG-only architecture to guarantee verifiable, jurisdiction-aware answers. The retrieval corpus comprises eighty (80) Texas Family Law PDFs that are pre-processed into section-anchored passages with normalized headings, deduplication, and rich metadata (including jurisdiction, chapter/section, effective date, and document version).

To preserve auditability and compliance, each response is accompanied by full provenance logs, including source identifier, version, timestamp, and jurisdiction filters applied at query time. Query normalization aligns layperson language with statutory terminology without altering intent, and temporal controls prefer passages that match the effective period implied by the question. These controls collectively reduce hallucinations, enforce jurisdictional fidelity, and enable downstream review. RAG performance is validated with RAGAS.

RAG is well-suited to this legal setting because it constrains answers to evidence retrieved from the eighty (80) Texas Family Law PDFs, ensuring each legal proposition is verifiable, jurisdiction-specific, and citable. By grounding generation in section-anchored passages and requiring standardized, pinpoint citations, RAG materially reduces hallucination

risk and keeps outputs aligned with the Family Code's language and scope. It is also operationally resilient: when statutes or guidance change, the system is updated by re-indexing the corpus rather than retraining a model, minimizing downtime and compliance exposure.

RAG also enables measurable quality control that maps cleanly to our existing evaluation stack. Retrieval quality and answer grounding are quantified with RAGAS (Faithfulness, Answer Relevancy, Context Precision/Recall), while our LLM-as-judge rubric assesses legal, factual soundness, clarity for non-lawyers, and responsiveness.

In practice, this combination of evidential grounding, maintainability, and auditable metrics makes RAG a robust default for producing reliable, compliant legal answers.

Findings

On the held-out evaluation set, fine tuning produced consistent gains over base models. Gemma 4B rose from 2.25 to 2.47 (+0.22; +9.8%), LLaMA 3.1 8B improved from 1.92 to 2.46 (+0.54; +28.1%), and Deepseek 7B increased from 1.52 to 1.77 (+0.25; +16.4%). However, DeepSeek's fine-tuned performance remains limited because the model was trained on a primarily Chinese corpus and its architecture is more prone to hallucination, which is especially disadvantageous for a legal QA task requiring strict factual grounding. Distributionally, LLaMA's Score 4 answers increased from 2 to 28, while Score 1 errors dropped from 30 to 23 (−23.3%), evidence that domain tuning reduced outright mistakes while pushing more responses into the fully correct and faithful band. For Gemma, Score 1 decreased from 20 to 8 (−60%) and Score 3 increased from 25 to 37 (+48%), indicating that fine tuning primarily converted low trust or partially off target replies into broadly correct and usable ones. Deepseek 7B showed a similar pattern, with Score 1 responses falling from 61 to 44(−50%), roughly halving the number of

outright errors. (See Figure 3 for details.) Similarly, the RAG model implemented on top of LLAMA-3.1-8B model gave an improvement of +80% for the average ratings, wherein the Score 1 responses falling from 30 to 1(−96%).

Figure 3. *Fine-tuning vs Base Model Results*

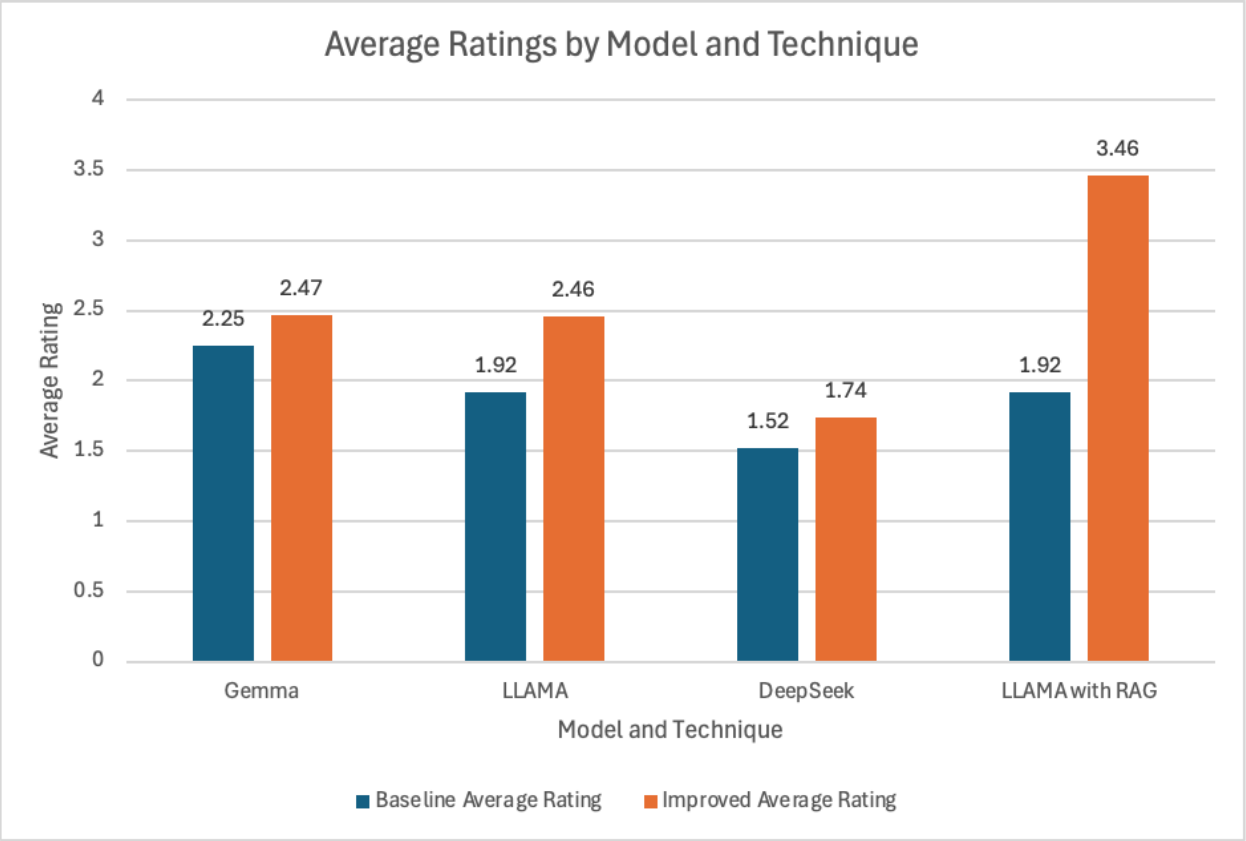
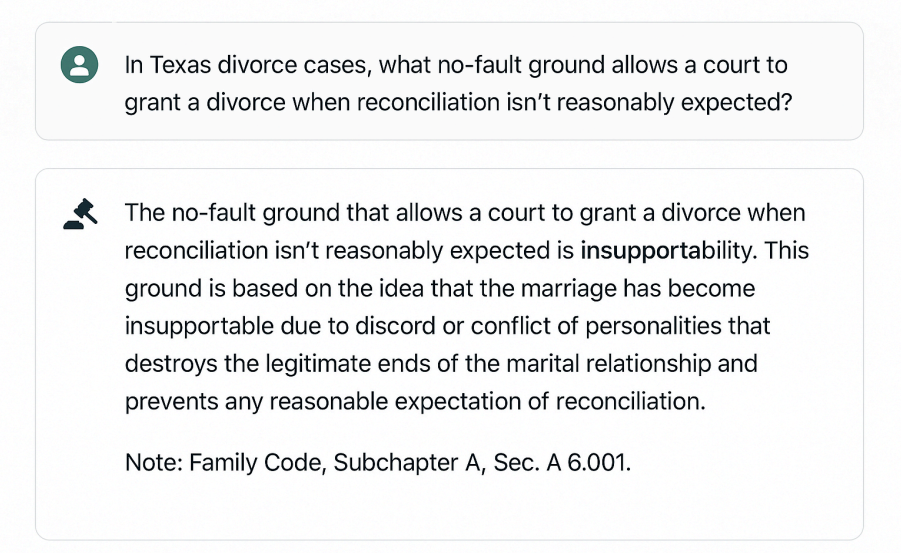


Figure 4. *Sample Retrieval from RAG*



Fine-tuning vs RAG

The system design required balancing the trade-offs between fine-tuning and retrieval-augmented generation (RAG). Fine-tuning primarily enhanced stylistic quality, tone, and overall response structure, while also providing deeper contextual understanding that supported greater consistency and faster response times. In contrast, RAG was optimized for high-precision information retrieval and offered greater flexibility when handling dynamic or frequently changing legal content. RAG further enabled transparent sourcing by grounding responses directly in statute-anchored passages. These contrasting strengths shaped the decision to rely on RAG for authoritative content while using fine-tuning selectively to refine clarity and presentation.

Contingency Plan Trigger and Path Forward

According to our contingency policy, fine-tuning must yield an improvement of at least 80% over the base to remain in scope for production. The observed lifts +9.8% (Gemma-4B), +16.4% (Deepseek-7B) and +28.1% (LLaMA-3.1-8B) are below the 80% threshold. Accordingly,

given limited computational resources and our advisor's recommendation, the team moved forward with a RAG-only production path. The FT results will continue to inform prompts and style constraints; however, generation in production will be strictly retrieval-grounded, with provenance.

Discussion

Interpretability vs. Legal Completeness

Fine-tuning pushed more answers into "usable" territory, largely by simplifying phrasing. However, simpler phrasing can omit legal qualifiers (e.g., unless, except, only if, effective-date limits), which reduces legal completeness even when the answer reads clearly. RAG, in contrast, gave answers which were more complexly phrased since they were referenced directly from the source texts. Although this gave more correct answers, they were not friendly for non-legal reader.

Risk Management & Compliance Testing

The models were designed to be helpful while avoiding the provision of inappropriate or unauthorized legal guidance. When a question requested personal legal advice, the system did not imitate the role of an attorney. Instead, the model issued a clear disclaimer and supplied general statutory information, along with resources for locating licensed legal counsel. This approach maintained user safety and ensured compliance with legal and ethical boundaries. Each answer "showed its receipts." Responses included the exact document name, relevant section number, the version of the document consulted, and the effective date of the law. When questions referred to past events, the system prioritized sourcing text that was valid at the

relevant time rather than relying solely on current law. This time-sensitive methodology enhanced the accuracy and trustworthiness of the outputs.

Before any component was deployed, it was required to pass a series of automatic validation checks. First, faithfulness: the answer had to align with the cited source text rather than rely on unsupported inference. Next, relevancy: the response needed to directly address the user's query. Additional checks ensured that the retrieved passages were correct and complete, avoiding irrelevant or incomplete excerpts. Finally, evaluation using a judge model needed to demonstrate strong average performance and a reduced rate of incorrect answers compared to earlier iterations.

Conclusion

The project addressed a real and measurable justice gap. Approximately 55 million Americans faced an estimated 260 million legal issues each year, with roughly 120 million remaining unresolved. The research investigated whether domain-specialized LLMs could clarify, improve accessibility, and enhance the safety of interacting with Texas family law. The team constructed a purpose-built QA corpus, fine-tuned open-source models using LoRA, and evaluated them with an LLM-as-judge rubric alongside retrieval metrics aligned with a planned production RAG architecture built over 80 Texas Family Law PDFs.

Empirically, fine-tuning improved answer quality but did not meet the production threshold. Gemma-4B improved from 2.25 to 2.47 (+9.8%), DeepSeek-7B from 1.52 to 1.77 (+16.4%), and LLaMA-3.1-8B from 1.92 to 2.46 (+28.1%). These gains demonstrated that targeted QA data and LoRA fine-tuning enhanced legal phrasing and common definitional accuracy; however, results remained below the ≥80% contingency threshold established for

deploying fine-tuning as the primary method. Consistent with the development plan and advisor recommendations—and taking into account compute constraints, retraining overhead, residual hallucinations, and jurisdiction drift—the team determined that a RAG-only approach would be used for production, with fine-tuning insights retained for shaping prompt style.

Technical implications: The results indicated that, under modest compute budgets, fine-tuning alone was insufficient to guarantee statute-faithful, jurisdiction-aware responses at production-level reliability. The findings also highlighted both the utility and the limitations of LLM-as-judge scoring: while highly efficient and consistent for rapid iteration, it exhibited constraints when evaluating edge-case legal reasoning. Future research should prioritize optimization of retrieval quality relative to generator constraints (e.g., brevity limits, templated outputs) to convert “mostly correct” answers into fully accurate, citation-aligned responses.

Practical deployment implications: A RAG-based architecture proved better aligned with legal-compliance requirements. Every claim could be grounded in section-anchored statutory passages, enabling formal audits and allowing updates to be executed through re-indexing rather than model retraining. This approach reduced operational cost, shortened update cycles following legislative changes, and provided transparent release criteria (e.g., RAGAS Faithfulness ≥ 0.85) before responses were delivered to end users.

Overall, the project established a defensible pathway toward reliable and auditable legal assistance: RAG served as the production backbone, while fine-tuning functioned as a lightweight enhancer for clarity and tone rather than as the primary source of truth. This structure balanced scientific rigor with practical deployability, advancing progress toward measurable impact on the justice gap.

Next Steps

For future work and scope of improvement, the project will focus on enhancing the robustness and reliability of the legal question-answering system. Key improvements include implementing an agentic RAG workflow orchestrated via LangGraph. This workflow will feature stateful agents, tool usage, and retries to ensure answers pass a verification loop before being returned to the user. It will start with a single agent and then introduce a reviewer sub-agent to specifically check citations and effective dates for enhanced accuracy. The knowledge foundation will be significantly enhanced by layering a knowledge graph index on top of the 80 PDFs. Utilizing GraphRAG, the system will extract entities (terms, parties, sections) and community summaries from the Texas Family Code, allowing cross-references and definitions to be resolved more effectively than with a vector-only search. Questions will be dynamically routed, sending "global" questions to the graph and "local" questions to document passages. Finally, the system's quality assurance will be production-grade, gating deployments with a rigorous RAGAS evaluation suite (targeting Faithfulness > 0.85, plus Answer Relevance and Context Precision/Recall). Beyond improvements in the Family Code domain, the project is designed for scalability, allowing for expansion to other sections of the legal domain and application to different jurisdictions within the U.S.

References

Above the Law. (2024, January). *Lexis+ AI and the power of good data.* https://abovethelaw.com/2024/01/lexis-ai-and-the-power-of-good-data/

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Amodei, D. (2022). Training a helpful and harmless assistant with RLHF. *Anthropic.* https://www.anthropic.com/research/hh-rlhf

By Design Law. (2024, July 1). *AI for legal services: Is it unauthorized practice of law (UPL)?* https://www.bydesignlaw.com/ai-for-legal-services-is-it-unauthorized-practice-of-law-upl

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient fine-tuning of quantized LLMs. *arXiv preprint arXiv:2305.14314.* https://arxiv.org/abs/2305.14314

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv.* https://arxiv.org/abs/1702.08608

Engstrom, D. F., & Engstrom, N. F. (2024, November 22). *We should focus on—and invest in—AI that serves people without lawyers.* ABA Journal. https://www.abajournal.com/voice/article/we-should-focus-onand-invest-inai-that-serves-people-without-lawyers

European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* Official Journal of the European Union. EUR-Lex.

Federal Trade Commission. (2024, September 25). *FTC announces crackdown on deceptive AI claims and schemes (Operation AI Comply).* Federal Trade Commission. https://www.ftc.gov/

Gebru, T., Morgenstern, J., Vecchione, B., et al. (2018/2021). Datasheets for datasets. *arXiv / Communications of the ACM.* https://arxiv.org/abs/1803.09010

Google Cloud. (n.d.). *Vertex AI documentation.* https://cloud.google.com/vertex-ai

Harvey AI. (n.d.). *Blog.* https://www.harvey.ai/blog

HiiL. (2021). *Justice needs and satisfaction in the United States of America.* https://iaals.du.edu/sites/default/files/documents/publications/justice-needs-and-satisfaction-us.pdf

Illinois Supreme Court. (2024, December 18). *Illinois Supreme Court announces policy on artificial intelligence.* https://www.lawnext.com/wp-content/uploads/2024/12/Illinois-Supreme-Court-AI-Policy.pdf

ISO/IEC. (2023). *ISO/IEC 42001: Artificial intelligence—Management system.* ISO.

Justia. (n.d.). *U.S. law and legal questions platform.* https://www.justia.com

LexisNexis. (2023, October). *LexisNexis announces launch of Lexis+ AI commercial preview.* https://www.lexisnexis.com/community/pressroom/b/news/posts/lexisnexis-announces-launch-of-lexis-ai-commercial-preview-most-comprehensive-global-legal-generative-ai-platform

LexisNexis. (2024, July). *LexisNexis enhances Lexis+ AI with new features, AI models, and graph technology.*

https://www.lexisnexis.com/community/pressroom/b/news/posts/lexisnexis-enhances-lexis-ai-with-new-features-ai-models-and-graph-technology-to-further-drive-high-quality-trusted-answers-for-legal-professionals

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems (NeurIPS)*. https://papers.nips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Mitchell, M., Wu, S., Zaldivar, A., et al. (2019). Model cards for model reporting. *arXiv / ACM FAT**. https://arxiv.org/abs/1810.03993

Morgan, R. (2024, December). *Can robot lawyers close the access to justice gap?* Colorado Lawyer. https://cl.cobar.org/features/can-robot-lawyers-close-the-access-to-justice-gap/

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST.

OECD. (2024). *OECD AI Principles (updated)*. OECD.

OpenAI. (2023). *GPT-4 technical report.* https://openai.com/research/gpt-4

Responsible AI in Legal Services (RAILS). (n.d.). *Responsible AI in legal services.* https://rails.legal/resources/resource-ai-orders/

Superior Court of Los Angeles County. (2024, January 24). *Court launches new partnership with Stanford Law School.* https://www.lacourt.org/newsmedia/uploads/142024124818124NR-01-24-2024-CourtLaunchesNewPartnershipwithStanfordLaw.pdf

Texas Legislature Online. (n.d.). *Texas statutes and codes.* https://statutes.capitol.texas.gov

Thomson Reuters. (2024, October 23). *Legal AI benchmarking: CoCounsel.* https://www.thomsonreuters.com/en-us/posts/innovation/legal-ai-benchmarking-cocounsel/

Thomson Reuters. (2024, October 23). *Legal AI benchmarking: CoCounsel – Addendum A.* https://www.thomsonreuters.com/en-us/posts/wp-content/uploads/sites/20/2024/10/LegalAIBenchmarkingAddendumA.pdf

Thomson Reuters. (2024, October 15). *Unlocking the full potential of professional-grade GenAI for your work.* https://www.thomsonreuters.com/en-us/posts/innovation/unlocking-the-full-potential-of-professional-grade-genai-for-your-work/

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... & Scialom, T. (2023). LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288.* https://arxiv.org/abs/2307.09288

Tam, D., Ju, D., Lin, Z., Xie, S., & Ji, Y. (2022). Evaluating factual consistency in summarization with scalable human review. *arXiv preprint arXiv:2211.08412*. https://arxiv.org/abs/2211.08412

United States District Court for the Southern District of Ohio. (2023, December 18). *Standing civil order.* https://www.ohsd.uscourts.gov/sites/ohsd/files/MJN%20Standing%20Civil%20Order%20eff.%2012.18.23.pdf

United States District Court for the Western District of North Carolina. (2024, June 27). *Standing order: In re use of artificial intelligence.* https://www.ncwd.uscourts.gov/news/standing-order-re-use-artificial-intelligence

Zhou, X., Wang, S., Dong, X., et al. (2023). AlignScore: Evaluating factual consistency with alignment function. *arXiv preprint arXiv:2305.16739*. https://arxiv.org/abs/2305.16739

Appendix

Appendix A: Available Texas Statutory Codes (Complete List)

For reference, the Texas statutes download page provides the following complete list of codes and related compilations available for downloadstatutes.capitol.texas.gov. These encompass the Texas Constitution, all codified state laws by subject, as well as certain uncodified laws.

Table 6. *Texas Family Law Codes*

Code/Compilation
The Texas Constitution
Agriculture Code
Alcoholic Beverage Code
Auxiliary Water Laws (a compilation of water-related statutes not included in the Water Code)
Business and Commerce Code
Business Organizations Code
Civil Practice and Remedies Code
Code of Criminal Procedure
Education Code
Election Code
Estates Code
Family Code
Finance Code

Government Code
Health and Safety Code
Human Resources Code
Insurance Code
Insurance Code – Not Codified (portions of Texas insurance law not yet integrated into the codified Insurance Code)
Labor Code
Local Government Code
Natural Resources Code
Occupations Code
Parks and Wildlife Code
Penal Code
Property Code
Special District Local Laws Code
Tax Code

Transportation Code
Utilities Code
Vernon's Civil Statutes (legacy laws that have not been incorporated into modern codes)
Water Code

Each of the above is a separate set of statutes focusing on a particular subject matter. They can be downloaded as PDF files from the Texas Legislature's website, ensuring we have an official copy of the text for training data. In the initial phase we are focusing on the Family Code, but this list illustrates the broad range of legal topics available for future expansion.

Appendix B: Sample Excerpt from the Texas Family Code

To understand the format and content of the statutes, consider a brief excerpt from the Texas Family Code. The codes are organized into hierarchical levels (Title, Subtitle, Chapter, etc.), followed by individual sections that contain the legal provisions. For example, the beginning of the Family Code includes structural headings and a section with definitions:

TITLE 1. THE MARRIAGE RELATIONSHIP

SUBTITLE A. MARRIAGE

CHAPTER 1. GENERAL PROVISIONS

SUBCHAPTER A. DEFINITIONS

Sec. 1.001. APPLICABILITY OF DEFINITIONS.

(a) The definitions in this subchapter apply to this title.

(b) Except as provided by this subchapter, the definitions in Chapter 101 apply to terms used in this title.

(c) If, in another part of this title, a term defined by this subchapter has a meaning different from the meaning provided by this subchapter, the meaning of that other provision prevails.

This snippet illustrates how statutes are formatted, with heading lines indicating the Title, Subtitle, Chapter, and Subchapter, which help contextualize the provisions. Then each section is labeled (here, Sec. 1.001) with a descriptive heading ("APPLICABILITY OF DEFINITIONS") followed by the substantive text of the law. Subsections (a), (b), (c) etc., break down the provision into smaller parts when needed. In the official publications, each section is typically followed by a citation of the act of the legislature that added or amended it (for example, a line like "Added by Acts 1997, 75th Leg., ch. 7, Sec. 1, eff. April 17, 1997" appears after Sec. 1.001 in the full text). These historical notes are included in the downloadable text and can serve as metadata, although they are not part of the law's current substantive content. This structured layout is consistent across the Texas codes, which makes it easier to parse and feed into the chatbot's training pipeline. The language itself is formal statutory language, which the model will learn to interpret and understand. By training on such text, the chatbot will become familiar with phrases like "except as provided by…", "shall mean…", or cross-references such as the one above, where Chapter 101's definitions are incorporated by referencestatutes.capitol.texas.gov. Overall, the Family Code data (and Texas codes in general)

are presented in a readily digestible format that clearly delineates each legal provision and its place within the broader structure of the code.

Appendix C: List of Abbreviations & Acronyms

AI — *Artificial Intelligence*: Computer systems performing tasks requiring human intelligence

BM25 — *Best Match 25*: Keyword-based (lexical) document ranking method

DeepSeek — *DeepSeek LLM*: Open-source large language model by DeepSeek

DPO — *Direct Preference Optimization*: Preference-based fine-tuning technique

EU AI Act — *Regulation (EU) 2024/1689*: EU rules for trustworthy, accountable AI

FTC — *Federal Trade Commission*: U.S. regulator enforcing truthful AI claims

FT — *Fine-Tuning*: Adapting a base model to a domain/task

Gemma — *Gemma LLM*: Google's open-source family of LLMs

GPT — *Generative Pre-trained Transformer*: Family of transformer-based language models

GraphRAG — *Graph-based RAG*: RAG using a knowledge-graph index

ICL — *In-Context Learning*: Teaching a model via examples in the prompt

ISO/IEC 42001 — *AI Management System Standard*: Certifiable standard for managing AI risks

KG — *Knowledge Graph*: Nodes/edges capturing entities and relations

LangGraph — *LangGraph framework*: Toolkit for agentic, graph-style LLM workflows

LLLT — *Limited License Legal Technician*: Washington State's limited legal license role

LLaMA — *Large Language Model Meta AI*: Meta's open-source LLM family

LLM — *Large Language Model*: Large neural model for natural language

LLM-as-judge — *LLM-based evaluator*: Model that scores answers on a 1–4 rubric

LPO — *Limited Practice Officer*: Washington State role with limited legal acts

LoRA — *Low-Rank Adaptation*: Parameter-efficient fine-tuning method

MVP — *Minimum Viable Product*: Earliest minimal product to validate approach

NIST AI RMF — *NIST AI Risk Management Framework*: "Govern–Map–Measure–Manage" risk playbook

OECD — *Organisation for Economic Co-operation and Development*: Global AI principles (transparency, accountability)

PEFT — *Parameter-Efficient Fine-Tuning*: Lightweight tuning methods (incl. LoRA)

QA — *Question Answering*: Task/dataset of questions with answers

RAG — *Retrieval-Augmented Generation*: Generate answers grounded in retrieved text

RAGAS — *RAG Assessment*: Metrics for faithfulness, relevancy, context quality

RCW — *Revised Code of Washington*: Washington State statutes (e.g., RCW 2.48.180)

RPC — *Rules of Professional Conduct*: Lawyer ethics rules (e.g., RPC 5.5)

SMART — *Specific, Measurable, Achievable, Relevant, Time-bound*: Goal-setting framework

SME — *Subject-Matter Expert*: Human expert reviewer/validator

UPL — *Unauthorized Practice of Law*: Giving legal advice without a license

Note: This list reflects terms used in the capstone and is intended for quick reference.