What Is the Law? A System for Statutory Research (STARA) with Large Language Models

Faiz Surani*
faiz@law.stanford.edu
Stanford University

Lindsey A. Gailmard*
gailmard@stanford.edu
Stanford University

Allison Casasola
allyc@stanford.edu
Stanford University

Varun Magesh
vim@law.stanford.edu
Stanford University

Emily J. Robitschek
erobitsc@stanford.edu
Stanford University

Daniel E. Ho
dho@law.stanford.edu
Stanford University

ABSTRACT

Each year thousands of bills are enacted into law by Congress, state legislatures, and municipalities. As new laws are enacted, legal codes have grown in length and complexity to such an extent that, in many domains, even the government lacks systematic knowledge of *what the law is*. This challenge has repeatedly stymied legal reform and research efforts. To address the problem, we develop the Statutory Research Assistant (STARA), an automated system capable of performing accurate "statutory surveys"—compilations of all provisions relevant to a particular legal issue, together with detailed annotations and reasoning. We validate STARA's accuracy on three existing human-compiled surveys, showing that it can reproduce them with high fidelity and in many cases identify hundreds of previously undiscovered provisions. We address the unique challenges of automated statutory research, and show that STARA's domain-specific architecture yields substantial improvements in accuracy over off-the-shelf language models. STARA can dramatically reduce the time for discerning the law, and we discuss its considerable implications for legal reform, academic research, and transparency. We make STARA's compiled statutory surveys available to the public and the tool available to researchers upon request.[1]

1 INTRODUCTION

In 1982, the U.S. Department of Justice attempted to count the number of federal crimes. After two years of work, a team of government lawyers could produce only a rough guess: about 3,000. The official who oversaw the effort later remarked that one could "die and [be] resurrected three times," and still not know the true number [17].

Similar challenges have frustrated other efforts to enumerate basic aspects of the law. Over the past century, Congress has enacted thousands of provisions requiring federal agencies to submit periodic reports to Congress. Recent efforts by the Congressional Research Service (CRS) and the Clerk of the House of Representatives to compile a comprehensive list of reporting requirements have variously turned up 3,359 and 2,500 requirements, respectively [14]. But, as both the CRS and the House Clerk acknowledge, no one is quite sure. The U.S. Code's size—more than 32 million words as of 2024—can obscure even unambiguous aspects of the law from legislators, bureaucrats, and citizens.

These challenges are not unique to the federal government. State and municipal codes may be just as dense and voluminous, posing similar challenges for systematic statutory research. In 2024, a civil grand jury investigation concluded that reducing the number of commissions in the City and County of San Francisco was of "vital importance," yet found that "no accurate and comprehensive list of commissions exists within city government" [11]. Despite two initiatives on the November 2024 ballot aimed at streamlining the number of city commissions [43], San Francisco city attorneys have invested considerable time and resources in identifying a comprehensive list, but different sources still disagree on the exact number of commissions authorized by its laws.

Statutory surveys, as such research efforts are known, are widely used in legal practice. Almost every major legal research service provides subscribers with a collection of '50-state surveys,' compilations of state-level laws on various topics, which require extensive human research input to compile. From an academic perspective, statutory surveys also provide rich empirical evidence on core questions of political science and administrative law. For instance, what authority has Congress delegated and what affirmative obligations has it placed on agencies? Finally, statutory surveys are a key element of legal reform. By systematically identifying all provisions relevant to a legal issue, such surveys reveal statutory schemes that may be duplicative, conflicting, or otherwise in need of revision.

But by their nature, statutory surveys are extraordinarily labor-intensive. Researchers must manually read and annotate thousands of provisions, each of which can be thousands of words long. Digitizing statutory codes has helped, with keyword and Boolean queries allowing researchers to narrow their scope of review. But if a legal question inherently implicates a vast number of provisions—as with the many attempts to enumerate criminal offenses—tedium reigns.

The rapid development of large language models (LLMs) offers a potential solution to this problem. LLMs have shown remarkable prowess on a wide range of reading comprehension and legal analysis tasks in recent years. Companies offering legal AI products have raised more than a billion dollars in venture capital in the last year alone [25], and now offer language model-based products for legal tasks ranging from contract review to legal research.

But there are some indications that the problem of automated statutory research remains unsolved. For one, statutes may pose unique challenges for language models. That is because statutory

*These authors contributed equally to this work.
[1]Access to our statutory survey datasets and more information about STARA can be found at https://reglab.github.io/stara/. A shorter version of this paper is forthcoming as: *What Is the Law? A System for Statutory Research (STARA) with Large Language Models*, Proceedings of the 20th International Conference on Artificial Intelligence and Law (ICAIL '25).

provisions have little in common with traditional prose. They contain highly precise language, terminology that may be defined in distant sections, and a web of non-obvious interdependencies. In many cases, individual subsections are nothing more than sentence fragments, understandable only with reference to language contained in a parent provision. Table 1 illustrates these structural complexities, showing how a general-purpose AI system could fail to understand even simple provisions (in left two columns) without a larger sense of statutory structure.

To address these challenges, we introduce the Statutory Research Assistant (STARA). STARA is an end-to-end tool for performing autonomous statutory surveys. Its architecture consists of three primary components. First, we develop a pipeline to annotate important features of a statutory code's structure: its hierarchical organization, defined terms, and internal cross-references. The right column of Table 1 illustrates how STARA exploits the tree-based structure of statutory codes to subdivide and contextualize their provisions for optimal accuracy. We show that this architecture performs robustly across arbitrary statutory codes, including the United States Code and the San Francisco Municipal Code.

Next, STARA's research engine reviews a statutory code for matches and annotates them accordingly. We use a 70 billion-parameter language LLaMA-3 model to classify provisions, extract structured data, and provide reasoning for its decisions. We show that STARA can self-improve its accuracy by first reformulating the researcher's annotation criteria into a specific multi-step rubric and then reasoning with reference to that rubric.

Using three validation datasets, we assess STARA's performance on three benchmark tasks core to governance: identifying what conduct is prohibited (federal criminal offenses), what agencies exist (SF commissions/agency establishment), and what agencies are required to do (congressionally mandated reports). We show that STARA performs comparably to human researchers and extends these existing datasets with many new discoveries. Our analysis of federal crimes identifies 514 additional criminal statutes not captured in the most comprehensive previous survey.

We also demonstrate gains in performance from STARA's domain-informed system design. By comparing STARA's performance to a general-purpose retrieval-augmented generation (RAG) approach, we show the value of context-specific preprocessing, challenging the view that scaling LLM capabilities is the only viable path to improving automated legal reasoning [2, 19, 35].

Last, we develop a web application through which users can create tasks, run them with STARA's research engine, and iterate on its results. We will make STARA publicly available upon publication to assist with statutory research, advance evidence-based legal reform, and promote transparency and accountability in government.

Our work makes five major contributions. First, we develop an end-to-end system capable of accurate, well-reasoned statutory research. We show that STARA is able to reproduce existing survey datasets with high fidelity and supplement them with hundreds of additional discoveries. Second, we demonstrate high precision and recall on each application with reference to existing (though non-comprehensive) datasets. Third, we demonstrate the value of domain-informed preprocessing for the performance of LLM-based research with a set of design ablation experiments and comparisons against commercial AI research systems. Fourth, we make available three STARA-compiled datasets of statutory provisions related to matters of public and academic concern, including congressionally mandated reports, federal criminal offenses, and advisory commissions in the City of San Francisco. Finally, we demonstrate concretely STARA's value for legal research and reform, through a first-of-its-kind assessment of agency compliance with reporting requirements. We find that the majority of reports required to be released publicly are never filed, illustrating a promising area for statutory reform.

2 BACKGROUND

2.1 Legal Codes

Federal, state, and local governments maintain and publish compilations of statutes or ordinances in effect in their jurisdiction in the form of a legal code—a "collection of laws or regulations gathered under one whole corpus, containing a more or less complete system of rules on one of several legal matters" [4]. In the United States, early codification movements were rooted in a democratic desire to make the law more accessible and legible [4, 41] by assembling a "definitive, coherent, and systematic statement of the law" [30].

Statutory codification was first attempted at the federal level in 1866 when Congress authorized President Andrew Johnson to appoint a committee to order and harmonize all active statutes [28]. This effort was met with sharp criticism and, ultimately, failure [28]. A renewed effort began in 1919 [1], resulting in the 1926 Code.

States similarly moved towards codification during the nineteenth century in an effort to provide more clarity and transparency around legal requirements. Louisiana, for instance, sought to use codification to clarify conflicts in inherited legal traditions, civil law, and common law, reforming its criminal, civil, commercial, and procedural laws [34].

Though the codification process varies by jurisdiction, generally new laws are first collected in a supplemental document before being incorporated into the legal code by subject. The U.S. Code is maintained by the Office of Law Revision Counsel of the U.S. House of Representatives (2 U.S.C. § 285b), which publishes an updated edition every 6 years [7].

2.2 Challenges for Statutory Research

Previous efforts to enumerate statutory requirements in a given domain have produced only lower-bound estimates of prevalence [37]. Recent efforts to count the number of distinct federal crimes or the number of congressionally-mandated reports, for instance, have repeatedly failed to produce a complete enumeration of statutory requirements [14, 17].

Three common features of legal codes, in particular, have made comprehensive statutory surveys challenging. First, the sheer volume of statutory provisions poses a challenge for systematic research. Legal codes are massive documents—the U.S. Code consists of 54 titles and approximately 33 million words. The Internal Revenue Code, just one component of the U.S. Code, is 3.4 million words long. And the problem is not unique to the federal code—the San Francisco municipal code and published resolutions consist of 27 volumes and nearly 16 million words.[2]

[2]Indeed the search problem may be worse at the local level, since these codes are not available as a matter of course in common research platforms (e.g., Lexis, Westlaw).

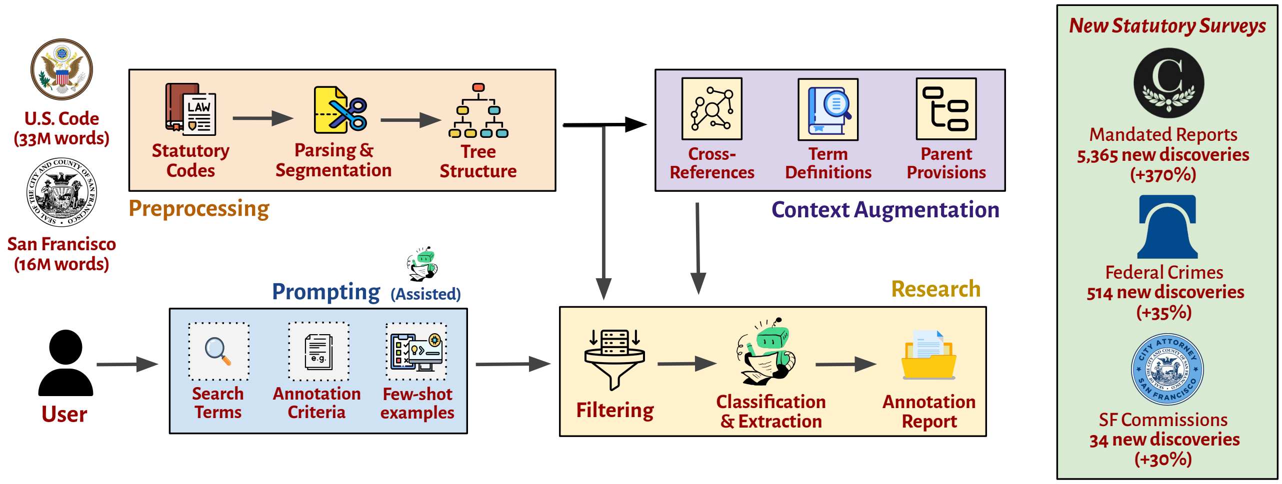

Figure 1: Left: Illustration of the STARA system's components. Right: STARA's contribution to statutory research, showing the number of newly-documented provisions (and percentage improvement over previous datasets) across three validation tasks: congressionally-mandated reports and criminal statutes in the United States Code, and commissions established by the San Francisco Municipal Code.

Second, provisions relevant to a legal issue are often not clustered within a single subdivision of a code. While legal codes are generally organized into titles or chapters by subject, specific types of requirements are not necessarily confined to a particular section of the code. For instance, statutes criminalizing specified actions appear in some 1,500 sections within the code. Similarly, statutes identifying congressionally mandated reports appear in nearly every title.[3] To construct a comprehensive list of requirements, therefore, requires an exhaustive search of the full legal code.

Third, the hierarchical structure and cross references embedded in the code complicate automated processing. And this complexity has only continued to increase. The U.S. Code, for instance, increased by approximately one million words per year between 2008 and 2010 [7]. Deciphering specific provisions requires understanding definitions and terminology possibly contained in distant sections [21]. An automated approach requires the ability to provide the appropriate context to a model, which, in turn, requires incorporating cross-referenced material and an understanding of the hierarchical structure of the code itself.

2.3 Previous Studies

Despite the challenges, scholars and practitioners have nonetheless undertaken many statutory surveys. These efforts fall into two general categories. First, there have been manual efforts to catalog statutory requirements in select titles or subsections of the U.S. Code, including: private causes of action [16]; prohibitions on ex parte communications [40]; judicial review [3]; extent of agency independence [38]; and features of executive agencies [39]. While this strategy potentially provides complete coverage of a particular subset of the Code, it is both labor- and time-intensive. Moreover, this approach fails to identify relevant provisions enumerated in other parts of the Code.

Second, there have been automated attempts to exhaustively search the code. By specifying certain keywords or phrases, researchers can identify provisions that invoke the same phrasing and thereby produce a population estimate of such provisions [8]. Yet this approach is highly sensitive to the keywords chosen by the researcher. Despite parsing an entire legal code, text searches miss requirements that do not conform to the specified language. As a result, this search process is known to be error-prone and incomplete—unable to draw on relevant context elsewhere in the code or systematically surface ambiguous language that may nonetheless characterize provisions of interest.

More recently, researchers have looked to large language models to bridge the gap between manual search techniques and automated search methods to achieve statutory understanding at scale. Researchers have attempted to use LLMs, such as GPT-3 and GPT-4, to perform legal reasoning tasks, including legal examinations [9, 23] and basic statutory reasoning [5]. Fine-tuned versions of GPT-3 have even been found to accurately classify legal rules [26]. While research has pointed to potential gains of these tools to streamline legal tasks, studies have also identified a number of challenges [22], including hallucinations [13] and sensitivity to prompt construction [5] that threaten the reliability, accuracy, and reproducibility of model output. To our knowledge, no prior work has used LLMs to perform large-scale statutory surveys.

3 SYSTEM DESIGN

Off-the-shelf information retrieval methods are not well suited to the task of performing statutory research. Statutes differ from caselaw, prose, and other 'normal' retrieval documents because they are deeply *structured* and *cross-referential*.

[3]Based on our analysis, congressionally mandated reporting requirements appear in 50 titles.

We develop a custom information-retrieval pipeline uniquely tailored to performing exhaustive searches of legal codes. Before performing research, a code is segmented into individual provisions. Each provision is enhanced with structural context and relevant definitions. After this preprocessing, each research task is executed as follows:

(1) A user describes a research question that defines relevant provisions.
(2) With the user's input, an LLM generates a prompt with specific multi-step annotation criteria to identify relevant provisions.
(3) STARA creates a broad keyword filter based on the user's task to create a shortlist of several thousand candidate provisions.
(4) An LLM is prompted to apply the definition to each candidate provision in turn, classifying it as positive (relevant) or negative and extracting relevant information.

This section describes each step of the search and the user interface we have developed to facilitate the process and help refine research tasks.

3.1 Preprocessing

Domain-informed preprocessing is a key step of our pipeline. Legal codes can be hundreds of thousands of provisions long. In order to analyze them with language models, they must be split into smaller components. But effectively chunking legal codes is a challenge; statutory meaning is dependent on surrounding (or distant) context to an unusual degree. Our preprocessing pipeline segments long legal codes into granular provisions and then augments these chunks with precisely chosen context from relevant statutory text.

Segmentation. In order to be analyzed, legal codes must be split into smaller components. Large and complex codes often have many different levels of divisions; the U.S. Code for example has titles, divisions, chapters, parts, sections, clauses, paragraphs, and more. Many codes (e.g., the U.S. Code and the S.F. Municipal Codes) use the 'section' as the primary unit of display. However, this unit of reading convenience doesn't directly correspond to a good unit of analysis; some sections (e.g., 44 U.S.C. § 2907) are just a few words long and serve only to modify other sections, while other sections (e.g., 42 U.S.C. § 1395f) are tens of thousands of words long and contain many subdivisions within them.

In order to maximize flexibility and accuracy in later stages of the pipeline, our segmentation process uses a robust parsing system to divide legal codes into the smallest enumerated units possible. For example, 8 U.S.C. § 1154 contains 8 levels of nested structure, from subsection to sub-sub-item. We split § 1154 into the most granular units possible, capturing numbering, text, and hierarchical information (parents and children in the enumeration hierarchy) for each chunk. The parsing methodology varies by code—for the U.S. Code, we leverage well-annotated XML published by the House Clerk that provides structural hints, while for other codes we use regular expressions to identify common structural markers as well as definitions and cross-references.

After initial parsing, we cluster or split provisions to optimize for LLM analysis. Through experimentation, we found that provisions shorter than 100 words often lack sufficient context for effective extraction, while those longer than 2,000 words tend to be too complex for reliable LLM processing. We therefore employ a clustering system that combines small related provisions and splits larger ones to target an optimal size of approximately 1,000 words per chunk, while preserving the structural relationships captured during parsing.

Statutory Trees. Legal codes are organized in a hierarchy of divisions. When we split codes into chunks, we retain information about the chunk's position in the surrounding hierarchy. This is a key part of the process that allows us to intelligently include the context required to enhance each chunk with context and make it a self-contained unit of analysis.

Consider § 41.4 of San Francisco's Administrative Code: "For purposes of this Chapter 41, the following terms shall have the following meanings…" A research method which naively split up and searched through provisions of the Code would not apply these definitions and could misclassify provisions as a result. STARA's preprocessing system extracts each of the term definitions defined in § 41.4 and associates them with all provisions within the "Chapter 41" scope. Doing so requires a pipeline of regular expression- and LLM-based semantic parsing combined with scope resolution logic and a hierarchival model of the code. High-quality statutory research requires nonlocal knowledge of a code's structure, which STARA provides.

In the tree, each node has information about its own text, position in the hierarchy, and its siblings, children, and parents in the hierarchy. At runtime, STARA traverses the structure to build optimally-sized chunks for LLM inference and enrich those chunks with the necessary context for analysis.

Statutory Context. Before being classified by a large language model or otherwise analyzed, each chunk is enhanced with context to make it a self-contained unit of analysis. As illustrated in the right column of Table 1, each chunk is enhanced with up to four pieces of information:

(1) A heading describing the chunk's citation, position in the code, and titles/descriptions of the chunk's parent divisions.
(2) A list of definitions for terms which appear in the chunk and have explicit definitions given in a relevant part of the code.
(3) 'Lead-ins' or 'chapeaus'. Sometimes, sentences are broken up over multiple numbered chunks, especially when listing or enumerating conditions in the law. A 'chapeau' is the first part of the sentence which describes the purpose of the list, and is necessary to interpret items in the list.
(4) Suffixes, or provisos, continuations, or other modifications of the legal meaning of a chunk that are described in following chunks.

As shown in Table 1, this context allows the reader (and the model) to fully understand provisions that would be incomprehensible in isolation. The example provision in the table - discussing efficiencies and cost savings from agency transfers - would be difficult to interpret without knowing it is part of a reorganization plan that must be submitted to Congress, or without understanding which specific "Corporation" is being referenced. The enhanced version provides this critical context by including the full hierarchical path showing this is part of a reorganization plan section,

Table 1: Context retrieved for 22 U.S.C. § 9682(a)(1)(C) using a general-purpose RAG approach and STARA. STARA retrieves relevant definitions and lead-ins from cross-references.

Statutory Text	General-Purpose RAG	STARA
(C) Any efficiencies or cost savings achieved or additional costs incurred as a result of the transfer of agencies, personnel, assets, and obligations to the Corporation pursuant to this subchapter, including reductions in unnecessary or duplicative operations, assets, and personnel.	(B) Any consolidation, reorganization, or streamlining of agencies transferred to the Corporation pursuant to this subchapter. (C) Any efficiencies or cost savings achieved or additional costs incurred as a result of the transfer of agencies, personnel, assets, and obligations to the Corporation pursuant to this subchapter, including reductions in unnecessary or duplicative operations, assets, and personnel. (2) Consultation Not later than 15 days before the date on which the plan is transmitted pursuant to this subsection, the President shall consult with the appropriate congressional committees on such plan.	**Headings**: Root: United States Code Title: Title 22—FOREIGN RELATIONS AND INTERCOURSE Chapter: CHAPTER 103—BETTER UTILIZATION OF INVESTMENTS LEADING TO DEVELOPMENT Subchapter: SUBCHAPTER VI—TRANSITIONAL PROVISIONS Section: §9682. Reorganization plan Subsection: (a) Submission of plan Paragraph: (1) In general **Definitions**: "*appropriate congressional committees*" — (1) Appropriate congressional committees. The term "appropriate congressional committees" means— (A) the Committee on Foreign Relations and the Committee on Appropriations of the Senate; and (B) the Committee on Foreign Affairs and the Committee on Appropriations of the House of Representatives. "*Corporation*" — There is established in the executive branch the United States International Development Finance Corporation (in this chapter referred to as the "Corporation"), which shall be a wholly owned Government corporation for purposes of chapter 91 of title 31 under the foreign policy guidance of the Secretary of State. **Lead-Ins**: Paragraph: Not later than 120 days after October 5, 2018, the President shall transmit to the appropriate congressional committees a reorganization plan regarding the following: **Text**: (C) Any efficiencies or cost savings achieved or additional costs incurred as a result of the transfer of agencies, personnel, assets, and obligations to the Corporation pursuant to this subchapter, including reductions in unnecessary or duplicative operations, assets, and personnel.

the statutory definition of the Corporation as the U.S. International Development Finance Corporation, and the lead-in text explaining this is part of a required report to congressional committees (which are themselves precisely defined). This systematic approach to context enables significantly improved statutory understanding ability, as we find in Section 6.

3.2 Running a Search

Prompting. Each STARA search begins with a user description of the research task and detailed information about which provisions should be considered relevant. The user is prompted, but not required, to add examples of statutory provisions that are relevant and some examples of those that are irrelevant. The user also specifies 'fields', or pieces of information that should be extracted from each matching provision. An LLM assists the user in turning this information into a longer, highly detailed prompt to be used in research inference.

Filtering. Our system allows the user to apply the LLM across all provisions, or supplement or replace LLM-generated filters with their own criteria. The regular expressions can be as general as the user desires (e.g., "`.*`" to search every provision) to trade off computing and recall. For example, if the user wants to surface only provisions that mention a "tax," the use of regular expressions avoids running the tool unnecessarily on all provisions. In practice, we find that some broad regular expressions can significantly reduce inference costs without reducing recall.

Users can also supplement or replace these LLM-generated filters with their own criteria. This capability is particularly valuable when researchers need to ensure, for methodological rigor, that STARA has reviewed all provisions meeting specific criteria. Every provision that matches at least one query is analyzed by the research model. For exhaustive searches, it is possible to add a regex (`.*`) to capture truly every provision; however, even very broad queries like the ones we reference above tend to include nearly every relevant provision, while reducing inference costs substantially.

Classification. For each provision selected by the regular expressions, STARA prompts an LLM (in our searches, Meta's LLaMA-3.3 70B, but any sufficiently capable LLM can be used) to compare the full, context-enhanced statutory language against the description and examples given by the user. As opposed to static methods like semantic similarity, this method takes advantage of the reasoning capabilities of larger LLMs to apply complex definitions to precise statutory language. If the provision is relevant, the model also extracts fields of interest from the provision. The positive (relevant) and negative (irrelevant) provisions, along with the model's reasoning statements and any extracted fields, are presented to the user in a report generated after the search is complete.

3.3 User Interface

The web interface guides users through the process of performing the search and enables users to easily customize their search parameters. Users can quickly refine searches by choosing positive and negative results and including them as examples in either direction, powering new research tasks with refined examples of target provisions. In addition, the interface enables researchers to filter results by specified fields to more easily review the output from the model and add fields to extract after running the initial search. These post-retrieval features enable researchers to more easily categorize and organize retrieved results after inspecting the output. Figure 2 shows an example of one provision's reasoning and extracted fields for the congressionally-mandated reports task.

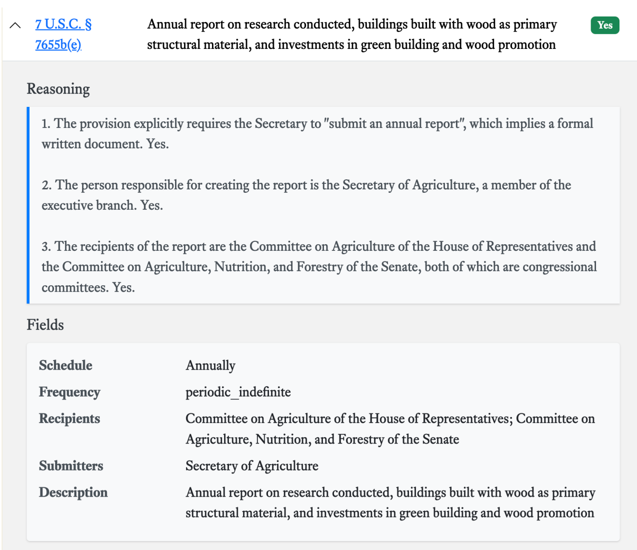

Figure 2: One provision's reasoning and extracted fields, displayed on STARA's web interface.

4 BENCHMARK DATA

To benchmark performance, we collect three tasks related to core functions of government: what conduct is prohibited (federal criminal offenses), what agencies are required to do (congressionally mandated reports), and what bodies exist (San Francisco city commissions). Each task constitutes a substantively significant application of our tool, as previous efforts to enumerate provisions are known to be incomplete.

4.1 Federal Criminal Offenses

In 1982 Justice Department officials set out to count the number of federal crimes in an attempt to convince Congress to revise the federal criminal code. After two years of combing through criminal statutes, Justice Department lawyers had managed to identify only enough provisions to provide an educated guess that the true number of federal crimes was around 3,000 [17]. Building on this effort, in 1998 the American Bar Association conducted a rudimentary computer search of the criminal code in an attempt to produce a more detailed count, only to conclude that the true number of federal offenses was likely much higher than the 1982 estimate [8].

The federal criminal code—Title 18 of the U.S. Code—is notoriously dense and complex, consisting of 1,510 sections that span tens of thousands of pages of text [29]. While Title 18 is the main federal criminal code, there are federal offenses that are not included in Title 18.

The most recent effort to enumerate the Code's criminal provisions was conducted in 2019 by researchers at the Heritage Foundation and Mercatus Center [29]. That work, known as the "Count the Code" project, relied on searching for a set of phrases commonly associated with criminal offenses to identify relevant provisions–terms like "Be punished by a fine" or "Shall be fined or imprisoned" [29].

Despite this limitation, we use the Count the Code dataset as our evaluation benchmark, as it represents the most thorough and well-documented enumeration of federal criminal statutes currently available.

4.2 Congressionally Mandated Reports

Each year, the President, federal agencies, and other federal government entities are required to transmit reports, studies, updates, and other briefings to Congress. Strikingly, reporting has shown that even Congress is not exactly sure what reports are required [15].

Since 1983, the Office of the Clerk of the House has published an annual list of mandated reports with accompanying statutory reference in *Reports to be Made to Congress*. In 2024, the publication enumerated 5,778 required reports across 1,449 statutory sections in effect during the 118th Congress [32]. However, even the list maintained by the House Clerk is known to be incomplete [14].

As recently as 2020, the Congressional Research Service, studying the issue, noted that it was "unaware of a search method that can obtain an exact accounting of all reports required to be submitted to Congress" [14].

In December 2022, Congress passed the "Access to Congressionally Mandated Reports Act,"[4] which required federal agencies to submit certain congressionally mandated reports for publication on GovInfo by the Government Publishing Office (GPO). The GPO now maintains a centralized repository of reports submitted to Congress and select committees or subcommittees on GovInfo. As of October 1, 2023, agencies are required to submit reports to the GPO portal within 60 days of the report being submitted to Congress [31].

To assess recall on this task, we use the House Clerk's dataset of 1,449 sections as our evaluation dataset. We further use the set of reports currently available on GovInfo to assess agency noncompliance based on STARA's enumeration of reports.

4.3 San Francisco City Commissions

In 2024, residents introduced two ballot propositions to reduce the number of city commissions: Proposition D would limit the number of commissions to 65, and Proposition E would establish a task force to streamline the city's commissions. While Proposition D failed, Proposition E passed [43], requiring San Francisco to assess whether commissions should be eliminated and suggest improvements.

[4]Pub. L. No. 117-263, sections 7241-7248.

However, there is no complete list of commissions that exist and what functions they are designed to serve [10]. An op-ed in the *San Francisco Chronicle*—calling Proposition D “deeply flawed”—noted this discrepancy: "Depending on whom you talk to, San Francisco has anywhere from 115 to 135 commissions, boards, and advisory committees" [6].

This ambiguity has persisted despite several efforts to compile a comprehensive list. In 2022, the San Francisco City Attorney’s office released a list of 132 entities [12]; two years later, a civil grand jury impaneled by the city found yet more [36]. We partnered with the San Francisco City Attorney’s office to help construct an exhaustive list of the city’s commissions using STARA.

The difficulties of identifying commissions are evident in the definition provided by San Francisco officials for STARA’s commission search:

> For our purpose, the definition of "commission" is any board, commission, committee, council, task force, working group, or other body, whether permanent or temporary, decision-making or advisory, that is created by the Charter, ordinance, resolution, state or federal statute, or formal action of any board or commission. There are a few exceptions to the definition: it does not include (1) the Board of Supervisors, (2) any standing or special committees of the Board of Supervisors, (3) committees of a commission consisting entirely of that commission’s members…

Even with the help of traditional search, a legal researcher would need to sift through thousands of mentions of not only “commissions,” but “boards,” “working groups,” etc. to find the code sections establishing each and every commission. This challenge is compounded by the fact that commissions can be established not only in the Municipal Code but also in city resolutions—of which there are 15,833. In each instance where a keyword relates to a commission, the researcher must check several factors, including whether the provision establishes a commission or merely references it, the composition of its membership, and its relationship to other bodies.

5 EVALUATION METHODOLOGY

STARA processes legal codes to identify and analyze provisions matching user-defined criteria. The system combines automated classification of provisions with structured extraction of key fields, producing a dataset for expert review. We evaluate this approach across three distinct statutory research tasks, measuring both the system’s ability to surface known relevant provisions and to identify previously undocumented ones.

To this end, the metric we are most interested in is STARA’s recall, or the percentage of relevant elements which are surfaced by the tool. Also of interest, but less important, is STARA’s precision, or the percentage of surfaced elements which are in fact relevant to the researcher. It is hard to measure STARA’s performance in absolute terms, since each search is different. Some research tasks may be unambiguously defined, while others may be more difficult to define and even involve disagreement between human labelers. We use the three test datasets discussed in Section 4 as case studies to evaluate representative precision and recall metrics for STARA: federal criminal statutes, congressionally mandated reports, and San Francisco commission establishment.

5.1 Recall

We estimate STARA’s recall by comparing it against pre-existing, human-created datasets. For federal criminal statutes and congressionally-mandated reports, we follow this process:

(1) Enumerate all positive provisions from existing datasets
(2) Filter out provisions that have been repealed, expired, or moved
(3) Extract title and section numbers from the filtered set
(4) Compare against titles and sections flagged by STARA

For these two datasets, the point estimate of recall is the proportion of matching provisions which STARA identifies as relevant. Some provisions may contain multiple relevant items (e.g., three separate mandated reports), but for the purposes of automated evaluation, we treat a single match to a particular section as sufficient. For the San Francisco commissions dataset, since the existing dataset enumerated commission names rather than statutory citations, we calculated recall at the commission name level—that is, whether STARA identified the sections establishing each known commission.

To validate our recall estimates, we reviewed STARA’s apparent false negatives to ensure they represented genuine misses rather than dataset false positives.[5] For all three tasks, the small number of potential false negatives (less than 150) allowed for exhaustive review. Where a provision was determined not to be relevant, we excluded it from the dataset and adjusted our recall estimate accordingly.

This method does not assume that the human-created datasets are complete or perfect, but it does assume that their false negatives are not correlated with STARA’s false negatives.[6]

5.2 Precision

While recall can be estimated against pre-existing datasets, precision is more challenging. That is because existing datasets (e.g., the House Clerk’s list of mandated reports) are themselves incomplete attempts at enumeration. Naïvely measuring precision against these datasets would therefore treat a new discovery of a true positive as a false positive.

To better estimate STARA’s precision, we randomly sampled 50 positive results for each task and manually annotated whether the retrieved result conformed to the task’s definition of a relevant provision.

5.3 Efficiency gains

We compare the rate of review, total duration, and cost of comprehensive statutory review using STARA versus manual review for the San Francisco commission identification task. This provides a

[5] We encountered several categories of dataset “false positives”: (1) provisions which were relevant but had since been repealed, moved, or expired; (2) provisions which were not relevant to the defined scope of the task but had been included in the dataset for other reasons; and (3) errors, typographical or substantive, in the source datasets.

[6] Because STARA uses some techniques common among human researchers, such as keyword filters, the false negatives may be correlated. Nevertheless, we believe these numbers provide useful guidance on the relative quality of STARA searches compared to previous human efforts.

concrete benchmark for the efficiency gains associated with our tool.

For manual review, we had two city employees search for commissions in a randomized series of the subset of codes with any commission keywords while tracking their time and success rate. We measured their completion rate and time taken to estimate the total effort required for comprehensive review. We calculated a cost estimate for a human review based on an hourly rate derived from publicly available city attorney job postings.

For STARA, we measured the total wall-clock time from initiating the search to receiving complete results across both the Municipal Code and all 15,833 city resolutions. We estimated compute costs based on current market rates for GPU compute time ($2.58/hour for our configuration of two NVIDIA A100 GPUs [24]).

5.4 Ablation Study

To validate our system design choices, we conduct an ablation study on the federal criminal statutes task, progressively adding components of our pipeline to demonstrate their impact. We evaluate four configurations:

(1) Base: Raw text with basic formatting instructions
(2) +Prompt Engineering: Adding criteria rewriting and few-shot examples
(3) +Basic Context: Including lead-ins and suffixes
(4) Full System: Adding headings, definitions, and cross-references

For each configuration, we measure precision and recall as described above. To evaluate more complex statutory understanding capabilities, we also measure **extraction accuracy**—the system's ability to correctly identify and extract three key components of each criminal statute: the prohibited behavior, severity classification (misdemeanor, felony, or both), and prescribed penalty. Similar to our precision measurement methodology, we randomly sampled 50 provisions for each configuration and manually assessed whether all components were both complete and correct, with no hallucinated details. By measuring performance on this more open-ended task across different system configurations, we can better assess how different components contribute to deeper statutory understanding beyond simple classification.

5.5 Comparison with Deep Research and Legal AI Systems

To contextualize STARA's performance relative to other state-of-the-art AI research systems, we evaluate against recently released Deep Research products from OpenAI [33] and Google [18] (referred to as OpenAI Deep Research and Gemini Deep Research, respectively), as well as Westlaw's AI Jurisdictional Surveys tool [42]. The Deep Research systems represent the current frontier of general-purpose AI research tools, designed to autonomously search the web, synthesize information, and produce comprehensive research reports on complex topics. Westlaw AI Jurisdictional Surveys is a specialized legal research tool that uses generative AI to "expedite and streamline the process of finding relevant laws across multiple jurisdictions" and is integrated within Thomson Reuters' legal database.

For the federal criminal statutes task, we instructed each of the systems to return all statutes defining criminal offenses in the United States Code or its associated statutory notes. For the Deep Research systems, which access the open internet, we also instructed them to not access the specific benchmark dataset used as ground truth (Count the Code) or a similar compilation produced by the National Association of Criminal Defense Lawyers, but otherwise use any primary or secondary source desired to compile criminal statutes.

Table 2: Performance on Validation Tasks

Task	# Known	# New	Prec.	Recall
Mandated Reports	1,449	5,365	0.96	0.995
Criminal Statutes	1,469	514	0.98	0.998
S.F. Commissions	111	33	0.94	1.0

During our evaluation, we observed significant variation in performance between different runs of the tools. To best represent the ceiling of these systems' capabilities, we ran our query three times on each system and report the best-of-3 recall, along with the precision of the highest-recall run. In some cases, OpenAI's Deep Research system disregarded instructions not to access the benchmark dataset; we exclude these runs and present results from the best of three runs in which it did not access the prohibited sources.

We report two metrics for all systems: precision and recall. We compute an upper bound on precision by manually labeling all provisions flagged by the comparison systems that were detected by neither STARA nor contained within the ground truth dataset, and assuming that all other entries are true positives. That is, where a comparison system flagged a provision also flagged by Count the Code or STARA, we assume that that provision is a true positive. We measure recall against the ground truth Count the Code dataset.

6 RESULTS

We now discuss results for three tasks: enumerating congressionally mandated reports, federal criminal offenses, and commissions in the City of San Francisco. As shown in Table 2, STARA demonstrates strong performance across all tasks with high precision (94-98%) and near-perfect recall (99.5-100%), while identifying hundreds of previously undocumented instances in each domain.

6.1 Federal Criminal Offenses

The most comprehensive effort to date identified 1,469 federal criminal provisions in 2019 [8]. Using STARA, we identify 1,983 criminal provisions with 98% precision—over 500 more provisions than previously documented. Our tool achieves 99.8% recall on known provisions.[7]

[7] An earlier version of this paper reported 2,305 criminal provisions. This was due to an error in de-duplicating STARA predictions by title and section. STARA's precision and recall are unaffected by this error, as precision was computed on a subset which contained no duplicate sections, and the the presence of duplicate sections does not bear on a recall calculation.

6.2 Congressionally-Mandated Reports

STARA identifies 6,814 sections in the U.S. Code containing reporting requirements. When evaluated against known reporting requirements, STARA achieves 99.5% recall and 96% precision.

STARA located 1,119 reports within the Code subject to public reporting by the Access to Congressionally Mandated Reports Act.[8] However, only 42 (less than 4%) were available through GovInfo and just 98 (less than 9%) were noted in the Congressional Record. Of a random sample of 55 reports not included in the Congressional Record or published on GovInfo, just 12 reports were available elsewhere online. One former member of Congress admitted he never reviewed many of the reports Congress had required agencies to submit: "To be honest, a lot of the reports that have been mandated from these federal agencies are so overwhelming that I [didn't] generally look at them" [15].

The low publication rate for these reports raises important practical and legal questions. First, do the reports actually exist in some form? And if not, do agencies even realize that they must be submitted? The low rates of compliance observed may also partially stem from vague statutory deadlines [14] or other statutory ambiguities.

Finally, it is unclear whether each report is necessary absent a complete inventory of requirements—are certain reports redundant or even irrelevant? For instance, 31 U.S.C. § 5112(p)(3)(B) requires the Federal Reserve Board of Governors to submit an annual report on obstacles to the circulation of $1 coins, along with progress towards goals and recommendations for legislative action. The Federal Reserve has dutifully filed its annual report to Congress on the "Presidential $1 Coin Program" each year since 2007 despite, as the 2023 report points out, the suspension of production of $1 coins for circulation as of 2012. Since 2012, the Board of Governors has recommended the elimination of the reporting requirement, but to no avail. A comprehensive list of reporting requirements allows lawmakers and agency personnel to more systematically identify ambiguous, conflicting, redundant, or obsolete requirements.

6.3 San Francisco City Commissions

Comparing our results to the 111 documented commissions known to be defined within the search text,[9] the STARA tool finds all 111 documented commissions (100% recall), as well as 33 additional commissions, with 94% precision.

STARA demonstrates dramatic efficiency gains for the commission identification task. In approximately 2 hours of manual review, two evaluators found 60% and 86% (21 and 30 out of 35 known), respectively, of the known bodies in the Administrative Code, identifying between 1-8 out of 24 new possible bodies. While it is possible that there is significant variation in difficulty between code sections with more or fewer relevant provisions, if human evaluation identifies 11–17 commissions/hour at a steady rate, it would take approximately 8.5–13 hours to find the 145 commissions identified by STARA. This estimate is likely conservative, as search costs are not linear and multiple passes might be needed to ensure completeness. Additionally, this evaluation does not capture whether moving from 60-86% to 100% recall of commissions is significantly harder for human evaluators.

[8]Inclusion within the category of reports subject to ACMRA was determined by STARA analysis; we randomly sampled 50 such determinations and found no errors.

[9]While the City's released list contains 134 entities, not all meet the definition provided by the City Attorney's office in Section 4; 111 remain after filtering to this definition.

Table 3: Comparison of task duration and cost of human annotation vs. STARA for commission identification.

	Human Annotator	STARA
Duration	8-13.5 hours	20 minutes
Cost	$\sim \$3,000$	$\sim \$0.86$

In contrast, STARA completed comprehensive analysis of both the Municipal Code and all city resolutions in just 20 minutes—finding *all* known commissions and 33 additional ones not uncovered in the city's earlier efforts. The cost difference is equally stark—what would cost approximately $3,000 in human annotation time costs just $0.86 in compute resources. Most importantly, the gains were significant to practitioners. In the assessment of the Deputy City Attorney, who spearheaded these implementation efforts, STARA was "outrageously helpful." Table 3 summarizes these efficiency gains. The methodology for this comparison is detailed in Section 5.

6.4 Ablation Study Results

We conducted an ablation study to quantify the benefits of our system design choice, as shown in Table 4. While standard prompt engineering techniques (few-shot examples and criteria rewriting) provide meaningful improvements in precision and recall, our domain-specific enhancements for statutory text provide much more substantial gains. In particular, the addition of lead-ins and suffixes yields a substantial improvement in extraction accuracy, as these elements often contain critical details about the nature and severity of criminal offenses—precision rises from 0.58 in the baseline system to 0.98.

Interestingly, despite stark differences in precision and extraction accuracy, all ablations of the system still achieve very high recall. We hypothesize that this is because it is relatively easy to identify statutes which *appear* to define criminal offenses, but much more difficult to distinguish actual criminal offenses from other similar provisions such as civil penalties and administrative disciplinary measures.[10]

In particular, the extraction accuracy results (accuracy in identifying components of criminal statutes) reveal the significant benefit of a domain-specific architecture for *more complex* statutory analysis. Without access to these features, the system struggles to synthesize complete offense descriptions and penalties that are distributed across multiple parts of a provision. The gains in extraction accuracy with each additional component from 0.28 to 0.76 suggest that higher-level statutory analysis is particularly aided by domain-specific enhancements This mirrors the complexity of actual legal research tasks, where practitioners must not only find relevant provisions but understand their complete meaning in context.

[10]This dynamic may not be present in other statutory analysis tasks, where the task of locating potentially relevant provisions may be more difficult.

Table 4: Ablation study results on federal criminal statutes task. Extraction measures accuracy and completeness in identifying offense descriptions and penalties.

Configuration	Precision	Recall	Extraction
Base	0.58	0.990	0.28
+Prompt Engineering	0.76	0.987	0.32
+Basic Context	0.96	0.984	0.70
Full System	**0.98**	**0.998**	**0.76**

Table 5: Comparison of STARA with other AI systems on the federal criminal offenses task

System	# Found	Precision	Recall
STARA	**1,983**	**0.98**	**0.998**
Gemini Deep Research	282	0.890	0.144
OpenAI Deep Research	84	0.881	0.044
Westlaw AI Jurisdictional Survey	113	0.415	0.072

6.5 Comparison with Deep Research and Legal AI Systems

To benchmark STARA against leading AI research tools, we compared performance on the federal criminal statutes task with OpenAI Deep Research, Gemini Deep Research, and Westlaw AI Jurisdictional Surveys. As shown in Table 5, STARA substantially outperforms all comparison systems across key metrics.

All comparison systems dramatically underperform STARA in recall. OpenAI Deep Research identified only 84 of the 1,469 known criminal statutes (0.044 recall), while Gemini Deep Research found 282 provisions (0.144 recall) and Westlaw AI Jurisdictional Surveys found 113 provisions (0.072 recall). In contrast, STARA identified nearly all known provisions (0.998 recall) while discovering over 500 additional criminal statutes not documented in previous efforts. Overall, STARA locates 17 times as many provisions as the best-performing comparison system.

The Deep Research systems achieve reasonable precision (0.88–0.89), but Westlaw AI Jurisdictional Surveys exhibited both poor recall and concerning reliability issues. Despite being a specialized legal research tool with access to a legal database, Westlaw achieved only 41.5% precision—not only mis-classifying non-criminal provisions, but also hallucinating 121 statutory provisions in the U.S. Code that do not actually exist.

These results highlight a fundamental difference in approach: most AI research systems are designed to gather some number of relevant documents and then summarize them, but STARA is designed for *systematic* research tasks, where one must locate all statutory language meeting specified criteria. While Deep Research tools excel at synthesizing information from web sources and specialized legal AI tools provide access to curated legal databases, both approaches struggle with the large-scale enumeration and reliability required for comprehensive statutory surveys.

7 LIMITATIONS

We now articulate some of the chief limitations of STARA.

Pre-filtering Limitations. STARA can be run on entire legal codes for exhaustive searches where completeness is desired. However, as a cost and time saving measure, users can optionally pre-filter provisions using keyword-based regular expressions to focus analysis on likely relevant sections. This optional pre-filtering step currently relies on keyword matching rather than semantic search capabilities.

Our experiments showed that embedding-based semantic search were effective at finding *some* positive matches, but struggled to capture the long tail of sometimes thousands of relevant provisions, for many of the same reasons that off-the-shelf LLM systems struggle to understand statutory text discussed in Section 3. Developing statute-specific semantic search techniques or models could improve this optional filtering step. However, this limitation can always be avoided entirely by running STARA on complete legal codes without pre-filtering, albeit at a higher cost in time and compute.

Lack of Automated Deduplication. Another limitation is that STARA does not automatically deduplicate or cluster semantically equivalent provisions. When searching for entities like commissions or reporting requirements, the same entity may be referenced or established in multiple sections or subsections of the code. STARA will extract each reference as a separate result, which can create additional work for researchers who must manually identify and consolidate duplicate entries. While this approach ensures no relevant provisions are missed—especially given the tendency to borrow similar language across code provisions—it means that raw counts of STARA's results cannot be used directly without further processing. In future revisions, we plan on implementing deduplication and/or clustering to make the consolidation easier.

Other Sources of Legal Requirements. STARA focuses on provisions enumerated in legal codes (U.S. Code, S.F. Municipal Code & Resolutions), but there are of course other sources of legal requirements. Executive orders, codified regulations, and agency guidance documents, not to mention judicial opinions, may be sources of law. To understand the full extent of legal requirements in many contexts, therefore, would require searching across all relevant sources of law. The general methodology we have outlined could be extended to encompass a range of executive actions, regulations, and agency materials to expand STARA's coverage.

8 CONCLUSION

Without a complete inventory of statutory obligations and rules, much of law and government can remain opaque—what conduct is prohibited, what authority is vested in government officials, and what tasks those officials are required to perform. STARA enables researchers, advocates, attorneys, and government officials to obtain a comprehensive understanding of statutory requirements in effect, enabling these actors to understand the full breadth of legislative mandates. Moreover, STARA highlights opportunities for statutory simplification, code reform, and surfaces legal ambiguity [20], which may assist statutory reform efforts and further highlight where human attention would be most valuable [27].

Finally, from a technical standpoint, STARA also demonstrates the potential of domain-specific systems built on top of large language models. While general-purpose LLMs and RAG implementations struggle with the structure of statutory text, STARA shows how careful system design that accounts for domain-specific challenges can unlock powerful new capabilities. This suggests a promising path forward for AI applications in complex domains: rather than relying solely on end-to-end learning or general-purpose tools, systems that combine the reasoning capabilities of LLMs with domain-specific design may prove most effective.

ACKNOWLEDGMENTS

We thank Ananya Karthik, Christopher D. Manning, Neel Guha, Emaan Hariri, Lucia Zheng, Isabel Gallegos, Mirac Suzgun, Elena Eneva, Jonathan Hennessy, Erin Maneri, Derek Ouyang, and Kit Rodolfa for helpful feedback and comments; and David Chiu, Jon Givner, Andrea Bruss, Leah Granger, and Rebekah Krell for the collaboration.

REFERENCES

[1] 1921. Codification and Revision of the General Laws of the United States. *Columbia Law Review* 21, 4 (April 1921), 373–376. https://www.jstor.org/stable/1112699

[2] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. 2024. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences* 121, 27 (June 2024). https://doi.org/10.1073/pnas.2311878121

[3] Leland E. Beck. 2013. Agency Practices and Judicial Review of Administrative Records in Informal Rulemaking. Washington, DC: Administrative Conference of the United States.

[4] Jean Louis Bergel. 1988. Principal Features and Methods of Codification. *Louisiana Law Review* 48, 5 (1988). https://digitalcommons.law.lsu.edu/lalrev/vol48/iss5/3

[5] Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can GPT-3 Perform Statutory Reasoning?. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law* (Braga, Portugal) *(ICAIL '23)*. Association for Computing Machinery, New York, NY, USA, 22–31. https://doi.org/10.1145/3594536.3595163

[6] Chronicle Editorial Board. 2024. Endorsement: S.F. needs to cut its bloated number of commissions. But Prop D is a first draft, not a solution. https://www.sfchronicle.com/opinion/editorials/article/prop-d-sf-commissions-19829525.php

[7] Michael J. Bommarito and Daniel M. Katz. 2010. A mathematical approach to the study of the United States Code. *Physica A: Statistical Mechanics and its Applications* 389, 19 (Oct. 2010), 4195–4200. https://doi.org/10.1016/j.physa.2010.05.057

[8] GianCarlo Canaparo, Patrick A. McLaughlin, Jonathan Nelson, and Liya Palagashvili. 2022. Count the Code: Quantifying Federalization of Criminal Statutes. *The Heritage Foundation Special Report* 251 (Jan. 2022). https://www.heritage.org/sites/default/files/2024-05/SR251.pdf

[9] Jonathan H Choi, Kristin E Hickman, Amy B Monahan, and Daniel Schwarcz. 2021. ChatGPT goes to law school. *J. Legal Educ.* 71 (2021), 387.

[10] City and Civil Grand Jury County of San Francisco. 2023-2024. Commission Impossible: Getting the Most from San Francisco's Commissions. https://www.sf.gov/sites/default/files/2024-06/Commissions%20Impossible%20Report.pdf

[11] City and County of San Francisco, Civil Grand Jury. 2023-2024. Press Release: Civil Grand Jury: Commission Impossible. https://www.sf.gov/sites/default/files/2024-06/Commission%20Impossible%20Press%20Release.pdf

[12] City and County of San Francisco Office of the City Attorney. 2022. List of City Boards, Commissions, and Advisory Bodies Created by Charter, Ordinance, or Statute.

[13] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis* 16, 1 (2024), 64–93.

[14] William T. Egar. 2020. Congressionally Mandated Reports: Overview and Considerations for Congress. https://www.everycrsreport.com/files/20200514_R46357_3c89b961380edcaceb4d4623833bf92f93e424b1.pdf

[15] David Fahrenthold. 2014. Unrequired Reading. *The Washington Post* (May 2014). https://www.washingtonpost.com/sf/national/2014/05/03/unrequired-reading/

[16] Sean Farhang. 2010. *The Litigation State: Public Regulation and Private Lawsuits in the United States*. Princeton University Press.

[17] Gary Fields and John R. Emshwiller. 2011. Many Failed Efforts to Count Nation's Federal Criminal Laws. *The Wall Street Journal* (2011). https://www.wsj.com/articles/SB10001424052702304319804576389601079728920

[18] Google. 2024. Gemini Deep Research. https://gemini.google/overview/deep-research/

[19] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. Deep Learning Scaling is Predictable, Empirically. arXiv:1712.00409 [cs.LG] https://arxiv.org/abs/1712.00409

[20] Meredith A Holland. 2017-2028. The Ambiguous Ambiguity Inquiry: Seeking to Clarify Judicial Determinations of Clarity versus Ambiguity in Statutory Interpretation. *Notre Dame Law Review* 93, 3 (2017-2028), 1371–1392.

[21] Michael J. Bommarito II and Daniel Martin Katz. 2010. Properties of the United States Code Citation Network. arXiv:0911.1751 [physics.soc-ph] https://arxiv.org/abs/0911.1751

[22] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169* (2023).

[23] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. GPT-4 Passes the Bar Exam. *Philosophical Transactions of the Royal Society A* 382 (2024). http://dx.doi.org/10.2139/ssrn.4389233

[24] Lambda Labs. 2024. Lambda GPU Cloud Pricing. https://lambdalabs.com/service/gpu-cloud#pricing Accessed March 2024.

[25] Steven Lerner. 2024. Legal Tech Funding Surges In Q3 Thanks To 3 Megadeals. *Law 360* (Oct. 2024). https://www.law360.com/pulse/articles/1889369/legal-tech-funding-surges-in-q3-thanks-to-3-megadeals

[26] Davide Liga and Livio Robaldo. 2023. Fine-tuning GPT-3 for legal rule classification. *Computer Law & Security Review* 51 (2023), 105864. https://doi.org/10.1016/j.clsr.2023.105864

[27] Kangcheng Luo, Quzhe Huang, Cong Jiang, and Yansong Feng. 2025. Automating Legal Concept Interpretation with LLMs: Retrieval, Generation, and Evaluation. arXiv:2501.01743 [cs.CL] https://arxiv.org/abs/2501.01743

[28] Michael J. Lynch. 1997. The U.S. Code, the Statutes at Large, and Some Peculiarities of Codification. *Legal Reference Services Quarterly* 16, 1 (1997), 69–84. https://doi.org/10.1300/J113v16n01_04

[29] Patrick A. McLaughlin and Liya Palagashvili. 2023. Counting the Code: How Many Criminal Laws Has Congress Created? https://www.mercatus.org/research/policy-briefs/counting-the-code-congress-criminal-laws

[30] Andrew P. Morriss. 1998. Codification and right answers. *Chi.-Kent L. Rev.* 74 (1998), 355.

[31] Office of Management and Budget. 2023. Access to Congressionally Mandated Reports Act Implementation Guidance. https://www.whitehouse.gov/wp-content/uploads/2023/06/M-23-17-Access-to-Congressionally-Mandated-Reports-Act-Implementation-Guidance.pdf

[32] U.S. Government Publishing Office. 2024. H. Doc. 118-97- Reports To Be Made To Congress. https://www.govinfo.gov/content/pkg/CDOC-118hdoc97/pdf/CDOC-118hdoc97.pdf

[33] OpenAI. 2024. Introducing Deep Research. https://openai.com/index/introducing-deep-research/

[34] Agustín Parise. 2012. Codification of the Law in Louisiana: Early Nineteenth-Century Oscillation Between Continental European and Common Law Systems. *Tulane European and Civil Law Forum* 27 (2012), 133–164. https://journals.tulane.edu/teclf/article/view/1816

[35] Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. 2024. Observational Scaling Laws and the Predictability of Language Model Performance. arXiv:2405.10938 [cs.LG] https://arxiv.org/abs/2405.10938

[36] San Francisco Civil Grand Jury. 2024. *Commission Impossible? Getting the Most from San Francisco's Commissions*. Civil Grand Jury Report. City and County of San Francisco.

[37] Peter H. Schuck. 1992. Legal Complexity: Some Causes, Consequences, and Cures. *Duke Law Journal* 42, 1 (Oct. 1992), 1–52. https://scholarship.law.duke.edu/dlj/vol42/iss1/1

[38] Jennifer L. Selin. 2015. What Makes an Agency Independent? *American Journal of Political Science* 59, 4 (2015), 971–987. http://www.jstor.org/stable/24582960

[39] Jennifer L Selin and David E Lewis. 2018. *Sourcebook of United States executive agencies* (second ed.). Administrative Conference of the United States.

[40] Esa L Sferra-Bonistalli. 2014. Ex Parte Communications in Informal Rulemaking. In *Report prepared for the consideration of the Administrative Conference of the United States, Washington, DC*.

[41] Norman W. Spaulding. 2004. The Luxury of the Law: The Codifification Movement and the Right to Counsel. *Fordham Law Review* 73 (2004), 373–376. https://ir.lawnet.fordham.edu/flr/vol73/iss3/10

[42] Thomson Reuters. 2024. Westlaw Precision with CoCounsel: Generative AI for lawyers. https://legal.thomsonreuters.com/en/c/westlaw/westlaw-precision-generative-ai

[43] Zhe Wu. 2024. San Francisco Proposition Results for Nov. 5 Election. *San Francisco Public Press* (Nov. 2024). https://www.sfpublicpress.org/san-francisco-proposition-results-for-nov-5-election/