

Software Engineering Methods for AI-Driven Deductive Legal Reasoning

Rohan Padhye

Carnegie Mellon University
Pittsburgh, USA
rohanpadhye@cmu.edu

Abstract

The recent proliferation of generative artificial intelligence (AI) technologies such as pre-trained large language models (LLMs) has opened up new frontiers in computational law. An exciting area of development is the use of AI to automate the deductive rule-based reasoning inherent in statutory and contract law. This paper argues that such automated deductive legal reasoning can now be viewed from the lens of software engineering, treating LLMs as interpreters of natural-language programs with natural-language inputs. We show how it is possible to apply principled software engineering techniques to enhance AI-driven legal reasoning of complex statutes and to unlock new applications in automated meta-reasoning such as mutation-guided example generation and metamorphic property-based testing.

CCS Concepts: • **Applied computing** → Law; • **Computing methodologies** → Natural language generation; • **Information systems** → Question answering; • **Software and its engineering** → *Software testing and debugging*; Search-based software engineering.

Keywords: computational law, generative artificial intelligence, large language models, legal reasoning, statutory reasoning, software engineering, property-based testing, mutation testing, example generation

ACM Reference Format:

Rohan Padhye. 2024. Software Engineering Methods for AI-Driven Deductive Legal Reasoning. In *Proceedings of the 2024 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software (Onward! '24)*, October 23–25, 2024, Pasadena, CA, USA. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3689492.3690050>



This work is licensed under a Creative Commons Attribution 4.0 International License.

Onward! '24, October 23–25, 2024, Pasadena, CA, USA

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1215-9/24/10

<https://doi.org/10.1145/3689492.3690050>

1 Introduction

Numeric computations in law, regulation, and contracts govern many aspects of our personal and professional lives—owing taxes, qualifying for student financial aid, receiving social security benefits, breaking a rental lease, paying down a credit card balance, demanding compensation for a delayed flight, and bequeathing property in a will, to name a few examples. In each of these cases, the logic of such computation is encoded in legally-binding natural language documents of sometimes monstrous complexity. The process of establishing how such legal rules apply to specific situations usually falls into the category of *deductive legal reasoning* [14] (as opposed to *analogical* or case-based legal reasoning that draws upon precedents set by judges).

As such, deductive legal reasoning is highly mechanical when all the relevant facts are precisely known. Yet, this is by no means straightforward. For example, Appendix A lists specific parts of the the *Internal Revenue Code* (IRC), the body of statutory law that deals with U.S. federal taxation. Consider the following (well-specified) problem:

Example 1. *Alice and Bob are married and filing taxes jointly for the year 2018. Alice was born on 1/1/1981 and Bob was born on 12/30/1975. In 2018, Alice and Bob’s adjusted gross income was \$216,350. Alice and Bob do not itemize their deductions and do not qualify for any deductions other than the standard deduction. What is their taxable income for 2018?*

Solution Approach. Although a seemingly simple question, correctly answering it requires applying multiple computational rules from the IRC¹ (ref. Appendix A): §63(b) and §63(c)(1) to identify the calculation of standard deduction and taxable income; §63(c)(2)(A)(i) and §63(c)(2)(C) to determine the formula for basic standard deduction applicable to joint returns; §63(c)(7)(A)(ii) which overrides the value \$3,000 in the preceding formula with the value \$12,000; as well as realizing that several other rules such as §63(c)(3) and §63(f)(1) do not apply because Alice and Bob are not 65 years old in 2018, and that inflation adjustments decreed in §63(c)(4) and §63(c)(7)(B) do not apply in 2018. The final answer is $\$216,350 - (2 \times \$12,000) = \$192,350$.

¹While there are numerous paper worksheets and software packages produced by tax preparation services, those are not the law. Someone has to manually interpret the statutes in the IRC, which is the ground truth specification. This paper concerns the underlying legal analysis.

The reasoning process thus involves evaluating the composition of multiple rules, resolving internal cross-references, handling exceptions that may invalidate seemingly applicable rules, and performing arithmetic calculations; crucially, however, this process assumes that all the facts (as they pertain to the rule-based reasoning) are clearly known and so it does not deal with subjective interpretation of potentially ambiguous terms [15]. Similar legal reasoning is required in other previously mentioned domains: for example, determining to what extent an insurance policy might cover a particular incident, or correctly executing a will of a decedent given specific circumstances about their assets and beneficiaries.

The complexity increases even further when considering the objectives of those drafting the relevant legislation or contracts, who may be interested in a form of *meta-reasoning* of edge cases. For example, government legislators might wonder if the tax law introduces any loopholes, whereas a testator might want to ensure that their will robustly expresses their intentions even in circumstances where some of their beneficiaries pass before themselves. The meta-reasoning problem is one of validating a legal text to satisfy some desirable properties across the universe of hypothetical scenarios.

Naturally, researchers have attempted to automate deductive legal reasoning to reduce the risk of costly errors; however, classical approaches have focused on formalizing the rules using special logics or domain-specific programming languages [18, 19, 26, 27, 30]—while these approaches are sound, the formalization process itself appears to be too tedious for these techniques to achieve widespread adoption.

The recent proliferation of generative artificial intelligence (Gen AI) technologies such as pre-trained large language models (LLMs) have made it possible to directly automate deductive legal reasoning using the natural language documents [4, 17, 31], with limited but progressively improving success rates. For example, prompting Anthropic’s Claude 3 Opus [1] with a preface (ref. Fig. 1) and the text from Example 1 is usually sufficient to solve the problem.²

In this paper, we argue that AI-driven legal reasoning is essentially a form of (noisy) program execution, where both the implementation code and the program input is represented in natural language—the language models are thus *interpreters*. As such, it should be possible to understand, debug, optimize, analyze, maintain, and evolve these documents *as if they were programs* using well-studied techniques from the field of *software engineering*. With the help of a running example from tax law, we propose several applications of software engineering to legal reasoning:

- Streamlined interpretation using *subroutine inlining*.
- Example generation using *coverage analysis* and *mutation testing*.
- Edge-case analysis using *property-based testing*.

²All LLM prompts and corresponding LLM responses related to examples listed in this paper are available at <https://github.com/cmu-pasta/onward24>.

You are an expert legal reasoning system, capable of reading legal statutes and applying them to given scenarios by identifying the correct subset of sections that apply to the particular situation.

In this conversation, we are only going to refer the following statutes. Do not use your knowledge of any other piece of legislation not provided in this conversation. If any undefined section is referenced in a question, please refuse to answer the question. The statutes referenced in this conversation are as follows:

```
...
<Excerpts from the IRC as provided verbatim in Appendix A>
...
```

Figure 1. LLM system prompt for all examples in this paper.

Our hope is that this paper opens up new research directions at the cross-sections of computational law, artificial intelligence, and software engineering. We anticipate that such research will enable the development of enhanced search and analysis tools to assist lawmakers and citizens alike in reasoning about complex legal documents.

2 Interpreting Substitutions via Inlining

One of the challenges with using LLMs for automated deductive legal reasoning is the cross-application of other sections with substituted values. For example, consider the task of determining the standard deduction for taxable year 2025, which requires calculating cost-of-living adjustments using hypothetical inflation data provided in Appendix B.³

Example 2. *Alice and Bob are married and filing taxes jointly for the year 2025. What is the basic standard deduction available to them?*

Solving this task correctly requires several considerations:

1. For joint returns, §63(c)(2)(A)(i) declares the basic standard deduction to be 200% of the value in §63(c)(2)(C), which is stated to be \$3,000.
2. §63(c)(4) applies “adjustments for inflation” to the value in §63(c)(2)(C), but later §63(c)(7)(B)(i) overrides this substitution by declaring that “Paragraph (4) shall not apply” for 2025.
3. For 2025, §63(c)(7)(A)(ii) replaces the value “\$3,000” in §63(c)(2)(C) with the value “\$12,000”.
4. As per §63(c)(7)(B)(ii), the value “\$12,000” in §63(c)(2)(C) is further increased using a cost-of-living adjustment as per §1(f)(3), but applying the latter rule itself with a substitution of “2017” for “2016”.

³At the time of writing (June 2024), the inflation data for calendar year 2024 is not yet fully known and consequently the standard deduction for 2025 has not yet been published by the IRS; so, we are confident that an LLM will not be able to use external knowledge to solve this problem.

```

<... System prompt (ref. Figure 1) ... >
<... Text of problem (e.g., Example 2) ... >

Let's think step by step as to which rules apply, and stop
if you come across an applicable rule R that requires
substitution of some value or phrase X in place of Y in
rule S. In this case, do not attempt to apply any other
rules or compute the final answer. Instead, I only want to
perform a single substitute. Please only return a modified
version of the relevant statute S with the substitution
performed. End your response with the following:

Modified rule: S
Modified by: R
Updated text for rule S: <new text replacing Y with X>

```

Figure 2. LLM Prompt incorporating subroutine inlining.

5. §1(f)(3)(A) provides a formula for computing a cost-of-living adjustment using the C-CPI-U value for the preceding calendar year (i.e., 2024) and the CPI for calendar year 2016 multiplied by the ratio of the C-CPI-U for 2016 to the CPI for 2016 as in §1(f)(3)(B). However, §1(f)(3)(C) comes into play, changing the calculation because we are using “2017” instead of “2016”, as determined in the previous step. Now, it appears we must only use the C-CPI-U values of 2024 and 2017; no multiplication is needed.
6. Plugging in the values from Appendix B, the adjustment is calculated to be 24.82%.
7. Applying the adjustment and the rounding specified in §63(c)(7)(B)(ii), the value “\$12,000” from §63(c)(2)(C) is therefore determined to change to “\$14,950”.
8. Finally, stepping back to §63(c)(2)(A)(i), the basic standard deduction for a joint return is therefore 200% of this adjusted value. So, the result is \$29,900.

In our experiments, modern LLMs like Claude 3 Opus or GPT-4 struggle to solve this task correctly even with few-shot in-context learning [6] and using *chain-of-thought* reasoning [25]. There are simply too many *subroutines*, and the LLM’s reasoning inevitably falters by using a wrong value or applying it in the wrong place. The end result is almost always incorrect—see Appendix C.2 for a sample response where Claude produces a wrong answer.

Taking inspiration from *inlining* in software optimization [10], we devised a prompting strategy where we asked the LLM not to calculate the final answer, but just perform one step of the substitution and reproduce a modified version of the statute verbatim (ref. Figure 2). For example, after step 2 above, the LLM will return the text in Appendix A except with §63(c)(4) completely removed. After applying step 3 above, the modified version of the statute will say the following in place of §63(c)(2)(C) (emphasis added):

- (2) Basic standard deduction
 - ...
 - (B) **\$18,000** in the case of a head of household (as defined in section 2(b)), or
 - (C) **\$12,000** in any other case.

Continuing in this way, applying step 5 above changes the text in §1(f)(3)(A) to say (emphasis added):

- (3) Cost-of-living adjustment

For purposes of this subsection—

 - (A) In general

The cost-of-living adjustment for any calendar year is the percentage (if any) by which—

 - (i) the C-CPI-U for the preceding calendar year, exceeds
 - (ii) **the C-CPI-U for calendar year 2017.**

and so on. Using a modified version of the statute with *inlined* substitutions in each subsequent step allows the LLM to slow down its chain of thought. With this process, Claude 3 Opus is able to correctly solve Example 2 after seven iterations.⁴

3 Diverse Example Generation

Consider the task of generating one-paragraph example scenarios (as in Example 1) such that certain rules in the statute apply in specific ways. This could be useful for documentation or education—for instance, the IRS embeds hundreds of examples across its various publications and form instructions in order to demonstrate how federal tax law applies across various circumstances. Additionally, high quality examples can also aid in training language models to perform more accurate legal reasoning—researchers have found that augmenting training data with additional examples improves accuracy on some tasks in the COLIEE dataset [37].

In software engineering, the task analogous to *diverse example generation* is that of constructing a high-quality suite of test cases. Software tests are useful in many ways including validation (i.e., checking whether a program behaves correctly), specification (i.e., writing a test before implementing functionality, as in *test-driven development* [3]), and documentation (i.e., the test code demonstrates the behavior of a certain application programming interface (API)).

3.1 Code Coverage

In software testing, the quality of a test suite is usually measured using criteria that track the diversity of program behaviors captured across all tests. The most widely used example is that of *code coverage*, which measures the fraction of source code (often counted at the level of *lines* of code) exercised across all tests. Code coverage is often used as an objective

⁴For this exercise, we manually ran the “iterations” by prompting the LLM multiple times and performing the substitution of rule *S* and cleanup of rule *R* as suggested by the LLM in its response (ref. Fig. 2). In general, we expect that this step can be automated with simple text-processing scripts.

function in automated test-input generation tools, such as *coverage-guided grey-box fuzzing* [5].

Using LLMs, we can use the same principles as in software testing to measure the coverage of available examples and to generate more examples by targeting specific coverage criteria. For example, asking Claude 3 Opus to solve Example 1 with a coverage trace (ref. Appendix C.3) correctly produces a list of rules whose conditions were satisfied in this example: “§63(b), §63(c)(1), §63(c)(7), §63(c)(7)(A)(ii), §63(c)(2)(A), §63(c)(2)(C)”. Moreover, we can ask the LLM to synthesize examples such that a particular rule is exercised in its coverage. For instance, asking Claude for an example such that §63(f)(1)(B) applies but §63(f)(1)(A) does not, produces:

Example 3. *Charlie and Diane are married and filing taxes jointly for the year 2018. Charlie was born on 3/15/1985 and Diane was born on 8/22/1953. In 2018, Charlie and Diane’s adjusted gross income was \$185,000. Charlie and Diane do not itemize their deductions and do not qualify for any deductions other than the standard deduction.*

Explanation (as per Claude). “In this scenario, Diane has attained age 65 before the close of the 2018 tax year, so §63(f)(1)(B) applies, entitling Charlie to an additional standard deduction amount for his spouse. However, Charlie himself has not attained age 65 before the close of the 2018 tax year, so §63(f)(1)(A) does not apply”.

3.2 Mutation Testing

In software testing, a criticism of the *code coverage* measurement of test-suite quality is that it only cares about whether certain lines of source code were *exercised*, but not the meaning of those lines themselves. High-coverage test suites do not necessarily find more faults than low-coverage ones [22]. Instead, what we really want is for our tests to be able to *fail* when a software bug is introduced. *Mutation testing* [12, 32] measures test-suite quality by artificially injecting “bugs” and determining if tests fail on the buggy mutated programs. The objective of a test writer is then to produce new tests that would distinguish the behavior of the original program from that of the mutation. This idea has also been used to automate test-input generation with the objective of maximizing mutation-detection ability [16, 36].

We can apply the same principles as in software mutation testing to generate examples that not only exercise certain legal rules, but also distinguish them from other incorrect rules. For example, we can ask an LLM to assume that §63(f)(1)(B) is modified to remove the second clause, as follows:

(B) for the spouse of the taxpayer if the spouse has attained age 65 before the close of the taxable year ~~and an additional exemption is allowable to the taxpayer for such spouse under section 151(b).~~

Then, we ask the model to synthesize an example such that §63(f)(1)(B) would apply with the modified rule but it would

not apply with the original rule. Claude 3 Opus responds to this challenge with the following example:

Example 4. *Ellie and Frank are married, but they are filing their taxes separately for the year 2018. Ellie was born on 5/10/1980 and Frank was born on 2/28/1952. In 2018, Ellie’s adjusted gross income was \$75,000 and Frank’s adjusted gross income was \$50,000. Neither Ellie nor Frank itemize their deductions and they do not qualify for any deductions other than the standard deduction.*

Explanation (as per Claude). “The modified rule [...] would apply to Frank because he has attained age 65 before the close of the 2018 tax year. This means Frank would be entitled to an additional standard deduction amount for being aged 65 or older, even though he and Ellie are filing separately. However, under the original rule, §63(f)(1)(B) would not apply to Frank because it requires not only that the spouse has attained age 65, but also that “an additional exemption is allowable to the taxpayer for such spouse under section §151(b).” [...] In this scenario, Frank does have gross income, so even though he is over 65, the original §63(f)(1)(B) would not apply to him because Ellie cannot claim an additional exemption for him under §151(b).”

Mutation testing thus helps with generating examples that not only exercise specific parts of relevant statutes but also provide a basis for reasoning about why certain rules are written in a certain way.

4 Metamorphic Property-Based Testing

Property-based testing, originally introduced in *Quickcheck* for Haskell [8], is a technique for writing generic tests that express what properties a function should satisfy given *any* imaginable input. Given a function $f : X \rightarrow Y$ and a predicate $p : X \times Y \rightarrow \text{bool}$, we want to ensure that $\forall x \in X : p(x, f(x))$. For example, given a `sort` function, we might want to ensure that the output is always sorted and contains a permutation of its input elements. The key idea of property-based testing is to check such generic properties using auto-generated inputs values via random sampling of the input space X .

A special class of property-based testing checks properties that should hold across multiple invocations of some program function when the inputs are changed in a specified manner—this is called *metamorphic testing* [7]. For example, given a `sort` function, we can check whether `sort(x)` and `sort(reverse(x))` are equal for all values x .

Meta-reasoning tasks in statutory or contract law often deal with ensuring that the legal text satisfies certain consistency properties. For example, a desirable property of a tax legislation may be that the same income dollar is not taxed twice, or that the same expense dollar is not allowed as a deduction in two different ways. In estate planning, consider a complex *last will and testament* having many conditional

branches to decide how assets should be distributed depending on which beneficiaries are alive at the time of the will’s execution; the testator might want to ensure that beneficiary X always receives more than beneficiary Y regardless of the circumstances and how the conditional branches are resolved. Such meta-reasoning is thus a form of edge-case analysis.

We argue that meta-reasoning tasks can potentially be automated using the key principles of metamorphic property-based testing and combining these with AI-driven techniques of interpretation. For example, consider §63(c)(7)(B) from the IRC (ref. Appendix A), which provides adjustments for inflation to the basic standard deduction applicable in tax years 2018–2025. One may want to analyze the code to ask: *Can the basic standard deduction available for single filers ever decrease year-over-year in this period?* Perhaps a legislator or other stakeholder is interested in understanding the effects of year-over-year deflation on the tax code. Answering this question involves not just interpreting the statutes but also reasoning about how they would apply in hypothetical edge-case scenarios.

Unfortunately, asking an LLM to analyze general properties of legal statutes directly does not always produce accurate results. When directly prompted with the aforementioned question, Claude 3 Opus determines “§63(c)(7)(B)(ii) states that for taxable years beginning after 2018, these \$18,000 and \$12,000 amounts shall each be increased by a cost-of-living adjustment. [...] Since the amounts can only be increased and not decreased according to this special rule for 2018–2025, the basic standard deduction *cannot decrease* year-over-year within this period.” (emphasis added). However, it turns out that this is not fully correct.

We can help Claude by expressing our question as a metamorphic property-based test to verify the *monotonicity* of the inflation adjustments—Given any two years X and Y , such that $2018 \leq X < Y \leq 2025$, and hypothetical C-CPI-U values for the correspondingly preceding calendar years I_x and I_y in the numeric range $[100, 200]$, calculate $D(X, I_x)$ and $D(Y, I_y)$ respectively, where function D is the computation of basic standard deduction for single filers in a given year assuming a given C-CPI-U value for the preceding calendar year; then check whether $D(X, I_x) \leq D(Y, I_y)$ is true. Now do this for random values of X, Y, I_x, I_y in a loop—in each iteration, use the LLM to calculate the inflation adjustment using the prompts shown earlier.

It takes about 2.4 iterations on average to discover that it is indeed possible to violate the stated property. Although the adjustments defined in §63(c)(7)(B) and §1(f)(3) do indeed only allow for *increases* to the basic standard deduction value listed in §63(c)(2)(C) (which is “\$12,000” in the period 2018–2025), the amount by which this value is increased can seemingly decrease year-over-year if there is deflation during those subsequent years. Property-based testing can reveal a

situation where the monotonicity property is violated: for instance, $D(2024, 168.1) = \$14,550$ and $D(2025, 167.0) = \$14,500$; that is, if there was a hypothetical 0.65% deflation in 2024, then the basic standard deduction for single filers in tax year 2025 as computed by §63(c)(2)(C) and §1(f)(3) appears to be \$50 lower than that computed for the previous year.

Property-based testing thus allows us to reason about invariants across many hypothetical scenarios.

5 Other Software Engineering Applications

Given our view of legal reasoning through the lens of software engineering, one can imagine applications of a number of other software analysis techniques.

Example Minimization. Delta debugging [38] is a technique for isolating the root cause of a program failure in terms of input data (e.g., think of a large file that causes a word processor to crash). The key idea is to repeatedly slice and dice the input until we obtain the smallest possible chunk that also reproduces the failure.

In legal reasoning or meta-reasoning, delta debugging can be applied to any task where the output is an example scenario (ref. Sections 3 and 4), and the objective is to find the *smallest example such that a particular rule applies in a specific way*. We only need to implement support for input reduction operators, which for natural language inputs (such as Examples 1–4) might mean deleting individual sentences or logical clauses separated by conjunctions. After that, input minimization via the *dmin* algorithm [38] is simply an application of interpretation (ref. Section 2) in a loop, with the use of coverage analysis (ref. Section 3.1) to determine whether reduced inputs exercise given rules (i.e., the *oracle* of reduction).

Amendment Validation. Consider the meta-reasoning task of understanding the scope of *amendments* to legislation or contracts, such as a bill introduced by the U.S. Congress that reforms the tax code, or a customized rider attached to an insurance policy.

In software, change impact analysis [2] is a technique for identifying which modules in a software get affected by a small patch applied to the program. The basic idea is to build a directed graph of various program modules and their interdependencies using both static source code information and dynamic hints from previous test executions. The impact of changes can then be identified by following paths through such a graph.

In legal reasoning, it is already possible to construct such static dependency graphs by mining explicit cross-references between various sections [11, 33]. With AI-driven interpretation and example generation with coverage analysis (ref. Section 3.1), we can also add test cases as nodes in this graph. An amendment that modifies existing rules could then be analyzed by re-running “test cases” whose coverage directly

or indirectly depends on the modified rule. Just like with the mutation testing approach (ref. Section 3.2), the goal here would be to identify example scenarios (i.e., the “test cases”) which are evaluated differently under the original and the amended rules.

Amendment Composition. A stretch goal for using AI in legal analysis would be to prototype new amendments that have a specific impact on selected input scenarios.

In software engineering, the field of *automatic program repair* (APR) [28] develops techniques for fixing bugs by synthesizing appropriate patches. In a classical formation, APR starts with a program and a suite of tests such that one bug-revealing test fails. The task is to find a software patch such that this test and all the other test cases in the original suite pass with the new program. The key idea behind search-based APR techniques is to search over a space of small changes to the program until all the tests pass in the expected way.

Generative AI could potentially be used to brainstorm amendments to legal texts (the “patches”) having a desired effect by providing natural language test scenarios such that certain rules should apply in desirable ways. We expect this application to be more useful in prototyping changes to contracts between business entities rather than drafting legislative bills.

6 Discussion

6.1 Threats to Validity

A key challenge with using LLMs to perform any sort of reasoning or analysis task is to ensure the reliability of their responses. LLMs are notorious for sometimes hallucinating [29] or providing incorrect responses for trivial reasoning tasks [4]. While we do not offer any silver bullets in this paper, we note that this is a problem being actively researched by the AI community [23, 35]; mitigation strategies include extracting explanations for traceability [9] and refining answers via self-reflecting feedback loops [13, 24].

One common failure mode we have observed pertains to cases when the LLM is pre-trained on legal documents from the domain. For example, when working with U.S. tax law, we can be certain that commercial LLMs such as Claude would have been trained on the official IRC statutes as well as thousands of web articles discussing contemporary tax issues. Consequently, we have found that applying techniques such as mutation testing or presenting questions about hypothetical new laws sometimes results in sub-optimal outcomes because the model continues to “believe” the current law instead of new rules or situations that we may provide in prompts. In future work, we plan to investigate a mitigation technique where key domain-specific terms in the statutes (such as “taxable income”) are replaced by domain-agnostic terms (such as “foobar”) in order to disconnect the model’s grounding in background knowledge.

6.2 Alternative Approaches and Related Work

In this paper, most of the examples considered directly prompt an LLM to perform numeric calculations based on natural language text in legal statutes. An alternative to this approach is to first prompt the LLM to synthesize an executable program that performs the same computation and then to execute such a program with inputs from a given scenario. We attempted to do so using Catala [30], a domain-specific programming language for formalizing statutes, as well as regular Python; however, our initial experience with this approach was quite unsatisfactory. One main problem we encountered is that legal statutes do not have a clear “entry point” and linear flow of logic; moreover, they do not upfront declare which terms may be re-interpreted by other sections. The task of manually transcribing statutes into Catala or Python programs often requires making multiple passes over the program to alter definitions. For example, when some rule R requires applying section S but replacing the term X in place of Y , we need to go back to the definition of S and convert constants or expressions into variables or subroutines, respectively, so that we may plug-in X as an input. In our experience, we found current LLMs to not be well-suited for this successive non-linear refinement of programs with standard prompting techniques, though future research may discover more effective solutions.

Holzenberger and Van Durme [20, 21] have proposed performing statutory reasoning by extracting the structure of legal statutes in the form of graphs and Prolog programs—the approach appears to be effective, but the representations are not designed to be easy for humans to read and debug. The software engineering techniques presented in our paper, such as coverage analysis, mutation testing, or property-based testing, could potentially be applied on top of these Prolog programs as well.

Metamorphic testing has previously been applied to tax preparation software [34]; it would be interesting to validate the metamorphic properties used in this prior work by applying them directly on the underlying natural-language statutes.

6.3 Implications and Aspirations of the Work

How will the techniques proposed in this paper apply in the real world? First, we want to clarify that we do not expect AI-driven legal analysis tools to completely replace any individual task performed by lawyers, primarily due to the lack of reliability in any statistical model. Instead, we anticipate that a software engineering approach to legal reasoning will enable the development of search & analysis tools that augment current legal processes that are performed predominantly manually.

For lawmakers and other legal professionals involved in drafting laws, policies, or contracts, SE+AI-enabled tools could provide useful insights into the breadth and limitations

of specific legal rules under consideration. Such tools can help drafters avoid producing inconsistent documents or making unintentional mistakes.

For ordinary citizens, legal analysis tools could potentially help provide useful insights that are only otherwise available to the more resourceful clients who can afford to retain legal experts. For example, we can imagine tools that can help individuals compare and demystify documents such as insurance policies, rental leases, and credit card agreements by providing “what-if” example generation and edge-case analysis capabilities; while not a substitute for legal advice, these insights could potentially form the basis for subsequent negotiation unlocked by the increased transparency.

7 Conclusion

Generative AI has unlocked new ways of automating deductive legal reasoning. The intricacies and peculiarities of legalese make it much more suitable to treat statutes and contracts as computer programs rather than as natural language databases of information. Consequently, we argue that this problem can benefit from the principled application of techniques well-studied in the field of software engineering such as code coverage, function inlining, mutation testing, metamorphic property-based testing, and more. We hope that this paper forms the basis for researchers to investigate further synergies across software analysis and computational law.

Acknowledgments

The author would like to thank the anonymous reviewers of the *Onward!* track for their valuable comments and suggestions for improvement. The paper also benefited from feedback provided by Sarah Lawsky, Travis Breaux, Sarah Santos, Vasudev Vikram, and Sam Estep. Finally, the author thanks Bogdan Vasilescu for encouraging him to pursue and publish these ideas.

A Excerpt from Internal Revenue Code

In this appendix, we list excerpts from 26 U.S.C. §63, §1, and §151 (as of April 2024) that are provided to an LLM for reasoning about tasks described in the sections above.

§63. Taxable income defined

- ...
- (b) Individuals who do not itemize their deductions
- In the case of an individual who does not elect to itemize his deductions for the taxable year, for purposes of this subtitle, the term “taxable income” means adjusted gross income, minus—
- (1) the standard deduction,
- ...
- (c) Standard deduction
- (1) In general
- Except as otherwise provided in this subsection, the term “standard deduction” means the sum of—
- (A) the basic standard deduction, and

(B) the additional standard deduction.

(2) Basic standard deduction

For purposes of paragraph (1), the basic standard deduction is—

- (A) 200 percent of the dollar amount in effect under subparagraph (C) for the taxable year in the case of—
- (i) a joint return, or
- (ii) a surviving spouse (as defined in section 2(a)),
- (B) \$4,400 in the case of a head of household (as defined in section 2(b)), or
- (C) \$3,000 in any other case.

(3) Additional standard deduction for aged and blind

For purposes of paragraph (1), the additional standard deduction is the sum of each additional amount to which the taxpayer is entitled under subsection (f).

(4) Adjustments for inflation

In the case of any taxable year beginning in a calendar year after 1988, each dollar amount contained in paragraph (2)(B), (2)(C), or (5) or subsection (f) shall be increased by an amount equal to—

...

(7) Special rules for taxable years 2018 through 2025

In the case of a taxable year beginning after December 31, 2017, and before January 1, 2026—

(A) Increase in standard deduction

Paragraph (2) shall be applied—

- (i) by substituting “\$18,000” for “\$4,400” in subparagraph (B), and
- (ii) by substituting “\$12,000” for “\$3,000” in subparagraph (C).

(B) Adjustment for inflation

(i) In general

Paragraph (4) shall not apply to the dollar amounts contained in paragraphs (2)(B) and (2)(C).

(ii) Adjustment of increased amounts

In the case of a taxable year beginning after 2018, the \$18,000 and \$12,000 amounts in subparagraph (A) shall each be increased by an amount equal to—

- (I) such dollar amount, multiplied by
- (II) the cost-of-living adjustment determined under section 1(f)(3) for the calendar year in which the taxable year begins, determined by substituting “2017” for “2016” in subparagraph (A)(ii) thereof.

If any increase under this clause is not a multiple of \$50, such increase shall be rounded to the next lowest multiple of \$50.

...

(f) Aged or blind additional amounts

(1) Additional amounts for the aged

The taxpayer shall be entitled to an additional amount of \$600—

- (A) for himself if he has attained age 65 before the close of his taxable year, and

(B) for the spouse of the taxpayer if the spouse has attained age 65 before the close of the taxable year and an additional exemption is allowable to the taxpayer for such spouse under section 151(b).

§1. Tax imposed

(f) Adjustments in tax tables so that inflation will not result in tax increases

(3) Cost-of-living adjustment

For purposes of this subsection—

(A) In general

The cost-of-living adjustment for any calendar year is the percentage (if any) by which—

- (i) the C-CPI-U for the preceding calendar year, exceeds
- (ii) the CPI for calendar year 2016, multiplied by the amount determined under subparagraph (B).

(B) Amount determined

The amount determined under this clause is the amount obtained by dividing—

- (i) the C-CPI-U for calendar year 2016, by
- (ii) the CPI for calendar year 2016.

(C) Special rule for adjustments with a base year after 2016

For purposes of any provision of this title which provides for the substitution of a year after 2016 for “2016” in subparagraph (A)(ii), subparagraph (A) shall be applied by substituting “the C-CPI-U for calendar year 2016” for “the CPI for calendar year 2016” and all that follows in clause (ii) thereof.

(4) CPI for any calendar year

For purposes of paragraph (3), the CPI for any calendar year is the average of the Consumer Price Index as of the close of the 12-month period ending on August 31 of such calendar year.

(6) C-CPI-U

(B) Determination for calendar year

The C-CPI-U for any calendar year is the average of the C-CPI-U as of the close of the 12-month period ending on August 31 of such calendar year.

§151. Allowance of deductions for personal exemptions

(a) Allowance of deductions

In the case of an individual, the exemptions provided by this section shall be allowed as deductions in computing taxable income.

(b) Taxpayer and spouse

An exemption of the exemption amount for the taxpayer; and an additional exemption of the exemption amount for the spouse of the taxpayer if a joint return is not made by the taxpayer and his spouse, and if the spouse, for the calendar year in which the taxable year of the taxpayer begins, has no gross income and is not the dependent of another taxpayer.

B Inflation Data for Cost-of-Living Adjustments

The following table provides (hypothetical) values for the Chained Consumer Price Index for All Urban Consumers (C-CPI-U) for calendar years 2017–2024, calculated using the averaging methodology in §1(f)(6)(B).

| Year | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Value | 138.2 | 141.0 | 143.2 | 144.8 | 149.1 | 159.9 | 168.3 | 172.5 |

C Sample Transcripts

This section contains some samples from our interactions with Anthropic’s Claude 3 Opus LLM, when prompted with examples from the paper. The detailed transcripts are available at <https://github.com/cmu-pasta/onward24>.

C.1 Determining Standard Deduction without Inflation Adjustments

When given the system prompt (ref. Fig. 1) followed by a prompt containing Example 1 and the phrase “Let’s think step by step” for zero-shot chain-of-thought [25], Claude responds:

Okay, let's break this down step-by-step based on the given statutes:

1. Determine the standard deduction for Alice and Bob in 2018:
 - According to §63(c)(7)(A), for taxable years 2018 through 2025, the basic standard deduction for a joint return is \$24,000 (\$18,000 substituted for \$4,400 in § 63(c)(2)(B), and \$12,000 substituted for \$3,000 in § 63(c)(2)(C)).
2. Determine if Alice and Bob are entitled to any additional standard deduction amounts:
 - Per §63(f)(1)(A) and (B), taxpayers are entitled to an additional \$600 each if they have attained age 65 before the close of the taxable year.
 - In 2018, Alice was 37 years old (born in 1981), and Bob was 43 years old (born in 1975). Neither had attained age 65.
 - Therefore, they are not entitled to any additional standard deduction amounts.
3. Calculate their total standard deduction:
 - Their total standard deduction is the basic standard deduction of \$24,000, as they do not qualify for any additional standard deduction amounts.
4. Calculate their taxable income:
 - According to §63(b), for individuals who do not itemize their deductions, taxable income means adjusted gross income minus the standard deduction.

- Alice and Bob's adjusted gross income in 2018 was \$216,350.
- Their taxable income = \$216,350 - \$24,000 = \$192,350.

Therefore, Alice and Bob's taxable income for 2018 is \$192,350.

C.2 Calculating Inflation Adjustments to Standard Deduction without Inlining

When asked to solve Example 2 with only the system prompt (ref. Fig. 1), inflation data (ref. Appendix B), and using zero-shot chain-of-thought [25], Claude responds:

Okay, let's break this down step-by-step based on the information provided in the statutes:

- Determine the basic standard deduction:
 - According to §63(c)(7)(A), for taxable years 2018 through 2025, the basic standard deduction for a joint return is \$18,000.
- Adjust the basic standard deduction for inflation:
 - §63(c)(7)(B)(ii) states that for taxable years beginning after 2018, the \$18,000 amount shall be increased by the cost-of-living adjustment determined under §1(f)(3), substituting "2017" for "2016" in §1(f)(3)(A)(ii).
 - The cost-of-living adjustment for 2025 is calculated as follows:
 - * C-CPI-U for 2024 (preceding year): 172.5
 - * C-CPI-U for 2017: 138.2
 - * Adjustment = $(172.5 / 138.2) - 1 = 0.2482$ (24.82%)
 - Increase the \$18,000 basic standard deduction by 24.82%:
 - * $\$18,000 \times (1 + 0.2482) = \$22,467.60$
 - Round down to the nearest multiple of \$50:
 - * Adjusted basic standard deduction = \$22,450
- Determine the additional standard deduction:
 - According to §63(f)(1)(A) and (B), taxpayers are entitled to an additional \$600 each if they have attained age 65 before the close of the taxable year.
 - In 2025, Alice is 44 years old, and Bob is 49 years old, so they are not entitled to the additional standard deduction.

Therefore, for the tax year 2025, Alice and Bob's standard deduction is \$22,450.

This answer is incorrect. Although Claude correctly calculates the inflation adjustment of 24.82%, it multiplies this with a figure of \$18,000 (corresponding to the basic standard deduction for a head-of-household) instead of multiplying by \$12,000 and then doubling it for a joint return.

C.3 Coverage Tracing

The following is a user prompt for coverage tracing on Example 1—the usual system prompt containing the statutes (ref. Fig. 1) was also provided.

Alice and Bob are married and filing taxes jointly for the year 2018. Alice was born on 1/1/1981 and Bob was born on 12/30/1975. In 2018, Alice and Bob's adjusted gross income was \$216,350. Alice and Bob do not itemize their deductions and do not qualify for any deductions other than the standard deduction. What is their taxable income for 2018?

Let's think step by step. In each step, reference the section or paragraph number and justify your reasoning. If you encounter any cross-references to other sections or paragraphs, perform the reasoning in a sub-list and return to the previous step. Do this recursively for any subsequent cross-references. If you encounter any substitutions of values, perform the substitution and reproduce the substituted sentences verbatim before deciding if they actually apply. End your response with two special lines as follows:

Answer: <dollar amount calculated>

Coverage: <comma-separated list of rules referenced, identified by their most specific section/paragraph numbers>

Relevance: <comma-separated list of rules that applied to this case because the conditions were true, again identified by their most specific section/paragraph numbers>

Claude's response:

Okay, let's break this down step-by-step:

- According to §63(b), since Alice and Bob do not itemize their deductions, their taxable income is their adjusted gross income minus the standard deduction.
 - They have an adjusted gross income of \$216,350.
 - We need to calculate their standard deduction.
- §63(c)(1) says the standard deduction is the sum of the basic standard deduction and the additional standard deduction.
- For the basic standard deduction, §63(c)(2) provides dollar amounts, but §63(c)(7) says for a taxable year beginning after December 31, 2017 and before January 1, 2026, paragraph (2) shall be applied with some substitutions.
 - §63(c)(7)(A)(i) says to substitute "\$18,000" for "\$4,400" in §63(c)(2)(B).
 - After substitution, §63(c)(2)(B) reads: "\$18,000 in the case of a head of household (as defined in section 2(b))". This does not apply as Alice and Bob are filing jointly, not as head of household.
 - §63(c)(7)(A)(ii) says to substitute "\$12,000" for "\$3,000" in §63(c)(2)(C).
 - After substitution, §63(c)(2)(C) reads: "\$12,000 in any other case".

3. §63(c)(2)(A) says the basic standard deduction is 200% of the amount in §63(c)(2)(C) for a joint return.
 - Since §63(c)(2)(C) is \$12,000 after the substitution, the basic standard deduction for Alice and Bob is \$24,000.

4. For the additional standard deduction, §63(c)(3) says it is the sum of the additional amounts the taxpayer is entitled to under §63(f).

1. §63(f)(1)(A) provides an additional amount of \$600 if the taxpayer has attained age 65 before the close of the taxable year. Bob was born on 12/30/1975, so he was 43 in 2018 and this does not apply. Alice was born on 1/1/1981, so she was 37 in 2018 and this also does not apply.
2. §63(f)(1)(B) provides an additional \$600 for the spouse of the taxpayer if the spouse has attained age 65 and an additional exemption is allowable under §151(b). Neither condition is met (Bob is under 65, and they are filing jointly so §151(b) does not provide an additional exemption). So this also does not apply.
 - Therefore, Alice and Bob are not entitled to any additional standard deduction under §63(f).

5. The total standard deduction is thus just the basic standard deduction of \$24,000.

6. Going back to §63(b), the taxable income is the adjusted gross income of \$216,350 minus the standard deduction of \$24,000.

Therefore:

Answer: \$192,350

Coverage: §63(b), §63(c)(1), §63(c)(2), §63(c)(7), §63(c)(7)(A)(i), §63(c)(7)(A)(ii), §63(c)(2)(A), §63(c)(2)(C), §63(c)(3), §63(f), §63(f)(1)(A), §63(f)(1)(B), §151(b)
 Relevance: §63(b), §63(c)(1), §63(c)(7), §63(c)(7)(A)(ii), §63(c)(2)(A), §63(c)(2)(C)

References

- [1] Anthropic. 2024. Claude. <https://www.anthropic.com/claude>.
- [2] Robert S Arnold. 1996. *Software change impact analysis*. IEEE Computer Society Press.
- [3] Kent Beck. 2003. *Test-driven Development: By Example*. Addison-Wesley Professional.
- [4] Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can GPT-3 perform statutory reasoning? *arXiv preprint arXiv:2302.06100* (2023).
- [5] Marcel Böhme, Van-Thuan Pham, and Abhik Roychoudhury. 2016. Coverage-based greybox fuzzing as markov chain. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 1032–1043.
- [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [7] T.Y. Chen, S.C. Cheung, and S.M. Yiu. 1998. *Metamorphic Testing: A New Approach for Generating Next Test Cases*. Technical Report HKUST-CS98-01. The Hong Kong University of Science and Technology.
- [8] Koen Claessen and John Hughes. 2000. QuickCheck: A Lightweight Tool for Random Testing of Haskell Programs. In *Proceedings of the 5th ACM SIGPLAN International Conference on Functional Programming (ICFP)*.
- [9] Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. *arXiv preprint arXiv:2104.08661* (2021).
- [10] Jack W. Davidson and AM Hollersnm. 1992. Subprogram inlining: A study of its effects on program execution time. *IEEE Transactions on Software Engineering* 18, 2 (1992), 89.
- [11] Emile De Maat, Radboud Winkels, and Tom Van Engers. 2006. Automated detection of reference. In *Proceedings of the 2006 conference on Legal Knowledge and Information Systems: JURIX 2006: The Nineteenth Annual Conference*, Vol. 152. 41.
- [12] R. A. DeMillo, R. J. Lipton, and F. G. Sayward. 1978. Hints on Test Data Selection: Help for the Practicing Programmer. *Computer* 11, 4 (1978), 34–41. <https://doi.org/10.1109/C-M.1978.218136>
- [13] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495* (2023).
- [14] Phoebe C. Ellsworth. 2005. Legal Reasoning. In *The Cambridge Handbook of Thinking and Reasoning*, K. J. Holyoak and R. G. Morrison Jr. (Eds.). Cambridge University Press, New York, 685–704.
- [15] Damien Falato. 2013. Is it a trade or business? Or a hobby or investment? <https://www.thetaxadviser.com/newsletters/2023/may/is-it-a-trade-or-business-or-a-hobby-or-investment.html>. Online; Retrieved March 9, 2024.
- [16] Gordon Fraser and Andreas Zeller. 2010. Mutation-driven generation of unit tests and oracles. In *Proceedings of the 19th international symposium on Software testing and analysis*. 147–158.
- [17] Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2023. Summary of the competition on legal information, extraction/entailment (COLIEE) 2023. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*. 472–480.
- [18] Oliver Goodenough and Susan Salkind. 2022. Computable Contracts and Insurance: An Introduction. *MIT Computational Law Report* (August 2022). <https://law.mit.edu/pub/computablecontractsandinsuranceanintroduction>.
- [19] Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. *arXiv preprint arXiv:2005.05257* (2020).
- [20] Nils Holzenberger and Benjamin Van Durme. 2021. Factoring Statutory Reasoning as Language Understanding Challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2742–2758. <https://doi.org/10.18653/v1/2021.acl-long.213>
- [21] Nils Holzenberger and Benjamin Van Durme. 2023. Connecting symbolic statutory reasoning with legal information extraction. In *Proceedings of the Natural Legal Language Processing Workshop 2023*. Association for Computational Linguistics, 113–131.
- [22] Laura Inozemtseva and Reid Holmes. 2014. Coverage is not strongly correlated with test suite effectiveness. In *Proceedings of the 36th international conference on software engineering*. 435–445.
- [23] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

- [24] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 1827–1843.
- [25] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.
- [26] Sarah B Lawsky. 2016. Formalizing the code. *Tax L. Rev.* 70 (2016), 377.
- [27] Sarah B Lawsky. 2017. A logic for statutes. *Fla. Tax Rev.* 21 (2017), 60.
- [28] Claire Le Goues, Michael Pradel, Abhik Roychoudhury, and Satish Chandra. 2021. Automatic program repair. *IEEE Software* 38, 4 (2021), 22–27.
- [29] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661* (2020).
- [30] Denis Merigoux, Nicolas Chataing, and Jonathan Protzenko. 2021. Catala: a programming language for the law. *Proceedings of the ACM on Programming Languages* 5, ICFP (2021), 1–29.
- [31] John J Nay, David Karamardian, Sarah B Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H Choi, and Jungo Kasai. 2024. Large language models as tax attorneys: a case study in legal capabilities emergence. *Philosophical Transactions of the Royal Society A* 382, 2270 (2024), 20230159.
- [32] Mike Papadakis, Marinos Kintis, Jie Zhang, Yue Jia, Yves Le Traon, and Mark Harman. 2019. Mutation testing advances: an analysis and survey. In *Advances in Computers*. Vol. 112. Elsevier, 275–378.
- [33] Nicolas Sannier, Morayo Adedjouma, Mehrdad Sabetzadeh, and Lionel Briand. 2017. An automated framework for detection and resolution of cross references in legal texts. *Requirements Engineering* 22 (2017), 215–237.
- [34] Saeid Tizpaz-Niari, Verrya Monjezi, Morgan Wagner, Shiva Darian, Krystia Reed, and Ashutosh Trivedi. 2023. Metamorphic testing and debugging of tax preparation software. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*. IEEE, 138–149.
- [35] SM Tonmoy, SM Zaman, Viniya Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313* (2024).
- [36] Vasudev Vikram, Isabella Laybourn, Ao Li, Nicole Nair, Kelton OBrien, Raffaello Sanna, and Rohan Padhye. 2023. Guiding greybox fuzzing with mutation testing. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*. 929–941.
- [37] Masaharu Yoshioka, Yasuhiro Aoki, and Youta Suzuki. 2021. BERT-based ensemble methods with data augmentation for legal textual entailment in COLIEE statute law task. In *Proceedings of the eighteenth international conference on artificial intelligence and law*. 278–284.
- [38] Andreas Zeller and Ralf Hildebrandt. 2002. Simplifying and isolating failure-inducing input. *IEEE Transactions on Software Engineering* 28, 2 (2002), 183–200.

Received 2024-04-25; accepted 2024-08-08