

日本の刑事事件を対象とした法律相談システム

中下咲帆¹ 菊池英明¹ 藤倉将平² 則竹理宇²

¹早稲田大学人間科学研究科 ²LawFlow 株式会社

nakashitas@gmail.com kikuchi@waseda.jp

fujikura@lawflow.jp noritakebengo@keiji-bengosi.com

概要

本研究は、日本の刑法を対象とした法律相談システムの構築を目指して「罪名」「量刑」「量刑根拠」の3つの予測に取り組んだ。弁護士へのヒアリングを基に構築した罪名データベースを参照するように設計した罪名予測モジュールは、正解割合の平均が70%となり目標の80%には達しなかった。弁護士へのヒアリングを基に構築した量刑根拠データベースを参照するように設計した量刑根拠予測モジュールは、正解割合の平均が89%となり、目標の80%に達した。過去の量刑事例集を参照させるように設計した量刑予測モジュールは、正解割合の平均が43%となり、目標の80%を大きく下回った。

1 はじめに

1. 1 背景

2022年以降、ChatGPT^[1]をはじめとする大規模言語モデル(LLM)の進展により、法律分野でのLLM活用が期待されている。中国では幅広い法律相談に自動で回答するために、弁護士が行っている思考プロセスを反映させたLLMの開発研究^[2]が行われている。日本でも民間企業においてそのような取り組みが行われている事例^{[3][4]}はあるが、学術研究として公開されている取り組み事例は少ない。また、JURISIN^[5]やCOLIEE^[6]などの法的タスクを目的としたワークショップでは、司法試験や民事判決に関するデータを扱うタスク^{[7][8]}は存在するが、刑事事件に関するタスクはない。

1. 2 目的

本研究の目的は、日本の刑法を対象とした再現性のある法律相談システムの構築を通じて、刑事事件に関連する法的支援を提供することである。法的なトラブルは大きく民事と刑事に分かれるが、本研究

では類似の先行研究が少ない刑事事件に注目する。さらに複数のカテゴリが存在する刑事事件の中で件数が多く法的支援の需要が高い「交通」に関する罪に焦点を当てる。

具体的には、「罪名」「量刑」「量刑根拠」の3つの予測を行うことを目的とする。現状のLLMは類似する罪名の区別や妥当な量刑相場の予測において誤りを犯すことが多いため、本研究ではまずこの問題を解決し、よりの確な応答が可能な法律相談システムの構築を目指す。本研究で目的とする対話例を図1に示す。

事件内容

知り合いを殴ってしまい、顔に怪我を負わせてしまいました。相手は病院で診断書を貰っていて全治1週間だそうです。警察に捕まる可能性はありますか。



ユーザ

罪名

傷害罪で捕まる可能性があります。

量刑の根拠

今回は全治1週間という比較的軽微な怪我であるため

量刑

10万円程度の罰金刑が予想されます。

チャットボット

図1. 目的とする対話例

2 提案手法

2. 1 システム概要

本システムは、前述の目的にて示した各タスクに対応するために「罪名予測モジュール」「量刑根拠予測モジュール」「量刑予測モジュール」の3つの予測モジュールで構成されている。図2は提案システムにおける入出力の流れである。はじめにユーザからの法的な相談内容が罪名予測モジュールに入力され、罪名が出力される。続いて、相談内容と予測された罪名が量刑根拠予測モジュールに入力され、量刑根拠が出力される。最後に、相談内容と予測された罪名及び量刑根拠が量刑予測モジュールに入力され、量刑が出力される。これら3つのモジュール

で予測された罪名，量刑根拠，量刑の情報がまとめてユーザに返される。

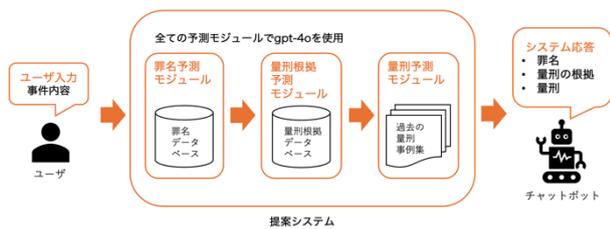


図 2. 提案システムの流れ

2. 2 罪名予測モジュール

罪名予測モジュールは，ユーザの事件内容を入力とし，それに対して最も可能性の高い罪名を3つ出力する．予測時のプロンプトでは，弁護士へのヒアリングを基に構築した罪名データベースを参照する形式になっている。

2. 2. 1 罪名の大分類

罪名は予測対象のカテゴリが多いため，大分類と詳細分類の2段階に分かれている．罪名データベースも各段階で分かれている．指示文は「ユーザが入力する法的案件と次の罪名データベースを照らし合わせて，確率の高い罪をJSON形式で出力してください」とし，大まかな罪名カテゴリを特定するための罪名データベースを参照させる．大分類のための罪名データベースには13カテゴリが含まれている．出力は，罪名データベースの中から選ばれた可能性の高い罪名カテゴリ（「交通に対する罪」など）となっている．複数のカテゴリが該当する場合には複数のカテゴリを出力する。

2. 2. 2 罪名の詳細分類

続いて，2段階目の詳細分類に進む．指示文は「ユーザが入力する法的案件と次の罪名データベースを照らし合わせて，確率の高い罪をJSON形式で確率の高い順に3つ出力してください」とし，詳細な罪名を特定するための罪名データベースを参照させる。

詳細分類のための罪名データベースは大分類のカテゴリ毎に12種類が存在する．出力は，罪名データベースに記載された詳細な罪名の中から，可能性の高い3つの罪が選ばれる形になっている。

2. 2. 3 最新罪名への対応

法改正によって改名された罪名が，罪名予測時に旧罪名のままで出力されてしまう課題があった．これに対応するために，新罪名を検索する過程を追加した．検索にはGoogle検索APIを使用し，「〇〇罪の最新罪名を検索してください」というクエリを指定した．抽出対象は，記事のタイトル，URL，スニペットの3つである．スニペットとは，Webページの要約文や説明文を指す．出力される記事の上限は10件である．続いてこの検索結果を元に，最新罪名の予測をgpt-4oで行う．検索結果を参照情報として，対象罪名の最新罪名を予測させる．プロンプトは「〇〇罪の最新罪名のみを予測してください」とした．これらの過程を踏まえて，最新罪名の確認を行い，それを罪名予測の最終的な出力とする．

2. 3 量刑根拠予測モジュール

量刑根拠予測モジュールは，ユーザの事件内容と罪名予測の結果を入力とし，量刑の根拠に影響する項目を出力する．指示文は「ユーザが入力する法的案件と次の量刑根拠データベースを照らし合わせて，量刑の判断根拠を出力してください」というような形で，弁護士へのヒアリングを基に構築した量刑根拠データベースを参照させる．量刑根拠データベースのカテゴリは「交通に関する罪」とし，このカテゴリに含まれる要素数は24となっている。

2. 4 量刑予測モジュール

量刑予測モジュールは，ユーザの事件内容と罪名予測及び量刑根拠予測の結果を入力とし，確率の高い量刑を2つ出力する．予測時の指示文は「ユーザが入力する量刑の相場表を参照して，根拠を示しながら具体的な量刑を予測してください」とし，過去の量刑事例集^{[9][10][11]}を参照させる．「交通に関する罪」の事例数は1003件である．量刑事例集はカテゴリ毎にテーブルが分かれているため，罪名予測の結果を元に該当カテゴリのテーブルだけを参照させる．出力は，量刑事例集を参照した上での相場感と，それを踏まえた量刑予測結果で構成されている。

3 評価

評価は「交通に関する罪」を対象とする。

3.1 本文テストデータの作成

テストデータには過去の裁判例を使用する。テストデータは、罪名データベースの「交通に関する罪」に含まれる全罪名の中から、裁判例が入手できた16罪名分のデータ16件である。第二審以降の判例は情報が省略されているケースが見られたため、情報が構造的かつ網羅的に記載されている第一審の判例のみを集めた。

続いて、収集した過去の裁判例から、テストデータとなる「事件内容」「罪名」「量刑」「量刑根拠」の4項目にあたる情報を Claude-3.5-Sonnet で抽出する。判例の中で、これらの4項目に対応する情報を表1に示す。判例には書誌情報と全文が存在するため、それぞれから抽出できる情報を分けて示す。

表1. 裁判例からのテストデータ抽出

テストデータの構成要素	書誌から抽出する情報	全文から抽出する情報
事件内容	要旨	罪となるべき事実
罪名	事件名/ 参照法令	法令の適用
量刑	裁判結果	主文
量刑根拠	要旨	量刑の理由

3.2 罪名予測の評価

罪名予測の評価は、罪名予測モジュールで最終的に予測された詳細な罪名に対して行う。予測された罪名の中に、正解の罪名が含まれている割合を評価する。この割合がテストデータの平均において80%を超えることを目標とする。

罪名予測の正解割合の平均は70%となり、目標の80%には達しなかった。3つのモジュールそれぞれについて、予測のミスケースをみていく。

まず、罪名の大分類については、不特定多数への被害があるケースに適用される「公共安全に対する罪」の出力がミスケースとして目立った。このように間違っ

たことで判別の精度が向上するのではないかと考える。

続いて、罪名の詳細分類のミスケースを以下に整理する。

1. 過失運転致死傷と危険運転致死傷の判別ミス
2. 罪名データベースに含まれる罪名が正式な罪名ではなく通称や総称の罪名になってしまっていたケース
3. 罪名データベースに記載されていなかったマイナーな罪名の予測ができていなかったケース
4. 最新罪名の検索で「検索できません」という出力になってしまったケース
5. 最新罪名の検索で旧罪名に対応しない全く別の罪名に変換してしまっ
6. 最新罪名の検索で旧罪名を旧罪名のまま出力してしまっ

1について、正確な判別は弁護士でも意見が分かれるほど難しいタスクであるため、初歩的な対応策として、それぞれの罪名の法条をプロンプトに記載し、ルールベースで参照させることが考えられる。

2・3への対応については、罪名データベースの構築基準を修正する必要がある。罪名は通称ではなく正式名称で記載することや、記載する罪名の粒度の調整が挙げられる。

4・5・6については、最新罪名の検索モジュールが期待通りに機能していない結果を示してしまっているため、最新罪名への対応方針を Google 検索から辞書検索に切り替えるべきだと考える。旧罪名と最新罪名のペアデータを含むような辞書を作成し、そちらを検索することで今よりも適切な変換が可能になるだろう。

3.3 量刑根拠予測の評価

量刑根拠予測では、テストデータの事件内容と前段で予測された罪名を入力とし、予測された量刑根拠と正解の量刑根拠を著者が比較評価する。予測された量刑根拠の各項目に、正解の項目が含まれている割合を評価する。この割合がテストデータの平均において80%を超えることを目標とする。

量刑根拠予測の正解割合の平均は89%となり、目標の80%に達した。評価値が目標を超えることができたのは、予測時に参照した量刑根拠データベースが網羅的に構築できていたからだ

からの確認をもらいながら、複数回のテストを繰り返し最適なデータベースの作成を行なった。

3.4 量刑予測の評価

量刑予測では、テストデータの事件内容と前段で予測された罪名及び量刑根拠を入力し、予測された量刑が弁護士の判断する量刑の範囲内に含まれる割合を評価する。前述した割合がテストデータの平均として80%を超えることを目標とする。

量刑予測の正解割合の平均は43%となり、目標の80%を大きく下回った。

ミスケースは16件中9件となった。ミスケース9件のうち、予測された量刑が正解よりも重くなったケースが5件、逆に正解よりも軽くなったケースが4件である。

正解と大きなズレのあるミスはいずれもシステムとして致命的であるが、実応用場面では正解よりも軽い量刑を予測してしまうケースの方が、ユーザにとっては問題であると考えられる。量刑予測におけるミスケースの原因は以下のパターンになっていた。

1. 罪名予測のミスが量刑予測のミスに繋がっているケース
2. 複数の罪名が正解となる場合に一部の罪名を正しく予測できないケース
3. 参照する過去の量刑事例集の情報量が少ないケース

1については、罪名予測で間違っただけで予測された罪名がそのまま量刑予測の入力として扱われているケースである。このケースへの対応は前述した罪名予測の改善になる。

2については、複数の罪が絡む併合罪のケースであるが、罪名予測の時点で、適用される罪が網羅的に予測できていないことで量刑を正しく加重できていない可能性がある。このケースへの対応としては、罪名予測の考察に記載した内容に重なるが、罪名データベースに記載する罪名を網羅的に拡充することが必要になる。

3については、今回プロンプトで参照している過去の量刑事例集において、該当罪名の事例がほとんどないケース(救護義務違反や酒酔い運転補助など)や、該当罪名の事例はあるが事件内容が簡略化されており、個別事例としての具体的な情報がほとんどないケースがあった。これらは参照情報がないという状態に等しいため、事例集の拡充が対応策となる。

4 おわりに

本研究では「罪名」「量刑」「量刑根拠」の3つの予測モジュールの構築に取り組んだ。各モジュールの正解割合を表2に示す。

表2. 各モジュールの正解割合

モジュール名	正解割合	目標割合
罪名予測モジュール	70%	80%
量刑根拠モジュール	89%	80%
量刑モジュール	43%	80%

罪名予測モジュール 罪名の大分類においては、主要なカテゴリである交通及び身体カテゴリの予測ができていた一方で、関係のないカテゴリの出力もあった。罪名の詳細分類においては、弁護士でも判断の難しい罪名の予測ミスや罪名データベースに記載のない罪名の予測ミスが目立った。また、最新罪名の検索モジュールが期待通りに機能しなかった。今後は、罪名データベースの修正、複雑な罪名に対するルールベースの規定、最新罪名の取得に向けた辞書の作成に取り組む。

量刑根拠予測モジュール 量刑根拠データベースが上手く機能している結果ではあったが、考察に記載した量刑根拠データベースの改善点を反映させることでさらに高精度な量刑根拠予測が可能になるだろう。

量刑予測モジュール 実際の量刑よりも軽い量刑を予測してしまうという致命的なミスがみられたため、罪名予測の改善に加え、参照する量刑事例集の拡充に取り組む必要がある。

参考文献

- [1] OpenAI. ChatGPT. (引用日:2025/01/09.).
<https://openai.com/blog/chatgpt>.
- [2] Jiayi Cui, et al. Chatlaw: A Multi-Agent Collaborative Legal Assistant with Knowledge Graph. Enhanced Mixture-of-Experts Large Language Model. arXiv:2306.16092v2 (引用日: 2025/01/09.)
- [3] 弁護士ドットコム. チャット法律相談(α). (引用日: 2025/01/09.) <https://chat.bengo4.com/>.
- [4] Legal AI. AI 弁護士法律相談ツール. (引用日: 2025/01/09.) <https://legalai.co.jp/#/>.
- [5] JURISIN2024. (引用日: 2025/01/09.).
<https://jurisinformaticscenter.github.io/jurisin2024/>.
- [6] COLIEE-2024 CALL FOR TASK PARTICIPATION. (引用日: 2025/01/09.).
<https://sites.ualberta.ca/~rabelo/COLIEE2024/>.
- [7] Hai-Long Nguyen, et al. Enhancing Legal Document Retrieval: A Multi-Phase Approach with Large Language Models. Proceedings of the 18th International Workshop on Juris-Informatics, pp.86-pp.100, 2024. (引用日: 2025/01/09.)
https://jurisinformaticscenter.github.io/jurisin2024/jurisin2024local_proceedings.pdf
- [8] Housam Khalifa Bashier Babiker, et al. Legal Yes/No Question Answering Through Text. Embedding, Fine-Tuning, and Prompt Engineering. Proceedings of the 18th International Workshop on Juris-Informatics, pp.170-pp.180, 2024. (引用日: 2025/01/09.)
https://jurisinformaticscenter.github.io/jurisin2024/jurisin2024local_proceedings.pdf
- [9] 第一東京弁護士会, 刑事弁護委員会. 量刑調査報告集Ⅲ, pp.12-pp.22, 2010.
- [10] 第一東京弁護士会, 刑事弁護委員会. 量刑調査報告集Ⅳ, pp.13-pp.19, 2015.
- [11] 第一東京弁護士会, 刑事弁護委員会. 量刑調査報告集Ⅴ, pp.12-pp.16, 2018.