

法律推論データセット学習による推論に強いモデルの構築

大南英理^{1,3} 岡崎直観^{1,2,3}

¹ 東京科学大学 ² 産業技術総合研究所 ³ NII LLMC
{eri.onami, okazaki}@comp.isct.ac.jp

概要

大規模言語モデル (LLM) を法実務に適用するには、専門的な知識を身に付けさせるだけでなく、長文自由記述形式の入力に対する推論能力を向上させることが不可欠である。法律の推論は法的三段論法という推論形式で行われ、人間の実務家はロースクールや実務経験を通してこれを学ぶが、LLM の学習に利用可能なオープンソースの判例には明示的に記載されない。本研究では、長文自由記述形式の法律推論能力向上のため、法的三段論法による推論タスクである LegalReasoningBench を提案する。また、法律推論に必須の 4 要素による評価手法を提案する。実験では、本データセットで学習済みモデルが学習無しのモデルの性能を上回ることを確認し、本データセットの有効性を示す。

1 はじめに

大規模言語モデル (LLM) を法律実務に応用するためには、二択または多肢選択式の質問に正確に回答するだけでなく、長文かつ自由記述形式のタスクにおいても法律に関する推論能力を向上させる必要がある。法律における自然言語処理 (LegalNLP) 分野において、LLM 登場以前は、要約 [1, 2], 固有表現抽出 [3, 4, 5, 6], テキスト分類 [7] など、様々な自然言語処理タスクが研究されてきた。LLM 登場後は、オープンソースで利用可能な各国の法律の条文・判例を収集したベンチマークデータセットが提案されている [8, 9, 10, 11]。しかし、これらのデータセットでモデルを学習しても、長文かつ自由記述形式の法律推論を完全に行えるモデルは構築できていない。判例の結果予測タスク [12, 13, 14, 15] は、LegalNLP の中でも最も難しいタスクの一つである。その理由は、判例が、裁判所が争点となっている事実や原告と被告の双方の主張を入力として、**法的三段論法**、または **IRAC framework** [16, 17] と呼ばれる推論形式に従って法律推論を行った結果だからであ

り、法的三段論法と呼ばれる推論形式がロースクール教育や実務経験を通して教えられ、判例や法律の条文に明示的に記載されないためである。

ところで、近年 LLM の性能は様々なタスクにおいて急激に向上しており [18], 多数の学習手法が提案され [19, 20], 複雑な推論 [21] が可能になりつつある。法律分野においても長文の法律質問応答タスクが提案されている [22, 23, 24] が、法律における推論タスクとは、単なる暗記問題を長文で出力することではなく、法的三段論法に従った推論が必要になる。既存の法律推論能力を図るデータセットとして LegalBench [16] が提案されているが、二択または多肢選択式の問題で、法的三段論法のうち一部の観点のみを測ることができ、法律の推論に関するデータセットは限られている。そこで本研究では、長文自由記述形式の法律推論能力を向上させるため、日本法及び国際法を対象として、法的三段論法に従った推論データセットである LegalReasoningBench を提案する。本データセットは、推論問題の解答に加えて、法律三段論法に必要な 4 つの要素と、根拠として参照されるべき法律の条文や判例が全て網羅されている。また、評価手法として、本データセットの法的三段論法に必要な 4 つの観点毎に、正解解答とモデル生成結果の比較による評価を行うことを提案する。実験では、本データセットで学習することにより、モデルの性能が学習無しモデルを上回ることから、データセットの有効性を示す。

2 データセット構築

法律推論 法律推論とは、法的三段論法の形式で行われる論理推論のことである。法的三段論法を自然言語処理において定式化するため、LegalBench [16] では IRAC framework という名称で法律推論から結論を導く方法を説明し、法律推論は Issue, Rule, Application, Conclusion の 4 段階により行われると定義される。すなわち、(I) 法律上議論すべき論点を特定して以降の段階で解答すべき

図 1: LegalReasoningBench データセットの問題例, 及び, アノテーション例を示す.

モデル入力	<p>甲は、食料品店主Aに対し、「指定した口座に400万円振り込まなければ、商品に毒を入れるぞ。」と電話で脅し、現金の振込先としてB銀行C支店の自己名義の普通預金口座を指定した。やむなくAが2回に分けて振込手続を行ったところ、Aの手続ミスにより、同支店に開設され、預金残高が37万円であった乙の普通預金口座に振り込まれてしまった。その後、乙が、30万円を通帳を使って窓口で引き出したところ、なお残高が207万円となっていたので、誤振込みがあったことを知り、更に窓口で100万円を引き出した。</p>	<p><実験設定></p> <div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid black; padding: 5px; text-align: center;"> <p>ICL</p> <p>プロンプト+具体例: 例1 ...</p> </div> <div style="border: 1px solid black; padding: 5px; text-align: center;"> <p>SFT</p> <p>学習データ</p> </div> </div> <p><法的三段論法によるアノテーション></p> <table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <tr><td style="background-color: #e0f0ff;">Issue</td><td style="background-color: #e0f0ff;">法律上の問題の特定</td></tr> <tr><td style="background-color: #fff0e0;">Rule</td><td style="background-color: #fff0e0;">法規範の解釈</td></tr> <tr><td style="background-color: #e0ffe0;">Application</td><td style="background-color: #e0ffe0;">規範の事実への当てはめ</td></tr> <tr><td style="background-color: #e0ffe0;">Conclusion</td><td style="background-color: #e0ffe0;">結論の導出</td></tr> </table>	Issue	法律上の問題の特定	Rule	法規範の解釈	Application	規範の事実への当てはめ	Conclusion	結論の導出
Issue	法律上の問題の特定									
Rule	法規範の解釈									
Application	規範の事実への当てはめ									
Conclusion	結論の導出									
アノテーション	<p>乙が、誤振込み金100万円を窓口で引き出した行為につき、詐欺罪(246条1項)が成立しないか。</p> <p>1100万円の占有が銀行になければ、詐欺罪は成立しないので、100万円の占有が乙と銀行のいずれにあるか問題となる。この点、誤振込みがあった場合、口座名義人には当該金銭を引き出す正当な権限はなく、銀行は所定の手続きを経て振込人へ組戻しをすることになる。とすれば、誤振込みされた金銭の占有は銀行にあると解すべきである。よって、本件でも、銀行に占有があり、詐欺罪が成立しうる。では、甲は、誤振込みがあったことを黙っていたに不義ないことから欺く行為があったといえるか。この点、欺く行為とは、相手方が財産的処分行為をするに当たり判断の基礎となる重要事項を偽ることをいう。</p> <p>本件で、誤振込みがあったことを知れば、銀行は上記のような所定の手続きを行わなければならないから、100万円が誤振込金であるか否かは、Aに100万円を交付するかどうかの判断の基礎となる重要事項といえる。そして、乙には、誤振込みであることを銀行に伝える信義則(民法1条2項)上の告知義務があるといえ、乙がこれを秘していたことは、不作為によって偽ったといえる。よって、欺く行為にあたる。</p> <p>そして、銀行は錯誤に陥り、乙に100万円を「交付」しているので、100万円につき詐欺罪が成立する。なお、乙が100万円を引き出した時点で口座には乙の預金が7万円存在していたが、乙は、誤振込みがあったことを知ってわざわざ100万円を引き出すとしたのであるから、100万円全体につき詐欺罪が成立すると考える。</p> <p style="background-color: #e0ffe0;">以上から、乙に詐欺罪、窃盗罪の共同正犯(後述)が成立し、併合罪(235条)となる。</p>									

表 1: データセット分割. L_{*-char} は平均文字数, L_{*-word} 平均単語数を示す.

分割	# 件数	# 日	# 英	$L_{JP-char}$	$L_{JP-word}$	$L_{EN-char}$	$L_{EN-word}$	# 法源
Train	5,694	5,516	178	574.04	377.88	1915.77	306.12	7,656
Validation	500	484	16	585.97	385.48	2230.75	363.25	679
Test	500	486	14	557.23	366.45	1818.43	289.07	624
All	6,694	6,486	208	573.67	377.59	1933.53	309.38	8,959

問いを立てる, (R) Issue に解答するために関連する条文や裁判例を参照し, 条文や判例などの抽象的なルールを解釈により具体化する, (A) Rule において具体化したルールを, 現在問題になっている事実当てはめる, (C) 当てはめにより得られた結論から, Issue での問いに解答する, という4段階である. また, Rule の段階で参照する条文, 裁判例, 法律の一般原則などを法源と呼び, 日本を始めとする国では法律の条文が主に参照されるが, 英米法体系の国では関連する裁判例が直接参照されることが多い. 本データセットは, 日本法と, 判例を直接参照することが多い国際法を対象としているため, Rule 段階の全ての形式を網羅していると言える.

タスク 日本の司法試験または国際法の模擬裁判の問題を入力として, 法律三段論法による推論を出力するタスクを導入する. 図 1 は, 要約された入力の問題文, 及び, 期待される出力のアノテーション例を示す.

データ収集 日本の司法試験形式の問題文と解答, および国際法模擬裁判の過去問と提出書面の収集を行った. 司法試験形式の問題文収集では, 司法試験, 予備試験及び, 司法試験類似の問題文と解答の収集を, 13 名のアノテーターの協力により行った.

作成したデータの 94% は法科大学院在学中の学生により作成され, データの 50% 以上は司法試験合格者により作られた. 国際法模擬裁判形式の問題文収集では, 投資仲裁の FDI Moot という英語模擬裁判を対象にデータ収集を行った. 模擬裁判では, 裁判の事実概要, 関連する架空の条約, 各国のロースクールからの過去参加者の提出書面の過去データが Web 上で公開されており, Web 上からデータを収集した.

データ分割 収集した問題は, 1 つの問題文に対して IRAC framework の I, R, A, C の推論をするべき検討箇所が複数あるため, 1 つの問題文に対して 1 つの法律上の論点特定され, 推論が行えるように, 収集した解答を論点毎に分割し, 特定の論点について解答できるように問題文を修正した.

以上のアノテーションにより, 日本法と国際法を対象とした, 17 の法分野, 8,959 件の法源情報を含む, 6,694 件の質問解答ペアを作成した. 表 1 はデータセット統計を示す.

3 評価

法律推論タスクの評価においては, IRAC framework に従って 4 つの観点別に評価をするのが一般的である [16]. 先行研究では, I, R, A, C のいづ

表 2: 文脈内学習および教師ありファインチューニングでの各モデルの性能を示す。LLM-as-a-Judge 評価モデルに GPT-5 を用いたため、(*) は潜在的なバイアス含まれる可能性がある。

モデル	LLM-as-a-Judge				人手評価				法源
	Issue	Rule	Application	Conclusion	Issue	Rule	Application	Conclusion	
ゼロショット推論									
Qwen3-1.7B	39.2	5.0	9.0	23.2	32	4	8	8	3.8
Qwen3-8B	50.8	11.2	21.4	30.6	32	0	0	12	6.7
Qwen3-14B	53.0	16.2	27.0	35.6	52	4	4	8	10.6
gpt-oss-20b	53.8	22.6	26.6	37.4	40	4	0	8	8.9
文脈内学習									
Qwen3-1.7B	35.0	3.4	8.0	20.4	28	0	0	12	4.39
Qwen3-8B	48.8	13.2	22.2	30.8	36	4	8	8	7.46
Qwen3-14B	52.0	16.2	25.0	32.2	32	4	0	0	9.34
gpt-oss-20b	50.8	22.4	26.8	38.2	16	0	0	0	7.13
教師ありファインチューニング									
Qwen3-1.7B	38.6	4.4	10.2	20.6	60	24	12	16	3.6
Qwen3-8B	48.6	13.8	21.6	31.4	68	28	12	32	7.9
Qwen3-14B	93.6	26.6	37.2	43.0	88	80	52	36	33.5
gpt-oss-20b	55.2	25.0	30.0	38.4	68	24	12	20	7.2
商用モデル									
GPT-4o	61.4	22.2	30.6	29.2	32	4	8	4	18.7
GPT-5	83.2*	57.6*	60.2*	57.4*	48	4	0	20	23.7

表 3: LLM-as-a-Judge の結果を法律の分野別で示す。

法律分野	刑法	民法	商法	民事訴訟法	刑事訴訟法	憲法	行政法	労働法	国際法(英)	国際法(日)	知的財産法	民事実務基礎	国際私法	倒産法	刑事実務基礎	経済法	環境法
問題件数	1,781	1,071	734	588	569	540	430	313	208	101	81	78	64	44	38	30	24
教師ありファインチューニング																	
Qwen3-1.7B	16.15	12.31	18.14	24.49	21.71	21.62	23.57	12.5	0.0	33.33	29.55	25.0	50.0	25.0	31.25	25.0	0.0
Qwen3-8B	24.31	21.15	30.88	37.24	30.26	35.81	34.29	20.54	3.57	58.33	40.91	46.43	75.0	25.0	37.5	25.0	25.0
Qwen3-14B	54.34	45.77	46.08	54.59	46.71	58.78	43.57	55.36	5.36	64.58	54.55	35.71	62.5	25.0	62.5	75.0	25.0
gpt-oss-20b	32.29	31.54	37.25	47.96	38.16	52.7	40.0	35.71	0.0	60.42	36.36	46.43	62.5	25.0	37.5	50.0	25.0

れかの要素に関する二択または多肢選択式の問題を答えさせることにより推論能力を評価していたが、本データセットは、自由記述形式の解答が、それぞれ IRAC のどの構成要素にあたるかを人手でアノテーションしているため、LLM-as-a-Judge、テキスト類似度、人手評価など様々な評価手法によりモデルの法律推論能力を評価することが可能である。そこで、LLM-as-a-Judge 評価、法源の参照能力の評価、人手評価の 3 つの評価を行う。

LLM-as-a-Judge IRAC framework に基づき、GPT-5 モデルを使用した LLM-as-a-Judge による評価を行う。0 または 1 のバイナリ評価により、I, R, A, C の各要素が正解データの内容と同じかどうかを評価するように指示した。

法源の参照 本データセットには、正解回答中で参照している法律の条文、判例、法律の一般原則などの法源のアノテーションが含まれるため、それらの法源がモデル出力で参照できているかどうかを厳密一致に基づく F1 スコアにより評価した。

人手評価 人手評価も IRAC の 4 つの観点に基づいて 0 または 1 のバイナリ評価により行った。人手評価は、各モデル出力毎に評価セットの 500 件中 25 件を抽出して行った。問題毎に解答すべき言語が決まっているため、ある言語で解答すべきなのに他の言語で解答した場合は、0 点として評価した。また、生成文の中に一部でも正解解答が含まれている場合は、正解として評価した。

表 4: 学習済みモデルのプロンプトによる精度比較.

モデル	LLM-as-a-Judge				法源
	I	R	A	C	
(1) SFT w/ 法律の論点 & 問題・解答例					
Qwen3-1.7B	89.6	6.6	11.2	26.4	31.6
Qwen3-8B	92.6	15.4	20.4	24.2	33.1
Qwen3-14B	90.0	30.2	37.4	52.8	33.3
gpt-oss-20b	91.4	22.8	25.8	34.2	31.4
(2) SFT w/ 法律の論点					
Qwen3-1.7B	39.8	3.0	7.6	21.2	3.8
Qwen3-8B	92.2	22.2	32.8	40.4	32.6
Qwen3-14B	94.0	25.8	33.6	38.6	32.5
gpt-oss-20b	89.6	24.2	25.0	30.6	32.9
(3) SFT w/ 問題・解答例					
Qwen3-1.7B	34.2	5.6	8.6	19.0	4.0
Qwen3-8B	47.0	12.2	19.6	32.6	7.0
Qwen3-14B	51.8	16.0	24.6	31.4	9.0
gpt-oss-20b	88.8	32.8	35.8	46.6	31.0

4 実験

公開モデルの LLM 及び商用モデルで実験を行った。公開モデルでは、推論能力と日本語及び英語の能力が一般的に高いとされる gpt-oss 及び Qwen 系モデルを使用した。また、商用モデルの中から、高い推論能力と言語能力を持つ GPT-4o 及び GPT-5 を使用した。ゼロショット生成では、法律の問題文のみを入力し、解答を生成するよう指示した。文脈内学習では、法律の問題文に加え、サンプル問題及びサンプル解答の例を入力した。法律の問題文は非常に長いので、LLM のプロンプト長の制約によりサンプルは 1 例のみ加えた。さらに、本データセットを用いた教師ありファインチューニングを行った。学習においては、人手で作成したデータセットに加えて、GPT-5 により類似問題の合成を行い、元のデータセットのおよそ 3 倍である合計約 15,000 件の学習データを作成した。生成においては、ゼロショットと同様のプロンプトを使用した。

5 結果

表 2 は、IRAC の 4 つの観点別の LLM-as-a-Judge 及び人手評価の結果を示す。LLM-as-a-Judge の評価を見ると、本データセットで教師ありファインチューニングを行うことで、特に Qwen3-14B においてモデルの推論能力が向上していることが分かる。Issue の性能は学習済み Qwen3-14B の性能が最も良く、Rule, Application, Conclusion の 3 つの観点においては GPT-5 の性能が最も性能が良い。また、全体的にゼロショット推論の方が文脈内学習より

も若干精度が高い傾向にある。これは、法律分野が 17 個あるのに対し、与えられるサンプルは 1 例のみのため、サンプル例で与えられたプロンプトに影響され、解答すべき分野がモデルにとって理解しづらくなっている可能性が考えられる。一方、人手評価を見ると、学習なしモデルの性能は低く評価され、教師ありファインチューニングにより Issue, Rule, Application, Conclusion のすべての観点で性能が向上している。これは、LLM-as-a-Judge に使用した GPT-5 自身の法律推論能力に限界があるため、各モデルの出力結果を正しく評価できていない可能性を示唆する。さらに、GPT-5 は自分自身の生成結果を良く見積もりすぎており、実際の手評価では重要な論点を説明せず、関係の無い論点を多く生成していることが分かった。LegalNLP においては、LLM-as-a-Judge による評価だけでは十分ではなく、当面は人手評価を併用する必要があるとされる。

表 3 は、法律の分野別に LLM-as-a-Judge スコアを示す。左側の列に行くほどデータセット中に占める問題件数が多い分野となり、右に行くほど件数が少ない分野となっている。全体的な傾向としては、データセット中の問題数が多いほどスコアが高くなっているが、国際法と国際私法は英語のオープンソースデータが入手可能な科目のため、スコアが高いと考えられる。また、民事実務基礎と民法、及び刑事実務基礎と刑法は同じ科目のため、それぞれスコアが高くなっていると考えられる。

表 4 は、学習済みモデルに対して、質問に加えて (1) 検討すべき Issue と問題・解答例を与えた場合、(2) 検討すべき Issue のみ与えた場合、(3) 問題・解答例を与えた場合の結果を示す。人間の場合、法律の問題に加えて検討すべき Issue が指定された場合、問題にヒントが与えられていることになり、後続の推論が簡単になる。LLM-as-a-Judge のスコアが高かった Qwen3-14B モデルは、検討すべき Issue と問題・解答例を与えることで R, A, C の精度が向上し、人間に近い振る舞いをしていると考えられる。

6 おわりに

日本法及び国際法を対象として法律の論理推論を行うタスクである LegalReasoningBench を提案し、法律推論の 4 つの要素を用いた評価手法を提案した。本データセットで学習されたモデルが、gpt-4o のような一般的に推論能力が高いモデルの法律推論能力を上回ることが示された。

謝辞

本研究は、JST ACT-X JPMJAX25C6, 国立研究開発法人情報通信研究機構 (NICT) の委託研究 (22501), HUMAI Program, 産総研及び AIST Solutions が提供する ABCI 3.0 の「ABCI 3.0 開発加速利用」, 東京科学大学のスーパーコンピュータ TSUBAME4.0 の支援を受けたものです。

参考文献

- [1] Mohamed Elaraby and Diane Litman. ArgLegalSumm: Improving abstractive summarization of legal documents with argument mining. In **Proceedings of the 29th International Conference on Computational Linguistics**, 2022.
- [2] Dennis Aumiller et. al. EUR-lex-sum: A multi- and cross-lingual dataset for long-form summarization in the legal domain. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, 2022.
- [3] Iosif Angelidis, Ilias Chalkidis, and Manolis Koubarakis. Named entity recognition, linking and generation for greek legislation. In **International Conference on Legal Knowledge and Information Systems**, 2018.
- [4] Pedro Henrique Luz de Araujo et. al. Lener-br: A dataset for named entity recognition in brazilian legal text. In **Computational Processing of the Portuguese Language**, 2018.
- [5] Vasile Pais et. al. Named entity recognition in the Romanian legal domain. In **Proceedings of the Natural Legal Language Processing Workshop 2021**, 2021.
- [6] Bonet de Gibert et. al. Spanish datasets for sensitive entity detection in the legal domain. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, 2022.
- [7] Ilias Chalkidis et. al. Extreme multi-label legal text classification: A case study in EU legislation. In **Proceedings of the Natural Legal Language Processing Workshop 2019**, 2019.
- [8] Sina Ahmadi et. al. ConLoan: A contrastive multilingual dataset for evaluating loanwords. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics**, 2025.
- [9] Joel Niklaus et. al. SwiLTra-bench: The Swiss legal translation benchmark. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics**, 2025.
- [10] Ilias Chalkidis et. al. LexGLUE: A benchmark dataset for legal language understanding in English. 2022.
- [11] Kasper Drawzeski et. al. A corpus for multilingual analysis of online terms of service. In **Proceedings of the Natural Legal Language Processing Workshop 2021**, 2021.
- [12] Joel Niklaus et. al. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In **Proceedings of the Natural Legal Language Processing Workshop 2021**, 2021.
- [13] Hai Ye et. al. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, June 2018.
- [14] Huajie Chen et. al. Charge-based prison term prediction with deep gating network. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing**, November 2019.
- [15] Chaojun Xiao et. al. Cail2018: A large-scale legal dataset for judgment prediction. **ArXiv**, Vol. abs/1807.02478, , 2018.
- [16] Neel Guha et. al. Legalbench: a collaboratively built benchmark for measuring legal reasoning in large language models. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, 2023.
- [17] Xiaoxi Kang et. al. Automating irac analysis in malaysian contract law using a semi-structured knowledge base. 2025.
- [18] Tom et. al Brown. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.
- [19] Rafael Rafailov et. al. Direct preference optimization: Your language model is secretly a reward model. In **Thirty-seventh Conference on Neural Information Processing Systems**, 2023.
- [20] Zhihong Shao et. al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.
- [21] Hunter Lightman et. al. Let’s verify step by step. In **The Twelfth International Conference on Learning Representations**, 2024.
- [22] Antoine Louis et. al. Interpretable long-form legal question answering with retrieval-augmented large language models. In **Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence**, 2024.
- [23] Joel Niklaus et. al. LawInstruct: A resource for studying language model adaptation to the legal domain. In **Findings of the Association for Computational Linguistics: NAACL 2025**, 2025.
- [24] Zhiwei Fei et. al. LawBench: Benchmarking legal knowledge of large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, November 2024.

表 5: 法律の分野別データ件数, 言語, 法源データ数を示す.

分野	言語	# Train	# Validation	# Test	# 合計	# 法源
憲法	JP	462	41	37	540	590
民法	JP	922	84	65	1,071	1,610
刑法	JP	1,510	127	144	1,781	2,087
行政法	JP	370	25	35	430	687
商法	JP	628	55	51	734	1,384
民事訴訟法	JP	490	49	49	588	753
刑事訴訟法	JP	483	48	38	569	733
知的財産法	JP	64	3	11	78	84
労働法	JP	263	22	28	313	354
倒産法	JP	39	4	1	44	86
経済法	JP	26	3	1	30	41
環境法	JP	23	0	1	24	60
国際私法	JP	58	4	2	64	67
国際公法 - 一般	JP	82	7	12	101	100
国際公法 - 投資仲裁	EN	178	16	14	208	189
民事実務基礎	JP	66	8	7	81	91
刑事実務基礎	JP	30	4	4	38	43

図 2: Qwen3-14B モデルによる文脈内学習とファインチューニング済みモデルの結果比較を示す.

文脈内学習によるモデル出力 (Qwen3-14B)	ファインチューニング済みモデルによる出力 (Qwen3-14B)
<p>甲は、食料品店主Aに対し、「指定した口座に400万円振り込まなければ、商品に毒を入れるぞ。」と電話で脅し、現金の振込先としてB銀行C支店の自己名義の普通預金口座を指定した。やむなくAが2回に分けて振込手続を行ったところ、Aの手続ミスにより、同支店に開設され、預金残高が37万円であった乙の普通預金口座に振り込まれてしまった。その後、乙が、30万円を通帳を使って窓口で引き出したところ、なお残高が207万円となっていたので、誤振込みがあったことを知り、更に窓口で100万円を引き出した。</p> <p>乙が誤振込み金100万円を窓口で引き出した行為につき罪責を論ずるに際し、まず刑法第235条（窃盗）の構成要件に照らして検討する。</p> <p style="text-align: right;">False</p>	<p>乙が誤振込み金100万円を窓口で引き出した行為について、詐欺罪（刑法246条1項）が成立するか否かを論じるにあたり、以下の点が検討される。</p> <p style="text-align: right;">True</p>
<p>1. 誤振込み金の性質 甲がAから誤って乙の口座に振り込まれた100万円は、本来Aの所有する物であり、乙には法定または約定上の所有権は存在しない。乙は、この金銭を「自分のもの」と認識したに過ぎず、実際には他人の財産であることを認識していた可能性が高い。2. 乙の認識と故意 乙は、通帳の残高が37万円であったにもかかわらず、200万円が振り込まれたことから、誤振込みが発覚した。この時点で、乙は「他者が誤って自分の口座に金銭を振り込んだ」という事実を認識し、その金銭が自らの所有物ではなく、Aの所有物であることを推測すべきである。これに加え、乙は100万円を引き出す際、自身の口座に不正に流入した金銭を「不当に占有し、これを自らの所有物として処分する」という意思を有していた。このような行為は、窃盗の「不正な所有の意図」（刑法第235条1項）に該当する。3. 誤認の抗弁の不成立 乙は、誤振込み金を「自分のもの」と誤認したが、その誤認が合理的な根拠を併せているとは言えない。例えば、通帳の残高が37万円であったにもかかわらず、200万円が振り込まれたことから、乙は自らの口座に不正な金銭が流入していることを確認できた。このような状況では、乙が「自己の所有物」と誤認する根拠が存在せず、誤認の抗弁（刑法第235条の誤認による罪責否定）は成立しない。4. 他の刑法条文との関係 乙の行為は、刑法第246条（背任）や第251条（詐欺）に該当する可能性もあるが、本件においては、乙がAの金銭を故意に不正に占有することを目的とした行為であり、窃盗罪の構成要件を満たす。</p> <p style="text-align: right;">False</p>	<p>1. 誤振込み金の占有主体 誤振込みされた金銭は、銀行が法的に「占有」していると解釈される。これは、銀行は口座への入金が誤りであることを認識していない限り、当該金銭を口座名義人に返還する義務があるため、法的な占有を銀行に帰属するとされる（最高裁判例平成21年12月17日）。したがって、乙が口座から金銭を引き出した際、銀行が占有していた金銭を不正に取得したと見なされる。2. 詐欺罪の構成要件 詐欺罪の成立には、「他人の財産を不正に取得する意思」と「虚偽の事実の陳述または隠蔽による欺罔」の二つの要素が必要である（刑法246条1項）。</p> <p style="text-align: right;">True</p>
<p>乙は、誤振込み金を自らの所有物と誤認したにせよ、その金銭が他人の所有物であることを認識しながら、それを不正に占有しようとした故意を有しており、刑法第235条の窃盗罪に該当する。したがって、乙の行為は罪責を有する。</p> <p style="text-align: right;">False</p>	<p>本件において、乙は銀行に対して虚偽の申告や隠蔽行為をせず、単に自己名義の口座から金銭を引き出したに過ぎない。これ自体は、銀行に対する欺罔が明確に認められない。したがって、詐欺罪は成立しない。</p> <p style="text-align: right;">False</p>
<p>乙は、誤振込み金を自らの所有物と誤認したにせよ、その金銭が他人の所有物であることを認識しながら、それを不正に占有しようとした故意を有しており、刑法第235条の窃盗罪に該当する。したがって、乙の行為は罪責を有する。</p> <p style="text-align: right;">False</p>	<p>4. 結論 乙の行為は、銀行が占有する金銭を不正に取得した点から窃盗罪の対象となり得るが、詐欺罪の構成要件（欺罔の存在）は満たされていない。したがって、乙の行為につき詐欺罪は成立しない。</p> <p style="text-align: right;">False</p>

法的三段論法の4つの要素が出力に含まれる場合、右のようにハイライトする。

論点	規範	当てはめ	結論
----	----	------	----

A データセット統計

表 5 は、法律の科目別のデータセット件数及び法源データ数を示す。

B 質的分析

図 2 は、Qwen3-14B における文脈内学習と教師ありファインチューニングの結果のうち、重要な部分を抽出したものである。文脈内学習では、検討するトピックが詐欺罪ではなく窃盗罪となっており、「3. 誤認の抗弁の不成立」の箇所など誤った内容を生成している。また、「4. 他の刑法条文との関係」というセクションでも、詐欺罪の条文番号が正しくない点や背任罪という今回の事例では成立が考えにくい罪の検討をしている点が見られる。他方で、教師ありファインチューニング済みモデルでは、誤振込金の引き出し行為について詐欺罪を検討できており、検討対象が正しい。また、誤振込金の占有についてと、詐欺罪の構成要件について検討できており、完璧ではないが正しい答えに近い内容を生成できている。しかし、当てはめの段階において欺罔行為が認められないという誤った判断を行い、誤った結論を導いている。

C 学習パラメータ

実験には、NVIDIA H100 GPU 4 基もしくは NVIDIA H200 GPU 8 基を使用し、DeepSpeed ZeRO によるメモリ最適化を適用した。学習率は Qwen 系モデルで $5e-5$ 、gpt-oss-20b で $1e-5$ を使用した。また、学習率スケジューラとして、コサイン波形による減衰を用いた。