LegalRikai: Open Benchmark - 法務ドメインの日本語ベンチマーク

藤田 正悟 [1]，楢木 悠士 [1]，Yiqing Zhu[1]，森 信介 [2]
[1] 株式会社 LegalOn Technologies, [2] 京都大学
shogo.fujita@legalontech.jp

概要

本論文では，法務ドメインの日本語ベンチマーク LegalRikai: Open Benchmark を提案する．本データセットは弁護士の監修の下，実際の法務業務を模した 4 つのタスクから構成されている．各タスクは長文かつ構造化された出力を要する 25 サンプルで設計されており，出力は複数の実務的観点に基づいて評価される．主要な LLM である GPT-5，Gemini 2.5 Pro，Claude Opus 4.1 を対象に人手評価および自動評価を実施した．人手評価の結果，抽象的な指示を要するタスクでは不要な出力が増加しやすいことが明らかとなり，既存のタスクでは捉えにくかった文書レベルの編集における LLM の弱点が示された．

さらに自動評価について分析した結果，文書構造の一貫性や専門用語の正確性に関する評価は困難である一方で，根拠が明確な評価観点においては人手評価と高い整合性を示した．これらの結果から，専門家の確保が難しい状況において，自動評価はスクリーニングや補助的評価手法として有用であることが示唆される．

1 はじめに

LLM の発展に伴い，法務ドメインにおいても，文書の生成・要約・校正といった業務の効率化に期待が寄せられている．実際の法務業務におけるワークフローは，単一かつ単純なタスクではなく，複数の工程が統合された複雑なタスクから構成されている．このような複雑なタスクに対する LLM の対応能力は，これまで限られた範囲でしか評価されてこなかった．

本研究が対象とする実務的な法務業務には，法令改正や関係者からの要請および将来的なリスク低減を背景として，契約書や関連文書を編集するワークフローが含まれる．例えば法令が改正された場合，

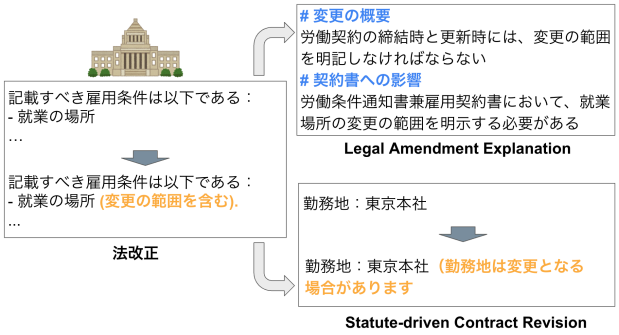

図 1 法務部における法改正に関する業務の概要

法務部門は図 1 に示すように，改正前後の法令内容の理解と差分把握，改正内容の要約，契約書への影響整理に加え，既存契約書の構造理解，修正対象条項の特定，および形式や整合性を維持した編集作業を段階的に行う．

既存の法務ドメインのベンチマークの多くは，法的知識を測定するための短文回答型 QA や分類タスクに焦点を当てている．そのため，契約書全体の整合性を保ちながら条項を修正するといった，実務に不可欠な長文編集能力を十分に評価できていない．この能力の欠如は重大であり，LLM が生成した契約書案は一見正しく見えるものの，条項番号の誤り，定義の不整合，文脈にそぐわない表現といった欠陥を含む可能性がある．そこで本研究では，実際の法務業務のワークフローをモデル化した 4 つの複雑なタスクから成るベンチマークを提案する．[1)] さらに主要な LLM でベンチマークを推論し，その出力に対して人手評価と LLM による自動評価を行なった．これにより，各評価観点においてどの能力が実用レベルに達しているのか，どこに顕著な弱点が存在するのかを明らかにする．

本研究の貢献は以下のとおりである．

1） https://huggingface.co/datasets/legalontech/Legal-Rikai-Open-Benchmark にてデータセットを公開している．実験で使用したデータは公開しているデータと一部異なる点がある．

- 法務業務のワークフローに対応した，4 つのタスクから構成される新規データセットの提案
- 本データセットに対する評価尺度の定義，および LLM を用いた自動評価の枠組みの構築
- 主要な LLM を用いた本データセットに対する推論と評価の実施および各モデルの性能と挙動の分析

2 データセット

提案するデータセットは日本の法務業務のワークフローを反映した 4 つのタスクによって構成される．評価尺度の詳細は表 5 に記す．

Legal Amendment Explanation (`AmendExp`) は法令の改正内容を要約し，契約書への影響を説明するタスクである．評価尺度は Coverage of Amendments (CA)，Accuracy of Amendments (AA)，Relevance of Amendments (RA)，Coverage of Impacts (CI)，Accuracy of Impacts (AI)，Relevance of Impacts (RI) を用いる．**Statute-Driven Contract Revision (`StatRev`)** は改正後の法令に適合するよう契約書を編集するタスクである．評価は，Instruction Following (IF)，Structural Consistency (SC)，Change Precision (CP)，Terminology Accuracy (TA)，Wording Appropriateness (WA) に基づいて行う．**Requirement-Driven Contract Revision (`ReqRev`)** は明示的に与えられた指示に従って契約書を編集するタスクである．評価は `StatRev` と同一の尺度で行う．**Risk-Driven Contract Reivison (`RiskRev`)** は法的リスクを軽減することを目的とした抽象的な指示に従って契約書を編集するタスクである．評価は `StatRev` と同一の尺度で行う．

弁護士の監修の下，各タスクにつき 25 件のデータを作成した．`RiskRev`，`ReqRev`，および `StatRev` のデータは実務で用いられる契約書およびテンプレートに基づいて構築した．また，`StatRev` および `AmendExp` で用いる法令は，実在する日本の法令に基づいている．各タスクの具体例をそれぞれ，`AmendExp` は表 6，`StatRev` は表 7，`ReqRev` および `RiskRev` は表 8 に示す．また，データセットの統計情報は表 9 にまとめる．

3 実験

4 つの各タスクについて，LLM で推論を行い，その結果について人手評価および自動評価を実施した．人手評価では，GPT-5[2] [24]，Gemini 2.5 Pro[3] [8]，Claude Opus 4.1[4] [1] の 3 モデルで推論を行い，弁護士の監督の下で出力を評価した．自動評価では，同様に GPT-5，Gemini，Claude の 3 モデルを評価モデルとして用いた．各タスクについて，出力形式および制約を指定したタスク固有テンプレートを用いた．契約書は Markdown 形式で，法令はデジタル庁が公開している法令 API[3] から取得した情報をテキスト形式に変換してモデルに入力した．[5] `AmendExp` では，モデルは改正要約と契約書への影響を Markdown で出力した．`StatRev` では，コンテキスト長の制約により，変更が必要な条項のみを Markdown で出力した．`ReqRev` および `RiskRev` では，編集後の契約書全体を Markdown で出力した．各モデルの temperature は 0.0 に設定した．GPT-5 は公式 API を通じて直接利用し，Gemini と Claude は Vertex AI Platform 経由で利用した．

3.1 人手評価

各タスクについて GPT-5，Gemini，Claude を用いて推論を行い，法務専門家 2 名により独立に評価を行なった．評価尺度は 2 章で定義した指標（`AmendExp`:CA/AA/RA, CI/AI/RI；他 3 タスク:IF/SC/CP/TA/WA）を用いた．評価スコアは区間尺度に変換し，$[0, 1]$ に正規化した（表 2）．評価者間でスコアが異なる場合には，その算術平均を用いた．評価者間一致度については，Cohen の κ スコアを用いて算出し，表 1 に示す．なお，すべてのサンプルで同一のラベルが付与された評価尺度については，κ スコアは `nan` となる．全体として高い一致度が確認された一方で，`ReqRev` および `RiskRev` における WA の κ スコアは低く，契約書的な言い回しの適切性に関する判断が評価者間で大きく異なることを示している．

表 2 に示すように，`AmendExp` では Gemini が全体的に最も高い性能を示した．一方，GPT-5 は網羅性を示す CA，CI は Gemini と同程度だが他の部分で劣っており，Claude は顕著な優位性を示さなかった．`StatRev` では，Gemini が IF と CP で高いスコアを示し，Claude は SC と WA に強みを示し，GPT-5 では CP が低く不要な修正が多く見られた．すべて

2） モデルバージョンは gpt-5-2025-08-07
3） モデルバージョンは gemini-2.5-pro
4） モデルバージョンは claude-opus-4-1-20250805
5） 入力長が LLM の最大コンテキスト長を超える条文については，改正前後で差分が存在しない末尾部分を切り詰めた．

CA	AA	RA	CI	AI	RI
1.00	1.00	1.00	0.98	0.98	0.91
Task	IF	SC	CP	TA	WA
ReqRev	0.78	1.00	0.74	1.00	0.49
RiskRev	0.78	1.00	0.69	0.53	0.60
StatRev	1.00	0.90	1.00	nan	1.00

表 1 各タスクにおける 2 名の評価者による評価結果に対する Cohen の κ スコア．上段:AmendExp，下段:ReqRev，RiskRev，StatRev．なお，StatRev における TA については，すべての評価結果が一致していたため nan として示している．

Model	CA	AA	RA	CI	AI	RI
GPT-5	0.66	0.74	0.76	0.52	0.44	0.22
Gemini	0.64	0.86	0.80	0.52	0.54	0.44
Claude	0.64	0.68	0.70	0.49	0.49	0.32
Task	Model	IF	SC	CP	TA	WA
StatRev	GPT-5	0.69	0.36	0.32	1.00	0.94
	Gemini	0.73	0.36	0.44	1.00	0.96
	Claude	0.57	0.40	0.20	1.00	1.00
ReqRev	GPT-5	0.85	0.92	0.58	0.94	0.79
	Gemini	0.85	0.80	1.00	1.00	0.90
	Claude	0.85	0.92	0.96	1.00	0.87
RiskRev	GPT-5	0.69	0.88	0.22	0.79	0.59
	Gemini	0.61	0.96	0.40	0.95	0.78
	Claude	0.65	1.00	0.56	0.96	0.75

表 2 LLM の出力に対する人手評価結果．上段:AmendExp，下段:StatRev，ReqRev，RiskRev。すべての評価値は区間 $[0, 1]$ に正規化している．

のモデルで TA は最大値であり，専門用語の扱いに問題は見られなかった．ReqRev では，明示的な指示が与えられるため IF は全モデルで同程度に高く，GPT-5 は CP が低く不要な修正が多く見られた．RiskRev は ReqRev と比較してモデル間の傾向は同じだが，全体的にスコアが低下した．

全体として，指示の具体性がモデルの挙動に大きく影響することが確認された．具体的な指示が与えられる場合は高い性能を示す一方，指示が抽象的で解釈を要するタスクでは不要な変更が増加した．また，Gemini は正確性，Claude は契約書書式の維持に強みを示し，GPT-5 は中間的な性能を示した．これらの結果は，法務業務における LLM の選択には，タスクの性質とモデル特性を考慮する必要があることを示している．

3.2 自動評価

本節では，LLM による自動評価が人間のアノテーションを代替または補助可能かを検証した．2 章で定義した評価尺度について，人手評価との間で Spearman の順位相関係数（表 3）および平均絶対誤差（表 4）を算出した．

評価モデルには GPT-5，Gemini，Claude を用い，temperature は 0.0 に設定した．評価に用いる LLM に，タスク定義，入力，モデル出力，評価基準を与え，尺度名とその値及び根拠から成る構造化出力を要求した．加えて，3 モデルの平均である Avg スコアも算出した．

表 3 に示すように，AmendExp では全体として有意な正の相関が確認された．これは，改正差分の特定や契約書への影響の整理といった，根拠が明確なタスクが自動評価に適していることを示す．

StatRev では IF， CP に中程度の相関が見られ，SC， WA は弱い相関が見られた．TA は人手評価が一様で相関を算出できなかったが，平均絶対誤差は IF，CP と同程度であった．ReqRev では IF， WA に中程度の相関，SC， CP，TA に弱い相関が見られた．RiskRev では CP，WA に中程度，IF，SC，TA に弱い相関が確認された．

総じて，AmendExp のようにテキスト内に明確な証拠があるタスクでは，自動評価と人手評価の整合性が高い．一方，契約書編集タスクでは，特に SC や TA で相関が弱く，LLM 評価の限界が示された．SC は文書全体構造の把握が難しい点，TA は法的文脈に依存する用語の判断が難しい点に起因する．以上より，自動評価は網羅性や指示遵守といった内容面のスクリーニングに有効であるが，契約書の書式や専門用語の評価には人手の関与が不可欠である．また，Avg スコアはモデル間のバイアスを相殺し，相関を改善する傾向が見られた．

4 関連研究

LLM の発展に伴い，法務領域における自然言語処理タスクに関する研究が活発化している．特に，LLM の法的知識や推論能力を評価するため，国や言語ごとの法制度および文書形式に対応した多様なベンチマークデータセットが提案されてきた．

法的知識や推論能力を評価する代表的なベンチマークとして，MMLU [10], LawBench [6], LegalBench [9], LegalAgentBench [17], LexGLUE [2] などがある．これらは，事実知識の想起，多段推論，判例分析など，法的理解の多様な側面を対象としている．LEXam [5] は 116 科目，340 件の法学試験問題から構成され，複数分野および学位レベルを網羅するベンチマークである．BriefMe [30] は，主張の要約，補完，判例検索といった訴訟準備タスクに焦点を当

Model	CA	AA	RA	CI	AI	RI
GPT-5	0.72†	0.32†	0.41†	0.50†	0.44†	0.23†
Gemini	0.58†	0.36†	0.47†	0.36†	0.31†	0.13
Claude	0.53†	0.37†	0.46†	0.53†	0.54†	0.41†
Avg	0.68†	0.38†	0.51†	0.55†	0.47†	0.31†

Task	Model	IF	SC	CP	TA	WA
StatRev	GPT-5	0.34†	0.07	0.47†	nan	-0.04
	Gemini	0.43†	0.11	0.46†	nan	0.08
	Claude	0.34†	0.08	0.58†	nan	0.19
	Avg	0.49†	0.12	0.65†	nan	0.14
ReqRev	GPT-5	0.17	0.09	0.15	-0.05	0.00
	Gemini	0.10	0.08	0.17	-0.06	0.27†
	Claude	0.26†	0.16	nan	-0.12	0.07
	Avg	0.23†	0.12	0.18	-0.15	0.23†
RiskRev	GPT-5	0.16	-0.08	0.36†	0.14	0.22
	Gemini	-0.13	0.01	0.26†	-0.02	0.20
	Claude	0.20	0.07	0.35	0.11	0.26†
	Avg	0.11	-0.07	0.41†	0.19	0.34†

表 3 人手評価と自動評価の間の Spearman の順位相関係数．上段:AmendExp，下段:StatRev，ReqRev，RiskRev。† は，人手評価との差が統計的に有意であること（$p < 0.05$）を示す．

Model	CA	AA	RA	CI	AI	RI
GPT-5	0.31	0.47	0.71	0.66	0.66	0.64
Gemini	0.39	0.44	0.47	0.69	0.75	1.01
Claude	0.44	0.53	0.56	0.58	0.55	0.63
Avg	0.37	0.47	0.58	0.58	0.60	0.71

Task	Model	IF	SC	CP	TA	WA
StatRev	GPT-5	0.54	0.60	0.24	0.52	0.31
	Gemini	0.58	0.57	0.25	0.28	0.36
	Claude	0.75	0.57	0.23	0.79	0.59
	Avg	0.56	0.58	0.24	0.53	0.41
ReqRev	GPT-5	0.33	0.15	0.21	0.09	0.31
	Gemini	0.35	0.21	0.19	0.13	0.28
	Claude	0.42	0.27	0.17	0.17	0.28
	Avg	0.34	0.21	0.17	0.17	0.28
RiskRev	GPT-5	0.43	0.15	0.33	0.25	0.52
	Gemini	0.62	0.27	0.42	0.24	0.52
	Claude	0.46	0.43	0.38	0.24	0.48
	Avg	0.49	0.28	0.36	0.23	0.49

表 4 人手評価と自動評価の間の平均絶対誤差．上段:AmendExp，下段:StatRev，ReqRev，RiskRev。

てている．また，韓国語および中国語など，非英語圏の法制度に基づくマルチタスク型ベンチマークも提案されている [12, 18, 20]．

一方，日本語 LLM の評価に関しては，LLM-jp[19] をはじめ，Vicuna QA 日本語版 [28] や Judge-free Benchmark[13] が提案されている．さらに，金融分野の FinBench [11]，医療分野の JMedBench [14]，形式論理推論を対象とする JFLD [22]，制御生成を評価する LCTG Bench [16] など，分野特化型のベンチマークも整備されつつある．日本の法令に関する多肢選択式 QA データセット [4] は，日本の法令に対する理解度を評価するデータセットである．これらのベンチマークはいずれも，日本の法務業務における具体的な文脈を十分に捉えていない．

近年の LLM に関するベンチマークでは，単純な知識想起や分類タスクではなく，複数の推論過程や工程を要する複雑なタスクが注目されている．GAIA [21] や Humanity's Last Exam [26] は一般的および専門的な推論能力を評価し，SWE-bench [15] や PaperBench [27] は，現実的設定における多段階問題解決能力を対象とする．ProcBench，ToolComp，Multi-LogiEval [7, 23, 25] などは，段階的推論やツール利用の重要性を強調している．これらの研究は複雑なタスク評価の有効性を示しているが，法務業務に特有の長文編集や契約書構造の一貫性を直接評価対象としていない．

以上の先行研究に対し，本研究は，日本の法務ドメインにおける実世界のワークフローを反映し，長文かつ構造化された契約書編集を伴う複雑なタスクを体系的に評価する点に特徴がある．

5 おわりに

本研究では，日本の法務業務に即した 4 つのタスク（AmendExp，StatRev，ReqRev，RiskRev）から構成される新しいデータセットを提案した．これらのタスクは，実務における長文編集や複雑な判断を含むワークフローを反映しており，従来の短文中心の法務ベンチマークでは十分に評価されてこなかった能力を対象としている．また，人手評価および自動評価の双方を用いて主要な LLM の性能を分析し，タスクの性質や指示の具体性によってモデルの挙動や弱点が大きく異なることを示した．

分析を通じて，法務業務における LLM の適用可能性と限界が明らかになった．本ベンチマークは，データ設計，評価設計，および実験設定を公開することで，日本の法務タスクにおける LLM 性能評価のための共通基盤を提供し，今後のモデル開発や評価手法の検討を促進することを目的とする．

6 謝辞

本研究におけるデータセットの作成にご尽力いただいた法務開発チームの小林司氏，今野悠樹氏，西海枝翔氏，軸丸厳氏，高澤和也氏，有沢慎之介氏，谷口香織氏に深く感謝申し上げます．

参考文献

[1]Anthropic. Claude opus 4.1. https://www.anthropic.com/news/claude-opus-4-1, 2025. Accessed on October 17, 2025.

[2]I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. Katz, and N. Aletras. LexGLUE: A benchmark dataset for legal language understanding in English. In S. Muresan, P. Nakov, and A. Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pages 4310–4330. Association for Computational Linguistics, May 2022.

[3]G. o. J. Digital Agency. e-gov laws and regulations search api swagger ui. https://laws.e-gov.go.jp/api/2/swagger-ui, 2025. Accessed on October 17, 2025.

[4]Digital Agency, Government of Japan. Legal System Related Dataset (LawQA_JP). GitHub Repository, 2025. Accessed on October 17, 2025.

[5]Y. Fan, J. Ni, J. Merane, Y. Tian, Y. Hermstrüwer, Y. Huang, M. Akhtar, E. Salimbeni, F. Geering, O. Dreyer, D. Brunner, M. Leippold, M. Sachan, A. Stremitzer, C. Engel, E. Ash, and J. Niklaus. Lexam: Benchmarking legal reasoning on 340 law exams. 2025.

[6]Z. Fei et al. LawBench: Benchmarking legal knowledge of large language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pages 7933–7962. Association for Computational Linguistics, Nov. 2024.

[7]I. Fujisawa, S. Nobe, H. Seto, R. Onda, Y. Uchida, H. Ikoma, P.-C. Chien, and R. Kanai. Procbench: Benchmark for multi-step reasoning and following procedure. 2024.

[8]Google. Gemini 2.5. https://developers.googleblog.com/en/gemini-2-5-thinking-model-updates/, 2025. Accessed on October 17, 2025.

[9]N. Guha et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **Advances in Neural Information Processing Systems**, volume 36, pages 44123–44279. Curran Associates, Inc., 2023.

[10]D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In **ICLR**. OpenReview.net, 2021.

[11]M. Hirano. Construction of a Japanese Financial Benchmark for Large Language Models. pages 1–9, 2024.

[12]W. Hwang, D. Lee, K. Cho, H. Lee, and M. Seo. A multi-task benchmark for korean legal language understanding and judgement prediction. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, **Advances in Neural Information Processing Systems**, volume 35, pages 32537–32551. Curran Associates, Inc., 2022.

[13]K. Imajo, M. Hirano, S. Suzuki, and H. Mikami. A judge-free llm open-ended generation benchmark based on the distributional hypothesis. 2025.

[14]J. Jiang, J. Huang, and A. Aizawa. JMedBench: A benchmark for evaluating Japanese biomedical large language models. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, **Proceedings of the 31st International Conference on Computational Linguistics**, pages 5918–5935. Association for Computational Linguistics, Jan. 2025.

[15]C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. R. Narasimhan. SWE-bench: Can language models resolve real-world github issues? In **The Twelfth International Conference on Learning Representations**, 2024.

[16]K. Kurihara, M. Mita, P. Zhang, S. Sasaki, R. Ishigami, and N. Okazaki. Lctg bench: Llm controlled text generation benchmark. 2025.

[17]H. Li et al. LegalAgentBench: Evaluating LLM agents in legal domain. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pages 2322–2344. Association for Computational Linguistics, July 2025.

[18]H. Li, Y. Shao, Y. Wu, Q. Ai, Y. Ma, and Y. Liu. Lecardv2: A large-scale chinese legal case retrieval dataset. In **Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '24, page 2251–2260. Association for Computing Machinery, 2024.

[19]LLM-jp et al. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. 2024.

[20]Y. Ma, Y. Shao, Y. Wu, Y. Liu, R. Zhang, M. Zhang, and S. Ma. Lecard: A legal case retrieval dataset for chinese law system. In **Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '21, page 2342–2348. Association for Computing Machinery, 2021.

[21]G. Mialon et al. Gaia: A benchmark for general ai assistants. In **Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)**, 2024.

[22]T. Morishita, A. Yamaguchi, G. Morio, H. Tomonari, O. Imaichi, and Y. Sogawa. JFLD: A Japanese benchmark for deductive reasoning based on formal logic. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pages 9526–9535. ELRA and ICCL, May 2024.

[23]V. Nath, P. Raja, C. Yoon, and S. Hendryx. Toolcomp: A multi-tool reasoning & process supervision benchmark. 2025.

[24]OpenAI. Introducing gpt-5 for developers. https://openai.com/ja-JP/index/introducing-gpt-5-for-developers/, 2025. Accessed on October 17, 2025.

[25]N. Patel et al. Multi-LogiEval: Towards evaluating multi-step logical reasoning ability of large language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pages 20856–20879. Association for Computational Linguistics, Nov. 2024.

[26]L. Phan et al. Humanity's last exam. 2025.

[27]G. Starace, O. Jaffe, D. Sherburn, J. Aung, J. S. Chan, L. Maksin, R. Dias, E. Mays, B. Kinsella, W. Thompson, J. Heidecke, A. Glaese, and T. Patwardhan. Paperbench: Evaluating ai's ability to replicate ai research. 2025.

[28]Y. Sun, Z. Wan, N. Ueda, S. Yahata, F. Cheng, C. Chu, and S. Kurohashi. Rapidly developing high-quality instruction data and evaluation benchmark for large language models with minimal human effort: A case study on Japanese. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pages 13537–13547. ELRA and ICCL, May 2024.

[29]H. Touvron et al. Llama 2: Open foundation and fine-tuned chat models, 2023.

[30]J. Woo, F. Hashemi Chaleshtori, A. Marasovic, and K. Marino. BriefMe: A legal NLP benchmark for assisting with legal briefs. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, **Findings of the Association for Computational Linguistics: ACL 2025**, pages 13139–13190. Association for Computational Linguistics, July 2025.

[31]H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun. How does NLP benefit legal system: A summary of legal artificial intelligence. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pages 5218–5230. Association for Computational Linguistics, July 2020.

Appendix

タスク	評価尺度名	略称	説明
AmendExp	Coverage of Amendments	CA	改正内容の要約が必要な情報を網羅しているかを評価する.
	Accuracy of Amendments	AA	改正内容の要約が正確であるかを評価する.
	Relevance of Amendments	RA	改正内容の要約に不要な内容が含まれていないかを評価する.
	Coverage of Impacts	CI	契約書への影響について，必要な観点が網羅されているかを評価する.
	Accuracy of Impacts	AI	契約書への影響の説明が正確であるかを評価する.
	Relevance of Impacts	RI	契約書への影響の説明に不要な内容が含まれていないかを評価する.
StatRev, ReqRev, RiskRev	Instruction Following	IF	与えられた指示にどの程度正確に従っているかを評価する.
	Structural Consistency	SC	契約書全体の構造や条文構成が一貫しているかを評価する.
	Change Precision	CP	本来変更すべきでない箇所が変更されていないかを評価する.
	Terminology Accuracy	TA	法律用語や専門用語が正確に使用されているかを評価する.
	Wording Appropriateness	WA	契約書として適切な文体・表現が使用されているかを評価する.

表 5　各タスクにおける評価尺度の概要

改正前の法令	...2 懲役は、刑事施設に拘置して所定の作業を行わせる。...
改正後の法令	...2 拘禁刑は、刑事施設に拘置する。3 拘禁刑に処せられた者には、改善更生を図るため、必要な作業を行わせ、又は必要な指導を行うことができる。 ...
正例 (変更点の概要)	刑事施設における受刑者の処遇及び執行猶予制度等のより一層の充実を図るため、懲役及び禁錮を廃止して拘禁刑を創設。
正例 (契約書への影響)	ストックオプション割当契約等、個人に何らかの権利を付与する契約において、その個人が刑罰に処せられた場合はその権利を喪失する旨が定められている場合がある。その規定において、"懲役""禁錮"という文言が使われていたら、"拘禁刑"に改正する必要がある。

表 6　AmendExp の例.

改正前の法令	... 第五条 使用者が法第十五条第一項前段の規定により労働者に対して明示しなければならない労働条件は、次に掲げるものとする。 (略) 一の三 就業の場所及び従事すべき業務に関する事項...
改正後の法令	... 第五条 使用者が法第十五条第一項前段の規定により労働者に対して明示しなければならない労働条件は、次に掲げるものとする。 (略) 一の三 就業の場所及び従事すべき業務に関する事項 (就業の場所及び従事すべき業務の変更の範囲を含む。) ...
入力契約書	... 従業員は、下記の就業場所において使用者の指示に従い誠実に行う。 就業場所: 雇入れ直後　東京本社 ...
正例 (Revised contract)	... 従業員は、下記の就業場所において使用者の指示に従い誠実に行う。 就業場所: 雇入れ直後　東京本社 (変更の範囲: 会社の別途指定する就業場所)...

表 7　StatRev の例.

ReqRev におけるユーザーからの要望	第 8 条 (再委託) を"委託者の事前の書面による承諾を得た場合に限り、再委託可能とする"旨に修正してください。
RiskRev におけるユーザーからの要望	再委託に関する条項について、契約書内に該当するテーマや項目が含まれている場合は、当該表現を委託側に有利に修正してください。削除することが有利になる場合は、削除しても良いです。
入力契約書	... 第 8 条 (再委託) 1. 乙は、乙の責任において、本委託業務の一部を第三者に再委託することができる。ただし、乙は、甲が要請した場合、再委託先の名称及び住所等を甲に報告しなければならない。...
正例	... 第 8 条 (再委託) 1. 乙は、甲の事前の書面による承諾を得た場合に限り、乙の責任において、本委託業務の一部を第三者に再委託することができる。...

表 8　ReqRev と RiskRev の例.

	ReqRev	RiskRev	StatRev	AmendExp
サンプル数	25	25	25	25
入力契約書の平均文字数	6,530	6,530	9,412	-
入力契約書の平均条文数	16	16	38	-
改正前法令の平均文字数	-	-	27,033	32,149
改正前法令の平均条文数	-	-	292	318
正解データの文字数	253	228	300	397
法令改正差分の文字数	-	-	3,122	4,924

表 9　4 つのタスクにおけるデータ統計量．ReqRev，RiskRev，StatRev における正解データの文字数は，入力契約書と正解契約書との差分の文字数を示す．AmendExp では，正解要約および契約への影響説明の文字数を示している．