LawQA-JP: 日本の法令に関する多肢選択式質問応答データセットの公開と分析

植松 幸生
デジタル庁 / 東京理科大学 創域理工学部/総合研究院
yukuemats@digital.go.jp/yukio@rs.tus.ac.jp

大杉 直也
デジタル庁
naoosugi@digital.go.jp

概要

本稿では，日本の法令を対象とした多肢選択式質問応答データセット LawQA-JP の公開について報告する．本データセットは，デジタル庁が行政および企業における生成 AI 利活用を推進する目的で，公共データ利用規約（第 1.0 版）の下に公開したものである．デジタル庁は AI 利活用におけるデータを「評価用」「コンテキスト用」「パラメトリックな学習用」の 3 種に分類しており，LawQA-JP はこのうち AI の能力を客観的に測定するための「評価用データ」として位置づけられている．本稿では，全 140 問の設問を対象に，設問カテゴリおよび選択肢の言語的特徴を定量的に分析するとともに，大規模言語モデルを用いた参照評価を実施した．その結果，関連法令条文をコンテキストとして付与することで，法的推論における回答精度が有意に向上することを確認した．さらに，現状の LLM にとって特に困難な設問タイプについても分析を行った．

1 序論

近年，大規模言語モデル（LLM）の発展により，法令や契約書といった専門文書を対象とした自然言語処理技術の社会実装が急速に進んでいる．一方で，生成 AI を行政実務や企業活動に導入するにあたっては，その正確性や信頼性を事前に検証するための評価基盤が不可欠である．特に，日本の法令を対象とした質問応答に関しては，体系的に整備された評価用データセットが限られているという課題があった．

日本の法令は，条・項・号から成る階層構造を持ち，多数の参照規定や例外規定を含む．また，「及び」「又は」といった接続表現や，条件節の解釈によって法的効果が大きく変化するなど，高度な構造的理解を要する特徴を有する．これらの特性により，英語圏で整備されてきた法領域ベンチマークを単純に翻訳・適用するだけでは，日本の法令に固有の推論難易度を十分に評価することは困難である．

こうした背景のもと，デジタル庁は，生成 AI の業務利用に関する検討を支援する目的で，日本の法令を対象とした多肢選択式質問応答データセットを公開した．本データセットは，実務で参照される法令条文を前提とした設問から構成されており，AI が法令理解においてどの程度適切な判断を行えるかを確認するためのベンチマークとして位置づけられている．

本稿の目的は，公開された LawQA-JP の設計方針および収録内容を明確に示すとともに，データセット分析を通じて，本データセットが評価対象とする推論能力の特性を明らかにすることである．また，LLM を用いた参照評価を行うことで，LawQA-JP が現状のモデルに対してどのような難易度と評価上の位置づけを持つかを示す．

2 関連研究

本節では，法領域における質問応答およびベンチマーク研究を概観し，日本法を対象とした多肢選択式 QA データセットの必要性を明確にする．

法領域の自然言語処理に関しては，英語圏を中心に多様なベンチマークが提案されてきた．代表的な LexGLUE [1] や LegalBench [3] は，法的推論や文書理解を含む包括的な評価基盤を提供している．また，CUAD [4] や CaseHOLD citepzheng2021casehold など，契約書解析や判例に基づく推論を対象としたデータセットも存在する．しかしながら，いずれも英語圏の法制度や文書構造を前提としており，日本法特有の条文構造や用語体系を考慮した多肢選択形式による実務的な判断を評価することは困難である．

国内においては，COLIEE [2] を中心に，判例検

索や法的含意認識に関する研究進められてきた．COLIEE は司法試験や判例を想定した高度な法的推論能力の評価に有用である一方，企業法務や行政実務において想定される「条文理解に基づく判断」を，多肢選択式で評価する枠組みは十分に整備されていない．このような背景から，日本の法令を対象とし，実務に即した多肢選択式 QA によって LLM の法令理解能力を評価するデータセットが必要であり，本研究ではそのギャップを埋めることを目的として LawQA-JP を構築し，公開した．

3 LawQA-JP データセット

本節では，デジタル庁が公開した日本の法令に関する多肢選択式質問応答データセット LawQA-JP について，公開形態，データ構造，および基本的な統計情報を説明する．本データセットは，生成 AI の業務利用における検討および研究用途での評価を想定して整備されたものである．

3.1 公開形態と利用条件

LawQA-JP は，デジタル庁が日本の法令に関する質問応答データとして作成・公開したものであり，GitHub 上の公開リポジトリ `https://github.com/digital-go-jp/lawqa_jp` を通じて提供されている．本リポジトリでは，多肢選択式質問応答データを CSV および JSON 形式の構造化データとして公開しており，外部の研究者や開発者が容易に取得・利用できるよう設計されている．公開されているデータは，公共データ利用規約（題一版）の下で提供されており，行政機関，企業，および学術研究コミュニティにおいて，研究目的および非研究目的の双方において利用可能である．

図 1 は，LawQA-JP を公開した GitHub リポジトリにおけるスター数の累積推移を示したものである．縦の点線は公開日 (2025 年 10 月 9 日) を示している．スター付与時刻（`starred_at`）は GitHub REST API を用いて取得し，日次で集計した．公開直後に急激な増加が見られ，その後は緩やかな増加に移行していることから，公開初期に高い関心を集めた後，継続的に利用・参照されていることが分かる．2026 年 1 月 7 日現在で 258 のスターと 6 つの fork レポジトリが存在しており，活用されている．

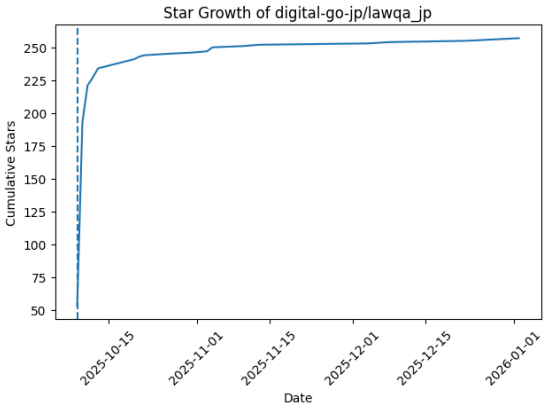

図 1 LawQA-JP GitHub リポジトリにおけるスター数の累積推移 (X 軸：スター数，Y 軸：日付)

表 1 LawQA-JP（selection.csv）の基本統計

項目	値
総問題数 (以下法令内訳)	140
金融商品取引法	80
医薬品医療機器等法	39
借地借家法	21
参照文数	
平均参照条文数	1.36
最大参照条文数	3

3.2 収録法令と問題数

LawQA-JP は，複数の日本法令を対象とした多肢選択式問題が収録されている．本稿で扱う公開版データセットには，作成を依頼した弁護士が実務で利用している金融商品取引法，医薬品医療機器等法 (以下，薬機法)，および借地借家法を対象とした合計 140 問の問題で構成されている．各問題はいずれも 4 択形式で構成されており，選択肢数は全問題で共通である．

表 1 に，LawQA-JP の基本統計情報を示す．本データセットは，特定の一法令に偏らない構成となっており，複数の条文参照を横断的に含む点にも特徴がある．

3.3 データ構造

LawQA-JP は，CSV および JSON 形式で提供されており，各問題は以下の情報を含む．

- 問題文
- 選択肢（4 択）
- 正答インデックス

表 2　問題カテゴリ別分布（selection.csv）

カテゴリ	問題数	割合 (%)
解釈・比較・総合判断	96	68.6
要件・適用条件	26	18.6
否定型設問（誤り選択）	14	10.0
定義・用語理解	4	2.9
合計	140	100.0

- 参照条文情報（法令名・条番号）

参照条文情報は，GitHub 上の公開リポジトリにおいて URL のリストとして明示されており，各設問が依拠する法令条文を第三者が容易に確認できる構成となっている．この点は，行政が公開するデータセットとしての透明性および再現性を担保する上で重要である．

3.4　設計方針と特徴

LawQA-JP は，日本の法令の単純な文言一致ではなく，条文構造や論理関係の理解を評価することを目的として設計されている．LawQA-JP に含まれる設問は，問題文および選択肢の表現に基づき，定義・用語理解，要件・適用条件，否定型設問（誤り選択），解釈・比較・総合判断の 4 つの問題カテゴリから構成されている．上記の分類に基づき，LawQA-JP に含まれる 140 問の設問を分類した結果を表 2 に示す．

表 2 から分かるように，LawQA-JP に含まれる設問の約 7 割は，「解釈・比較・総合判断」カテゴリに分類される．この結果は，本データセットが，単なる法令用語の知識確認ではなく，条文構造や論理関係を踏まえた実務的な法令解釈能力の評価を主目的として設計されていることを示している．一方で，「定義・用語理解」カテゴリの設問は全体の 3% 未満に留まっており，辞書的知識のみで解答可能な問題は重視しなかった．

4　LLM を用いた参照評価

本節では，LawQA-JP の難易度および評価上の位置づけを明らかにする目的で，複数の LLM を用いた参照評価を行う．ここでの評価は，特定の手法やプロンプト設計の有効性を示すことを目的とするものではなく，公開データセットとしての LawQA-JP が，現状の LLM に対してどのような課題を提示するかを定量的に把握することを目的としている．

```
【参考コンテキスト】
金融商品取引法 第 5 条第 6 項 …（略）

【問題】
金融商品取引法第 5 条第 6 項に基づき…

【選択肢】
a. …
b. …
c. …
d. …

次の形式で答えてください：
- answer: a/b/c/d
- 根拠: 簡潔に（日本語）
```

図 2　LawQA-JP において実験で使用した入力プロンプト例（コンテキストあり）

4.1　評価設定

評価対象として，近年広く利用されている表 3 に記載された複数の LLM を選定した．評価では，各設問に対してモデルが出力した選択肢のうち，正答を選択できたかどうかを基準として正答率（Accuracy）を算出した．また，設問に付随する参照条文の有無（コンテキスト有無）が回答結果に与える影響を比較した．なお，本評価は，LawQA-JP の設問が要求する推論特性を把握することを目的としており，各モデルに対する最適化や詳細なプロンプト調整は行っていない．図 2 は，本実験に用いたプロンプト（コンテキストあり）の一例である．このプロンプトを全ての LLM に対して用いた．

4.2　コンテキスト有無による正答率比較

表 3 は，コンテキスト（参照条文情報）の有無による各モデルの正答率を示したものである．すべてのモデルにおいて，コンテキストを付与した条件で正答率が向上する傾向が確認された．特に，比較的小規模なモデルでは，正答率の改善幅が大きい．

この結果は，LawQA-JP に含まれる多くの設問が，問題文単体からの推測ではなく，法令条文の構造や定義関係を踏まえた推論を必要としていることを示唆している．すなわち，関連する文脈情報を明示的に与えない場合，モデルは設問の意図を十分に捉えられず，誤答に至りやすいことが分かる．

また，コンテキスト有無によるモデルの判断変化を詳細に分析するため，同一設問に対する正誤の変化を McNemar 検定により評価した結果を*で示す．gpt-5-nano 以外は，コンテキスト付与による正答への変化が統計的に有意であることが確認された．

表 3 コンテキスト有無によるモデル別正答率（%）
コンテキストの有無で精度が有意に違う場合は*で表記 (McNemar 検定, $N = 140$)

モデル	コンテキスト なし	あり
gpt-4.1	60.7	96.4*
gpt-5-mini	53.6	89.3*
gpt-5-nano	46.4	55.7
gemini-2.0-flash	65.7	95.0*

4.3 参照評価からの示唆

以上の参照評価から，LawQA-JP は，最新の LLM に対しても単純な知識想起や表層的な文理解では安定して解答できない設問を含むことが明らかとなった．特に，条文構造や参照関係を明示的に考慮しない場合，誤答が多く見られた (詳細は付録参照).

これらの結果は，LawQA-JP が，日本の法令に基づく構造的推論能力を評価するための公開データセットとして，現状の LLM に対しても十分な難易度と評価価値を持つことを示している．

5 結論

本稿では，デジタル庁が公開した日本の法令に関する多肢選択式質問応答データセット **LawQA-JP** について，その公開背景，設計方針，データ構造，および内容分析を行った．LawQA-JP は，問題文，4 択選択肢，正答，参照条文情報を体系的に含む公開データセットであり，日本の法令に特有な条文構造や参照関係，例外規定の解釈を要する設問を多く含む点に特徴がある．

また，大規模言語モデルを用いた参照評価を通じて，LawQA-JP が，現状の LLM に対しても安定した正答が容易ではない設問を含むことを確認した．特に，法令条文に関する文脈情報を明示的に与えない条件では，誤答や判断の不安定が生じやすく，日本の法令における構造的推論が依然として難易度の高い課題であることが示唆された．

LawQA-JP は，生成 AI の業務利用における検討や，研究目的での評価を支援するために，公共データ利用規約（PDL）の下で GitHub 上に公開されている．本データセットが，行政機関，企業，および研究コミュニティにおいて，日本の法令理解に関する共通の評価基盤として広く活用されることを期待する．

今後は，対象法令や設問数の拡充に加え，参照条文の粒度や設問タイプの多様化を通じて，より広範な法令理解能力を評価可能なデータセットへと発展させていく予定である．また，検索拡張生成（RAG）手法や法務支援システムの評価への応用など，実務的な利用シナリオに基づく検証も今後の重要な課題である．

参考文献

[1]Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. LexGLUE: A benchmark dataset for legal language understanding in English. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[2]Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Calum Kwan, Ken Satoh, Hiroaki Yamada, and Yoshioka Masaharu. An overview of the coliee 2025 competition: Legal case law and statute law information retrieval and entailment. In *the 20th International Conference on Artificial Intelligence and Law (ICAIL)*, 2025.

[3]Neel Guha, Julian Nyarko, Daniel E. Ho, et al. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. In *Neurips 2023 : Conference on Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 4583531 of *Osgoode Legal Studies Research Paper*, pages 1–143, New Orleans (LA), United States, December 2023. SSRN, Toronto. 143 pages, 79 tables, 4 figures.

[4]Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. CUAD: An expert-annotated NLP dataset for legal contract review. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

表 4　正解に至った 根拠 の特徴量比較（モデル別・コンテキスト有無）

モデル	コンテキスト	平均回答長	第 X 条言及率	項言及率	号言及率	法令名言及率
gemini-2.0-flash	なし	91.2	0.32	0.21	0.09	0.24
	あり	101.7	0.63	0.57	0.33	0.32
gpt-4.1	なし	115.0	0.48	0.34	0.11	0.35
	あり	118.0	0.67	0.48	0.27	0.27
gpt-5-mini	なし	132.6	0.28	0.17	0.07	0.20
	あり	142.8	0.65	0.51	0.32	0.05
gpt-5-nano	なし	152.5	0.43	0.29	0.17	0.38
	あり	149.7	0.63	0.53	0.35	0.18

A　LLM 参照実験から得られた知見

A.1　正解根拠表現の特徴

表 4 は，正解に至った回答の根拠表現（根拠）に着目し，コンテキスト有無による特徴量の違いを示したものである．すべてのモデルにおいて，コンテキスト付与条件では，「第 X 条」「項」「号」といった条文構造への言及率が大きく増加している．

この傾向は，LawQA-JP が，法令構造に基づく推論を促す設問を多く含んでおり，文脈情報の有無がモデルの推論スタイルに直接的な影響を与えていることを示している．

A.2　現状の LLM における難問

参照実験で誤答だった問題の分析に基づき，法令 QA における難問をコンテキスト情報の有無による性能変化の観点から，以下の 2 種類に分類する．

- **Type-A**：コンテキストを付与した場合においても，複数の LLM が一貫して誤答する問題
- **Type-B**：適切なコンテキストを付与することで正答率が大きく改善する問題

Type-A の具体例　以下に，Type-A に分類される代表的な問題例 (抜粋) を示す．

> **問**金融商品取引業者等が顧客の注文について、政令で定めるところにより、最良執行方針等を定めなくてもよいものを教えてください。
>
> **a**　上場株券等の売買で、デリバティブ取引に該当するもの
>
> **b**　店頭売買有価証券の売買で、デリバティブ取引に該当しないもの
>
> **c**　取扱有価証券の売買で、デリバティブ取引に該当しないもの
>
> **d**　金融商品取引法施行令第 16 条の 6 第 1 項第 1 号イ、ロ及びハに掲げるものを除く有価証券の売買

本問では，コンテキスト付与しても 2 つのモデルのみが正答であった．回答は a である．この答えは、金融商品取引法施行令第 16 条の 6 第 1 項第 2 号により、デリバティブ取引に該当する場合は、最良執行方針を定める必要がないものとして除かれるという根拠を見つけ出さないと回答が難しい問題である．LLM は，自然言語的には合理的な連想に基づいて推論を行う一方，条文が課す形式的かつ例外的な要件を十分に優先できていないことが示唆される．

Type-B の具体例　次に，Type-B に分類される代表的な問題例を示す．

> **問**金融商品取引法第 25 条第 1 項により、内閣総理大臣が縦覧書類を受理した日から公衆の縦覧に供しなければならない縦覧書類及び期間の組み合わせとして、誤っているものを教えてください。
>
> **a**　有価証券報告書及びその添付書類 五年
>
> **b**　内部統制報告書及びその添付書類の訂正報告書 五年
>
> **c**　四半期報告書 五年
>
> **d**　自己株券買付状況報告書の訂正報告書 一年

本問では，コンテキスト無し条件において全てのモデルが誤答した一方，コンテキスト有り条件ではすべてのモデルが正答した．回答は c．これは，a,b,d が全て正しいことが確認されない限り，c が間違っているという回答にたどり着かないような事例である．逆に言うと，コンテキスト付与により，a,b,d が正しいことが確認できれば，全てのモデルにおいて残りの 1 つが間違いであるという論理的な解釈が出来たと言える．

以上の分析から，法令 QA における難問は，単にモデルの性能不足に起因するものではなく，問題文の構造や情報提示のあり方に強く依存していることが分かる．Type-A は，常識的推論と法令要件が衝突する問題であり，現行の LLM にとって本質的に解決が困難な課題である．Type-B は，複数の選択肢が正しいことを確認しなければ誤りを特定できない構造を持つ問題であり，その検証に必要な前提情報が問題文中に明示されていない設問である．