複数の LLM を用いた法令 QA タスクの Ground Truth Curation

植松 幸生 [1,2]　大杉直也 [1]
[1] デジタル庁　[2] 東京理科大学　研究推進機構
{yukuemats, naosugi}@digital.go.jp　yukio@rs.tus.ac.jp

概要

本論文では，専門家が作成した法令に関する QA の四者択一問題とその Ground Truth に対して，複数の大規模言語モデル (LLM) を用いて検証することで，Ground Truth に誤りがあることを発見した．発見する方式としては，Ground Truth を与えない状態で，3 種類の異なるファウンデーションモデルの LLM に 0-shot の設定で回答させた．生成された回答から majority voting などの方法で結果を統合し，すべての LLM の答えが一致しない問題を抽出したのち，再度専門家に確認を依頼し，一部問題で Ground Truth が誤っていることを突き止めた．

1　はじめに

大規模言語モデル（LLM）は，自然言語処理タスクのさまざまな分野で高い性能を発揮しており，特に質問応答（QA）タスクにおいても注目を集めている．法令に関する QA タスクでは，正確な Ground Truth（正解データ）が求められるが，人手で作成された正解データには誤りが含まれる可能性がある．誤った Ground Truth は，モデルの評価結果や学習性能に悪影響を与えるため，その信頼性を検証することが重要である．

本研究では，人手で作成した四者択一法令 QA タスク（全 140 問）に対して，複数の大規模言語モデル（LLM）を用いて回答を生成し，その結果から Ground Truth の妥当性を検証した．具体的には，複数の LLM の回答の一致度やアンサンブル手法を通じて「疑わしい」問題を特定した．その後，専門家に再検証したところ，一部の問題において，既存の Ground Truth が誤っていたことを発見した．

本研究の目的は，法令 QA タスクにおける Ground Truth の信頼性を高めるとともに，LLM の出力を用いたデータ検証手法の有効性を示すことである．本論文では，実験の設定，結果，および考察を通じて，法令 QA データの質向上に向けたアプローチについて述べる．

2　関連研究

本章では，Ground Truth Curation と，法令 QA タスクの関連研究について述べる．

Ground Truth の Validation や Curation は従来より多くの分野で実施されてきた．例えば，Yoo[1] 等は，in-context learning における入力とラベルの対応の重要性について述べている．また，古くから Crowd sourcing[2] を活用した，Ground Truth の生成，及びその中で起きるバイアスや共謀などが課題になっている．本研究もこれらの分野と類似した研究であるが，Curation に LLM を利用している点と法判定 QA タスクという特殊なタスクである点が異なる．

法令を扱った QA タスクについてはこれまで多くの関連研究やコンペティションが行われてきた．我々が扱うタスクに最も近いのは日本の国家試験の司法試験である[1)]．司法試験では，毎年短答式問題が出題され，Q に対して正誤を付与するタスクが実施されている．

コンピュータを対象にした法令に関する QA タスクはこれまで多くのコンペティションが開催されてきた．具体的には，Competition on Legal Information Extraction and Entailment (COLIEE) が 10 年以上にわたり開催されている[2)] この中のタスクは，択一問題も扱っており本研究と類似性が高い [3]．また，古くは Text REtrieval Conference(TREC) の Legal Track[4] がある．TREC では，電子情報開示（e-discovery）に関連する情報検索の課題に焦点を当て，法的文書から関連情報を効率的に抽出するタスクを提供するものである．COLIEE や法令 QA タスクと異なり，特定の法令や判例に基づく選択肢問題ではなく，大規模な文書集合内での情報探索が主題である．ただし，法的ドメインでの自動化技術の応用という点で共通点がある．

1） https://www.moj.go.jp/jinji/shihoushiken/jinji08_00241.html
2） https://coliee.org/overview

Oxford University Press では，会社法に関する多肢選択式問題が公開されている[3]．これらは，法令に関する四者択一形式の問題を含むものであり，本研究が扱う法令 QA タスクと形式的に類似している．ただし，この問題集は教育目的で設計されており，法令 QA タスクの研究や評価のためのデータセットとして直接利用されることを意図しているわけではない．

3 法令 QA タスク

本章では，本研究で扱う法令 QA タスクについて説明する．法令 QA タスクとは，法令に関連する質問に対して正確に回答することを目的とした自然言語処理タスクである．本研究では，主に 2 つの目的で利用するために法令 QA タスクの正確なデータセットを作成している．

- LLM の学習向けバリデーションデータセット
- 回答が四者択一形式の法令 QA タスクのデータセット

法令に特化した LLM を作成するために，継続事前学習などで利用するためのバリデーションデータセットが必要になる．その場合，法令を使った単純な QA のデータセットが必要になる．また，日本語で公開されている法令 QA タスクが少ないため，正確な法令 QA タスクのデータセット自体にニーズがあると考える．

3.1 本研究が扱う法令 QA タスク

本研究における法令 QA タスクでは，以下の特徴を持つ設問を対象とする：

形式 4 者択一形式の選択問題
対象範囲 法令条文および関連する政令・解釈指針に基づいた内容
質問の特性 文脈依存型の解釈を求める問題や，特定条文に関する知識の正確性を問う問題

具体的な QA は，コンテキスト，指示，問題文，選択肢，アウトプット記載欄 の 5 つで構成されている．コンテキストとは，問題文を解くために必要な法令を抜粋したものである．本タスクでは，基本的にこのコンテキストを読むことで正解を選択可能なタスクになっている．指示とは，コンテキストに記載した情報を用いて，選択肢を a,b,c,d から 1 つ選べという指示である．問題文には法令 QA の Q が記載され，その後に選択肢が a,b,c,d の形式で 4 つリストされる．アウトプット記載欄は空欄で，そこに LLM がアウトプットを入れる．

例えば，**金融商品取引法**に関連する問題では，法令条文および政令の内容を基に，特定の条項や適用範囲を正確に判断することが求められる．具体的には，以下のような質問が考えられる：

> 金融商品取引法第 2 条第 2 項第 5 号ニにおいて，有価証券とみなさなくても公益または出資者の保護に支障を生じないとされる権利を選択肢から選べ．

選択肢としては，弁護士の業務を行う事業のみを対象とする組合契約に基づく権利など，法令の詳細な理解が必要な内容が含まれる．

このように，法令 QA タスクは法令の文脈および条文解釈に基づいた正確な回答生成が求められるため，Ground Truth の信頼性が特に重要である．

3.2 法令 QA タスクにおける Ground Truth の作成とその問題点

前節で示した法令 QA タスクは高い専門性が求められるため複数の弁護士に作成を依頼する必要があるため非常にコストがかかる．また，Ground Truth を作成する過程で以下のような問題が起きる可能性がある．

1. コンテキスト (引用される文書) の間違い
2. 選択肢作成時の間違い
3. 単純な答えの間違い

引用される文書の間違いとは，コンテキストとして与える文書そのものが間違いだったり，不足している例である．4 択の答えの作成間違いとは，4 択の中に正答が存在しない，あるいは複数の答えが正解になる場合である．最後に，単純な答えの間違いとは，作成過程におけるバグ等の理由で，答えの指定が間違えている場合である．

4 複数の LLM を用いた Ground Truth Curation

本研究では，複数の LLM を用いて Ground Truth の間違いを発見する．これは，複数の LLM を用いることで，複数の Assessor に正解作成の依頼したことと同様の効果があるかを検証する．

4.1 複数の LLM のアウトプットを統合するアンサンブル手法

複数の LLM の回答を統合する方法は，従来から用いられている複数の機械学習を統合するアンサンブル学習 [5] を LLM に応用する．表 4 に，本研究で用いたアンサンブル手法について比較する．

4.2 Majority voting

各モデルが選択肢 j のいずれか 1 つに対して 1 票投じ，その投じた票が最も大きい選択肢 j を出力とする方式である．

$$S_j = \sum_{i=1}^{n} \mathbb{I}(y_i = j)$$

ここで:

$$\mathbb{I}(y_i = j) = \begin{cases} 1 & \text{モデル } i \text{ が選択肢 } j \text{ を選んだ場合} \\ 0 & \text{それ以外の場合} \end{cases}$$

3） https://global.oup.com/uk/orc/law/company/roach4e/resources/mcqs/

本研究では，3 つのモデルを利用しているため，2 票以上が投じられた選択肢が選ばれることになる．仮に 1 票ずつで 3 つの選択肢が同点となった場合は，回答が最も上にある 1 つが選択されたこととした．(a,b,c が 1 票ずつの場合，a が選択される)

4.3 Soft voting

各モデルが選択肢 j に対して出力する確率 $p_{i,j}$ を最終スコアとして算出する．具体的には，1 つの設問 i に対する 4 つの選択肢の合計が 1 になるように p_i を調整する

$$S_j = \sum_{i=1}^{n} p_{i,j}$$

この，p を得るために以下に示すプロンプトを支持に追加した．

> 以下の問題文に対する回答を、選択肢 a、b、c、d に対して各選択肢の確率を足すと 1 になるように 0-1 の間でコンマに続いて出力してください。(ex. a:0.55,b:0.35,c:0,d:0.1)

稀に，合計が 1 に満たないことや，答えが指定した形式でない場合等があったが，その場合は当該モデルの出力が無かったこととして実験を進めた．

4.4 モデルの信頼度を考慮した Soft voting

各モデルが選択肢 j に対して出力する確率 $p_{i,j}$ にモデルの信頼度 w_i を重み付けし，最終スコアを計算する方法である．前述した Soft voting では，各モデルの出力の信頼度が同じであることが前提となっていたが，各モデルの精度が異なるため本手法を採用した．

$$S_j = \frac{\sum_{i=1}^{n} w_i \cdot p_{i,j}}{\sum_{i=1}^{n} w_i}$$

ここで:

- $p_{i,j}$: モデル i が選択肢 j に出力する確率
- w_i: モデル i の重み（事前に設定された信頼度）

モデル i の信頼度 w_i を検証用データから得た正解率に基づいて算出した．

$$w_i = \frac{\text{正解数}_i}{\text{全検証問題数}}$$

表 3 に，検証データで算出した各モデルの信頼度を示す．この信頼度 w に基づいて，後述の実験を進めた．

5 実験

Ground Truth Curation の評価を実施するために大きく 2 つの実験を実施した．

実験 1 法令 QA タスクの難易度を確認する実験
実験 2 Ground Truth Curation の結果を確認する実験

まず，実験 1 として法令 QA タスクの難易度を確認するために，人手で作成した Ground Truth を利用して，各 LLM および前述したアンサンブル手法の正答率を算出する．次に，実験 2 として実験 1 を実施した結果，LLM の正答に一貫性があり，Ground Truth に疑義がある設問を抽出し，再度専門家に設問と Ground Truth の確認を実施する．

5.1 評価データセットの作成

本研究では，法令 QA タスクの評価データセットを，設問検証用として 70 問，実験に利用するテストデータとして 140 問のデータセットを作成した．各設問は，以下の情報を含む：

コンテキスト 必要な条文，政令の抜粋，簡単な解説文
質問文 「次の選択肢のうち正しいものはどれか」「誤っているものはどれか」など
選択肢 a,b,c,d 1 つが正解，3 つが誤り
正解（Ground Truth） 専門家が判断

コンテキストには，回答に必要な条文や政令の抜粋を入力し，専門家はこのコンテキストの情報のみで正解の選択肢を選ぶことが出来るようにした．後述する実験では，このコンテキストを与えて LLM に回答させた場合と，コンテキストを与えない場合の両方を実施する．

検証用データの 70 問は，前述した各モデルの信頼度算出のために利用した．テストデータ 140 問は，過去の文献で選択肢の Order sensitivity 問題 [6] が指摘されており，それを取り除くために選択肢 a,b,c,d をランダマイズして同じ設問を 4 つの設問にし，560 問の設問を作成しテストデータとして利用した．Ground Truth は，弁護士 2 名の合意により作成された．

5.2 実験に利用したモデル

実験では，以下の 3 つの LLM を用いた．各ファウンデーションモデル (以降モデル) を選定した理由は，現在広く利用されているモデルであることと，それぞれ独自のデータソースから作られていると想定しており，日本語も扱えることが確認されているからである．

モデル A (llama-based) llama 系列をベースに独自調整を施したモデル
モデル B (gpt4) OpenAI GPT-4 API を利用
モデル C (gemini1.5) Google 社が提供 API を利用

5.3 実験 1: 法令 QA タスクの正答率

実験 1 では，法令 QA タスクの難易度と，専門家の知識を有するような問題であるかを確認するために，専門家が作成した Ground Truth を使って，各モデルの正答率を算出した．モデル i の正答率 A_i は以下の式から算出される．

$$A_i = \frac{\text{正解数}_i}{\text{全設問数}} \times 100$$

ここで，正解数$_i$ は， モデル i が正解した設問数であり，全設問数 とは，データセット全体の設問数である 560 件となる．

表 1 に，各モデルの正答率を示す．コンテキスト

を与えた場合（関連する法令条文が提示されている場合）には，全モデルが大幅に正答率を向上させている．コンテキストがある場合は，gemini が最も高い正答率となり，次いで gpt4, llama based が続いている．コンテキストが無い場合は，相対的には変わらない結果となったが，正答率が約 50%程度のタスクであることが分かった．

本実験結果から，正答に必要な情報を与えることで，LLM は人間とほぼ同等の正答率で回答を返すことが出来るが，pre-trained の gpt では，法令に関する情報が不足していることが推測されるため，正確な答えを返すまでには至らないことが分かった．次に，各モデルの出力をアンサンブルした場合の結果を表 2 に示す．Majority voting では，gpt4 単体モデルよりも若干高い正答率を示し，Soft voting 及び Soft voting + 信頼度考慮によってさらにわずかな正答率向上が得られた．また，単体の性能と比較すると，gemini1.5 単体の方がアンサンブル手法よりも高い精度になることが分かった．

表 1　法令 QA タスクの正答率 (各モデル単体)

モデル	コンテキストあり	なし
llama based	81.4%	47.0%
gpt4	86.1%	49.3%
gemini1.5	**95.0%**	**54.3%**

本実験から，アンサンブル手法自体は有効ではあるが，単体性能がある程度高いモデル同士を利用する方が回答率のさらなる向上に寄与する可能性が高いことを示唆している．

表 2　法令 QA タスクの正答率 (アンサンブル)

モデル	コンテキストあり	なし
Majority voting	86.2%	51.3%
Soft voting	86.3%	51.4%
Soft voting+信頼度	**86.6%**	**51.6%**

5.4　実験 2: Ground Truth Curation の実験と結果

実験 1 で利用した 3 つのモデルを使って，Ground Truth の間違いが存在するのか実験を行った．実験 1 から，現状の LLM はコンテキストが無いと高精度に回答できないことが分かっているので，今回はコンテキストありのテストデータを利用した．本実験では，Ground Truth 以外の LLM のアウトプットが 2 つ以上一致している設問について抽出し，専門家への回答の見直しを実施した．例えば，以下の事例は，GT(Ground Truth) 以外すべて a と回答した事例である．

QID	GT	model1	model2	model3
金商法 第 6 章 27	d	a	a	a

結果を表 5 に示す．全部で 7 件 (a,b,c,d をランダムにする前の 140 件中) が抽出された．表中の *QID* が設問 ID で，それに対して修正の必要性があったかどうか，そして，修正の理由を記載している．表からわかる通り，全 7 件中 4 件で何かしらの問題があることが分かった．この中で，最も致命的な回答記載ミスが 1 件，その他 3 件も問題文の記載に不明瞭な点があり，修正をしたものが 3 件あった．誤検知（実際は Ground Truth があっていたもの）が 3 件あったが，その中には，問題文の解釈が難しいもの（ex. 当てはまらないものを択一式で選択する）や，コンテキストからは正しくないことしか判断できないもの等人間が見ても難しい問題が多く抽出されていた．

5.5　実験結果の考察

実験の結果，複数の LLM の回答を組み合わせることで，(1) 単独モデルでは見逃していた問題点を抽出しやすくなったこと，(2) Ground Truth 自体に含まれる不備を効率的に発見できたことが分かった．

具体的には，不一致問題に対して再度専門家が検証したところ，**条文引用不足**，**答えラベルの単純ミス**，**複数の選択肢が正解になる** といったケースが確認された．これは，法令 QA タスクのように専門性が高い領域において，人手で作成する Ground Truth にも一定のエラー率が含まれ得ることを示唆する．

6　結論

本研究では，複数の LLM を用いて法令 QA タスクの Ground Truth を検証するアプローチを提案し，実際に不備を含む問題を特定して修正した．法令のような専門領域では，問題作成の段階で誤りが混入するリスクが高く，従来は専門家による時間的コストのかかる再検証が必要だった．しかし，LLM の回答をアンサンブルし，不一致箇所を重点的に確認する手法により，Ground Truth の正確性を効率的に向上できることが示唆された．

今後の課題としては，(1) 法令ドメインに特化した LLM のさらなる性能向上，(2) モデルが生成する根拠の自動評価手法の確立，(3) 法解釈の揺れが大きい問題への対処などが挙げられる．いずれにせよ，本研究の結果は法令 QA タスクのデータキュレーションや LLM を活用した自動評価の有効性を示すものであり，他の専門領域（医療，特許など）にも応用可能な知見を提供する．

参考文献

[1] Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang goo Lee, and Taeuk Kim. Ground-truth labels matter: A deeper look into input-label demonstrations, 2022.

[2] Changyue Song, Kaibo Liu, and Xi Zhang. Collusion detection and ground truth inference in crowdsourcing for labeling tasks. **J. Mach. Learn. Res.**, Vol. 22, No. 1, January 2021.

[3] Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. Overview of benchmark datasets and methods for the legal information extraction/entailment competition (coliee) 2024. In Toyotaro Suzumura and Mayumi Bono, editors, **New Frontiers in Artificial Intelligence**, pp. 109–124, Singapore, 2024. Springer Nature Singapore.

[4] Stephen Tomlinson and Bruce Hedin. **Measuring Effectiveness in the TREC Legal Track**, pp. 167–180. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[5] Jinliang Lu, Ziliang Pang, Min Xiao, Yaochen Zhu, Rui Xia, and Jiajun Zhang. Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models, 2024.

[6] Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. Can multiple-choice questions really be useful in detecting the abilities of llms?, 2024.

表 3 検証データにおける各モデルの信頼度

モデル	コンテキストあり	なし
llama based	84.3%	45.7%
gpt4	84.3%	50.0%
gemini1.5	91.0%	58.6%

表 4 アンサンブル手法の比較

手法	数式	入力
Majority voting	$S_j = \mathrm{argmax}_j \left(\sum_{i=1}^{n} \mathbb{I}(y_i = j)\right)$	全モデルの出力
Soft voting	$S_j = \mathrm{argmax}_j \left(\sum_{i=1}^{n} p_{i,j}\right)$	モデルの確率出力 $p_{i,j}$
モデルの信頼度を考慮した Soft voting	$S_j = \frac{\sum_{i=1}^{n} w_i \cdot p_{i,j}}{\sum_{i=1}^{n} w_i}$, $w_i = \frac{\text{正解数}_i}{N}$	モデルの確率出力 $p_{i,j}$, 信頼度 w_i

表 5 Ground Truth Curation 結果

QID	修正	理由
金商法 第 6 章 73	No	誤検知
薬機法 第 5 章 38	No	誤検知
薬機法 第 5 章 46	Yes	問題文修正
金商法 第 6 章 27	Yes	問題文修正
金商法 第 2 章 34	No	誤検知
薬機法 第 15 章 19	Yes	回答記載ミス
借地借家法 第 3 章 19	Yes	問題文修正