マルチエージェント推論による 法令適用能力の向上に関する取り組み

前田 竜聖[1,a)]　内山 光彩[1,b)]　有村玲音[1,c)]　高橋 哲朗[1,d)]　小野 智司[1,e)]

概要：近年，大規模言語モデル（Large Language Models：LLM）の法務領域への応用が進む一方で，法的根拠との対応付けの不整合や例外規定の見落としにより，誤答が生じる点が課題となっている．本研究では，複数の LLM エージェントを組み合わせてディベート，分業，多数決を行うことで，法令適用能力の改善を試みる．デジタル庁が公開している法令 4 択問題を対象に，提案手法を LLM エージェント単体に対するプロンプティング方式と比較を行い，その有効性を検証した．実験の結果，マルチエージェント方式は単体方式に比べて一貫して高い正答率を示し，マルチエージェント方式に基づく推論過程の相互検証と集約は，法令 4 択問題における正解率の改善に寄与する可能性が示唆された．

An Attempt to Enhance Legal Statute Application through Multi-Agent Reasoning

Abstract: As the application of Large Language Models (LLMs) in the legal domain has advanced, challenges remain regarding errors caused by inconsistent grounding in legal authorities and the omission of exception clauses. This study aims to improve the capability to apply statutory provisions by employing a multi-agent framework that incorporates debate, role specialization, and majority voting. We evaluated the effectiveness of the proposed method by comparing it with a single-agent prompting baseline using four-option multiple-choice legal questions published by the Digital Agency of Japan. Experimental results demonstrated that the multi-agent approach consistently achieved higher accuracy than the single-agent approach, suggesting that this performance enhancement arose from the mutual verification and aggregation of reasoning processes within a multi-agent system.

1. はじめに

近年, 大規模言語モデル（Large Language Models; LLM）の発展により，一般的な言語タスクのみならず，司法試験問題の解答や契約書レビューといった専門的な法務領域への応用が模索されている．しかし，LLM を単体で用いる場合，学習データに含まれない最新の法令への追従が困難である点や，ハルシネーションを生じうる点から，正確性に課題が残る．これに対し，外部知識を検索，統合する検索拡張生成（Retrieval-Augmented Generation; RAG）は，生成時に根拠となる情報を参照できるため，法令知識を要するタスクへの有効性が期待されている．

本研究では，法令 4 択問題に対して単体エージェント方式とマルチエージェント方式を実装し，それぞれの有効性について検討する．具体的には，デジタル庁が公開する法令 4 択問題データセットを対象に，推論の方式による違いが正解率に与える影響を比較評価する．

本稿の構成は以下のとおりである．2 節では，法務領域における LLM 活用およびプロンプトエンジニアリングに関する関連研究について述べる．3 節では，本研究で実装する各推論方式（単体プロンプティング，ディベート，分業，多数決）について説明する．4 節では実験結果と考察を示し，5 節で結論を述べる．

2. 関連研究

本研究は，日本の法令知識を要する多肢選択式質問応答（MCQA）に対して，根拠提示を伴う RAG と，マルチエー

[1] 鹿児島大学
Korimoto, Kagoshima, Kagoshima 8900065, Japan
[a)] k5881533@kadai.jp
[b)] k7719654@kadai.jp
[c)] k2927194@kadai.jp
[d)] takahashi@ibe.kagoshima-u.ac.jp
[e)] ono@ibe.kagoshima-u.ac.jp

ジェント推論の集約によって正答率向上を目指す．以下では，(i) 法務領域における LLM の課題，(ii) 法令に関する質問応答（法令 QA）データセット，(iii) RAG と検索設計，(iv) 推論プロンプティングおよびマルチエージェントに関する研究を整理し，本研究の位置づけを述べる．

2.1 法務領域における LLM とハルシネーション

LLM は条文検索，文書要約，QA などの法務支援への応用が進む一方で，条文番号，要件，例外規定の取り違えなどのハルシネーションは重大なリスクとなる．ハルシネーションの類型や要因，評価上の課題はサーベイで整理されており [1]，厳密な根拠を要する法令 QA では，外部根拠に基づく生成と検証が重要となる．

2.2 法令 QA データセットとベンチマーク

英語圏では，法務 NLP のベンチマークとして LexGLUE [2], LLM の法的推論能力の測定を目的とする LegalBench [3] などが整備されている．日本語では，デジタル庁が e-Gov 法令への参照情報を付与した日本の法令に関する多肢選択式 QA データセット（法令 4 択問題）lawqa_jp を公開しており [4]，正答根拠の整備を含むデータ品質向上の取り組みも報告されている [5].

本研究は lawqa_jp を対象に，軽量モデルでも再現可能な構成で，単体プロンプティングと複数のマルチエージェント推論方式を同一条件で比較する．

2.3 RAG と検索設計

RAG は，生成モデルの内部知識（学習済みパラメータに埋め込まれた知識）だけに依存せず，外部文書集合から検索した文書を条件として回答を生成する枠組みである [6]. 検索器としては，密なベクトル表現に基づく DPR（Dense Passage Retrieval）が代表的であり，質問と文書をそれぞれエンコードして近傍検索を行う [7]．法令 QA では，参照すべき法令名，条，項，号を正しく特定する必要があるため，取得単位（条，項，号など），但書，例外の含め方，クエリ設計といった検索構成が性能に影響することが指摘されている [15, 16].

本研究においても，法令 4 択問題における根拠提示のため，根拠条文の取得と提示を重要な要素として扱い，推論過程に組み込む．

2.4 推論プロンプティングとマルチエージェント集約

推論過程を明示させる Chain-of-Thought（CoT）[8] や，複数の推論経路をサンプリングし最頻の解を採用する Self-Consistency [9] は，単発生成よりも推論の頑健性を高める方向性として知られる．さらに，複数エージェントの対話や役割分担を扱う枠組みとして AutoGen [10]，役割対話により探索を行う CAMEL [11] などが提案されている．集約方法の観点では，討論と投票の比較 [12] などが報告され，合意度の閾値設定などの集約設計が性能を左右し得ることが示されている．本研究では，正誤判定が容易な法令 4 択問題を対象に，ディベート，分業，多数決といった代表的な集約方式を実装し，合意閾値等の設計も含めて比較評価する．

3. 研究方法

3.1 概要

本研究の目的は，デジタル庁が公開する法令 4 択問題データセットに対して，複数の LLM エージェントによる推論，集約を導入することによる効果を，単体エージェントと比較した場合の正解率の変化に着目することで検証する．この目的のため，単体エージェントに回答手順を付与するプロンプティング方式（方式 1，2 および 3）と，複数エージェントを用いるマルチエージェント方式（方式 4，5，6 および 7）を実装し，法令 4 択問題データセットの正答率で比較する．なお，各方式で用いる LLM，検索設定，および評価問題は統一し，推論方式のみを比較対象とする．

比較対象は以下の 7 方式であり，プロンプト設計およびエージェントの利用形態が互いに異なる．

(方式 1) ベースライン：単体の法律エージェントにより法令 4 択問題を解く．

(方式 2) 法的三段論法プロンプティング方式：単体エージェントに法的三段論法の手順を付与して解く．

(方式 3) 消去法プロンプティング方式：単体エージェントに消去法の手順を付与して解く．

(方式 4) マルチエージェントディベート方式：複数エージェントが議論し，合意度に基づき最終回答を決定する．

(方式 5) 法的三段論法マルチエージェント方式：法的三段論法の工程を複数エージェントで分担して解く．

(方式 6) 消去法マルチエージェント方式：消去法の工程を複数エージェントで分担して解く．

(方式 7) 多数決方式：複数エージェントの回答を多数決で集約して解く．

各設問の入力は，問題文 q と選択肢文の集合 $\{c_l\}_{l\in\mathcal{L}}$ であり，$\mathcal{L} = \{\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}\}$ は選択肢ラベル集合である．出力は，正しいと判断した選択肢ラベル $\hat{l} \in \mathcal{L}$ とする．評価指標は正答率（Accuracy）であり，評価問題数を N，正解ラベルを l_i とすると，$\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\hat{l}_i = l_i]$ により算出する．

Example 1（金融商品取引法 - 正しいものを選ぶ):
【Legal Provisions】
金融商品取引法第 24 条第 1 項：有価証券報告書の提出について、外国会社は事業年度経過後六月以内に提出しなければならない.

【Question】
外国会社が有価証券報告書を提出する場合の期限として、正しいものを教えてください.
【Choices】
a 三月以内
b 六月以内
c 九月以内
d 一年以内
【Answer】b

図 1 ベースライン（方式 1）で用いた few-shot 例示（抜粋）

3.2 単体エージェント方式（方式 1，2 および 3）

3.2.1 ベースライン（方式 1）

方式 1 では，LLM に「法令 4 択問題を解く法律エージェント」としての役割を付与するため，few-shot 例示（少数事例の提示）を用いる．few-shot 例示の一部を図 1 に示す．プロンプトでは，(1) 問題文と 4 つの選択肢文を読み取り，(2) 参照情報として与えられる根拠条文（取得条文）を踏まえて判断し，(3) 最終的に選択肢ラベルのみを明示して回答する，という手順と出力形式を指示した．これにより，単体エージェントにおける基本性能をベースラインとして評価する．

3.2.2 法的三段論法プロンプティング方式（方式 2）

方式 2 では，法令判断の典型的な枠組みである法的三段論法 [13] に基づき，推論手順を明示することで，正解率の向上を図る．図 2 は，法的三段論法で推論を行うときの推論手順と出力フォーマットの指定を例示したものである．方式 2 は，次の手順を単体エージェントに明示し，規範（条文）と選択肢文の対応付けを段階的に行わせることで，推論の明確化を図る．

(1) 規範の確認（大前提）：提供された法令から本問に関わる条文を特定し，要件と効果を整理する．

(2) 事実のあてはめ（小前提）：問題文および各選択肢文が，要件に合致するかを順に検討する．

(3) 結論：検討結果に基づき，最も適切な選択肢ラベルを選択する．

また，根拠条文のどの部分に依拠したかが不明瞭にならないよう，大前提の記述では取得条文の参照箇所を明示するよう指示を行う．最終出力は方式 1 と同様に選択肢ラベルとし，評価は最終回答の正誤により行う．

解答プロセス:
■ 規範の確認:
金融商品取引法第 8 条第 1 項によれば、届出の効力発生の「原則」は、届出書を受理した日から「15 日を経過した日」である.
一方、同条第 3 項によれば、一定の要件（公衆に容易に理解される、または情報が広範に提供されている等）を満たす場合の「例外」として、内閣総理大臣は、(1)15 日に満たない期間の指定、(2) 直ちに効力を生ずる旨の通知、(3) 受理日の翌日に効力を生ずる旨の通知、を行うことができる.

■ 事実のあてはめ:
- 選択肢 a:「容易に理解される場合」に「15 日に満たない期間を指定」できるとする記述は、第 8 条第 3 項の規定と一致する.
- 選択肢 b:「情報が広範に提供されている場合」に「直ちに効力を生ずる旨を通知」できるとする記述は、第 8 条第 3 項の規定と一致する.
- 選択肢 c:特別な通知がない場合の効力発生時期を「10 日を経過した日」としているが、これは第 8 条第 1 項の「十五日を経過した日」という規定と矛盾する.
- 選択肢 d:特別な通知がない場合の効力発生時期を「15 日を経過した日」とする記述は、第 8 条第 1 項の原則規定と一致する.

■ 結論:
最終回答：c

図 2 法的三段論法プロンプティング方式（方式 2）で用いた例示（抜粋）

3.2.3 消去法プロンプティング方式（方式 3）

方式 3 では，法令 4 択問題における実務的な解法として消去法を明示し，誤選択肢文の排除を通じた正解率向上を図る．図 3 は，消去法を用いて推論を行うときの推論手順と出力フォーマットの指定を例示したものである．

消去法プロンプティングでは，次の手順を単体エージェントに明示し，候補の絞り込みと最終選択を分離することで，明らかな誤りの早期除外を試みる．

(1) 明らかに不適切な選択肢文を除外：4 択から不適切な選択肢を除外し，候補を絞り込む.

(2) 残りの候補の精査：残った候補を比較し，最も適切な選択肢文を選択する.

最終出力は方式 1，2 と同様に選択肢ラベルのみとし，推論過程の記述は評価対象に含めない.

【解答（出力フォーマット厳守）】
問題タイプ：誤っているもの
除外：a，b
除外理由：
- a：第 8 条 1 項の原則と一致しており「誤り」ではなさそう
- b：第 8 条 3 項の例外と一致しており「誤り」ではなさそう
二択：c vs d
比較：
- c：原則は 15 日経過（第 8 条 1 項）なので「10 日」は条文と矛盾しやすい
- d：第 8 条 3 項に「直ちに効力」を通知できる旨があり整合しやすい
最終回答：c

図 3 消去法プロンプティング方式（方式 3）で用いた例示（抜粋）

3.3 マルチエージェント方式（方式 4，5，6 および 7）

3.3.1 マルチエージェントディベート方式（方式 4）

方式 4 では，3 体の LLM エージェント（Debater A，Debater B，Moderator）を用いる．2 体の Debater が互いに主張を提示し，Moderator が合意度に基づいて最終回答を決定する．

ディベートの例を図 4 に示す．手順は以下のとおりである．

(1) Debater A が候補選択肢文と理由を提示する．

(2) Debater B が候補選択肢文と理由（例外，制約の観点）を提示する．

(3) Moderator が両者の主張から合意度を算出し，合意度が閾値 τ 以上なら最終回答を確定する．

(4) 合意度が τ 未満の場合，Debater A が反論に応答し，再議論する．最大 3 ラウンドまで繰り返し，打ち切り時は Moderator が最終回答を選出する．

ここで合意度は $[0, 1]$ のスコアとして定義し，方式 4 では，合意の閾値を $\tau \in \{0.8, 0.98\}$ とする．

3.3.2 法的三段論法マルチエージェント方式（方式 5）

方式 5 は，方式 2 で用いた法的三段論法の手順を，複数エージェントで分担して直列に実行する方式である．法的三段論法の各段階を 3 体のエージェントに割り当て，「規範の確認」，「要件のあてはめ（各選択肢文の評価）」，「結論の算出」を順に実行するよう設計する．

(Agent 1) 規範の確認（根拠条文から判断基準を要約）

(Agent 2) 要件のあてはめ（各選択肢文の評価）

(Agent 3) 結論の算出（最終回答の決定）

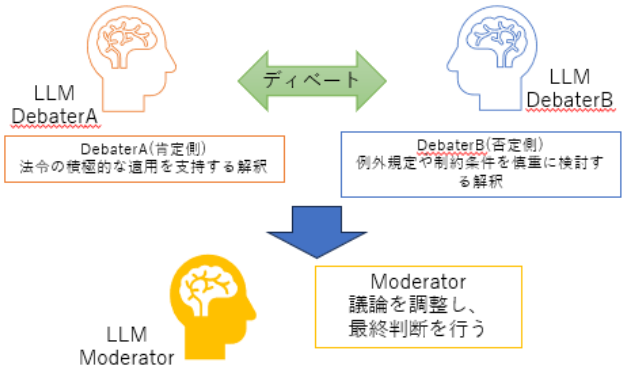

図 4 マルチエージェントディベート方式（方式 4）の概念図

前段の出力（規範，評価理由）を次段の入力として引き継ぎ，最終段のエージェントが選択肢ラベルを出力する．

3.3.3 消去法マルチエージェント方式（方式 6）

方式 6 は，方式 3 で用いた消去法の手順を，複数エージェントで分担して直列に実行する方式である．消去法の工程を 2 体のエージェントに割り当て，「不適切な選択肢文を除外」，「残りの候補の精査」を順に実行するよう設計する．

(Agent 1) 明らかに不適切な選択肢文を除外

(Agent 2) 残りの候補の精査

なお，方式 3 では「除外」により 2 択まで絞るよう指定するが，分業構成では除外担当が誤って正答を除外するリスクを考慮し，2 択に限定する指定は行わない．前段の除外理由と残候補を後段に引き継ぎ，後段のエージェントが選択肢ラベルを出力する．

3.3.4 多数決方式（方式 7）

方式 7 では，5 体の法律エージェント（Voter）を用い，各エージェントの回答を多数決で集約する．同一プロンプト，同一順序では回答が同質化しやすいため，図 1 の基本プロンプトに加えて，各エージェントに異なる役割（観点）を付与し，選択肢文の順序を変えるとともに，`temperature` をわずかに変更して多様性を確保する．各エージェントに付与する役割（観点）を以下に示す．

(Voter 1) 条文の文言一致重視（literal）

(Voter 2) 例外, 但書, 期限数値重視（exceptions/numbers）

(Voter 3) 定義, 要件構造重視（definitions/requirements）

(Voter 4) 「誤っているもの」なら矛盾探し優先（contradiction finder）

(Voter 5) 迷ったら最小断定（least-commitment）

各 Voter が法令 4 択問題を解いた後，選択肢ラベルに対して多数決を行う．最多得票の選択肢ラベルを最終回答と

する．最多得票が同数の場合（例： a に 2 票，b に 2 票，c に 1 票，d に 0 票）は，上位 2 択のみを対象として再度多数決を行い，最終回答を決定する（例：a および b のみで再投票を行う）．

4. 評価実験

4.1 実験設定

3 節に示した 7 種類の方式を比較検討し，(i) 単体エージェントに推論手順を付与するのみで性能が向上するか，および，(ii) 複数エージェントによる相互検証（ディベート），工程分割（分業），集約（多数決）が性能に寄与するかを検証するため，デジタル庁が公開する日本の法令に関する多肢選択式 QA データセット（lawqa_jp）[4] を用いて実験を行った．

各問題は，設問文，選択肢文，正答ラベルを含み，参考資料として e-Gov 法令 URL が付与される．計算資源の制約から，評価用に 40 問を抽出して比較評価を行った．40 問の抽出は，(i) e-Gov URL を `law_list.json` で法令タイトルへマッピングし，(ii) 法令タイトルごとに最低 1 問を確保した上で残りを出現比率に比例配分し，(iii) 各法令内では問題タイプ（正しい選択/誤り選択/組み合わせ/その他）の不足分を優先して貪欲に選択する，という層化抽出を行った．乱数シードは 42 とし，Python の hash に依存しない決定的なシャッフルにより再現可能とした．以上により，特定法令への偏りを抑えつつ，多様な問題タイプを含む 40 問を構成した．

各方式において，LLM モデルは Ollama を介して Qwen3:8b を使用した．生成は法令 4 択問題の決定性を重視し，評価時は `temperature`=0.0 で 1 回試行とした．

評価指標は 4 択の正解率（Accuracy）とした．すなわち，各方式について，40 問中の正答数を 40 で除した値を正解率として算出した．加えて，問題ごとの正誤分布を比較し，方式間の得手不得手の傾向を分析した．

4.2 実験結果

表 1 に実験結果を示す．単体エージェント群（ベースライン，法的三段論法，消去法）の正解率はいずれも 47.5%であり，三段論法および消去法をプロンプトに組み込むのみでは正解率の改善を確認できなかった．一方，マルチエージェント群はいずれも単体群を上回り，複数視点の相互検証，分業，集約が有効である可能性が示された．特に多数決方式は 60.0%で最も高く，単体の法律エージェントに対して 12.5 ポイントの改善がみられた．

方式 5，6 および 7 などの分業方式では，法的三段論法（分業）が 57.5%，消去法（分業）が 55.0%であり，単体方式（いずれも 47.5%）と比較して向上が確認された．このことから，工程分割により，条文要件と選択肢文の対応付けにおける誤りが抑制された可能性がある．また，方式

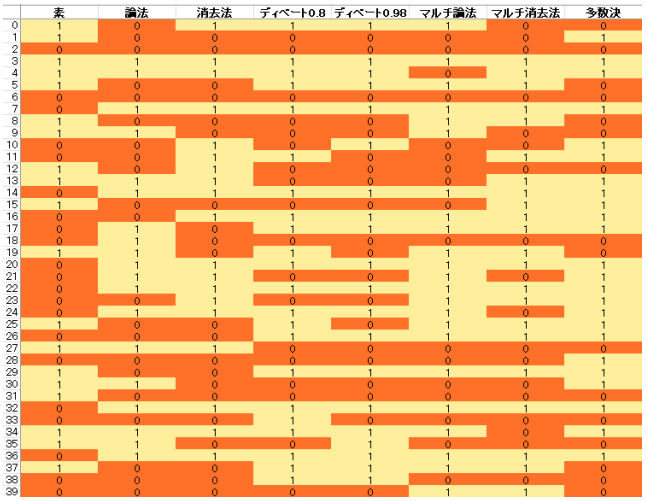

図 5 エージェントごとの正誤（黄色：正，オレンジ：誤）

4 のディベート方式は閾値設定により挙動が変化し，閾値 0.8 では 55.0%であった一方，閾値 0.98 では 50.0%に低下した．

表 1 に示す 8 方式のうち正答した方式の数に着目して問題数を集計した結果を表 2 に示す．本稿ではディベート方式を閾値 0.8 および 0.98 の 2 条件として扱うこととした．表 2 より，過半数（5 方式以上）が正答した問題は 20 問であり，全体の 50%を占める．一方で，1 つまたは 2 つの方式のみが正答した問題も 8 問存在し，方式ごとに得手不得手が異なることが示唆された．図 5 に，方式ごとの正誤を示す．0 は誤答を，1 は正答であることを示す．問 3，4，7 など，多くの方式が正解した問題がみられる一方で，問 2，6 のようにすべてのエージェントで誤答であった問題や問 18，28，31，33 のように 1 エージェントのみが正解した問題を確認した．方式別の誤答要因については，次節以降で考察する．

4.3 考察

4.3.1 方式 4（マルチエージェントディベート方式）による誤答分析と考察

本節では，方式 4 のディベート型マルチエージェント方式について，誤答の要因を整理する．誤答は主として，以下の 3 類型に分けられる（括弧内は件数）．

(1) 両ディベーターが誤答で，同一の選択：2 体が同じ選択肢ラベルを選び誤答となった（10 件）．

(2) 両ディベーターが誤答で，選択が不一致：2 体が異なる選択肢ラベルを選ぶが，いずれも誤答であった（6 件）．

(3) 片方は正答だが採用されない：一方が正答を選ぶが，もう一方の誤答が採用された（2 件）．

誤答 18 問のうち，両ディベーターが正答に到達できな

表 1 各方式の正解率（40 問）

種類	方式	Accuracy [%]
単一エージェント	方式 1: ベースライン（単体）	47.5
	方式 2: 法的三段論法プロンプティング（単体）	47.5
	方式 3: 消去法プロンプティング（単体）	47.5
複数エージェント	方式 4: マルチエージェントディベート（閾値 0.8）	55.0
	方式 4: マルチエージェントディベート（閾値 0.98）	50.0
	方式 5: 法的三段論法マルチエージェント（分業）	57.5
	方式 6: 消去法マルチエージェント（分業）	55.0
	方式 7: 多数決（5 エージェント）	60.0

表 2 問題ごとの正答数

正解した方式の数 (/8)	問題数 (/40)
0	2
1	4
2	4
3	7
4	3
5	7
6	4
7	8
8	1

かった問題が 16 問を占めた．したがって，合意度設計のみでの改善には限界があり，ディベーター自身の回答性能（根拠条文の特定や要件の対応付け）を高めることが重要であると考える．

4.3.2 方式 5（法的三段論法マルチエージェント方式）による誤答分析と考察

法的三段論法（分業）方式について，図 5 をもとに誤答と他方式の正誤分布の関係を比較すると，多くの方式が正解した問題では，本方式も正答しやすい傾向が見られる．一方で，本方式が誤答した 17 問のうち，途中段階のエージェントの出力が崩れたことで，後続処理が破綻した問題が 3 問確認された．このことから，分業方式では，各段階の出力検証や受け渡し形式の厳格化が性能安定化に寄与すると考える．

4.3.3 方式 6（消去法マルチエージェント方式）による誤答分析と考察

消去法（分業）方式について，誤答要因を整理する（括弧内は件数）．

(1) 絞り込み段階での誤り：除外担当が正答選択肢ラベルを除外していた（9 件）．

(2) 最終選択段階での誤り：残りの候補からの最終選択が誤りであった（9 件）．

詳細に確認すると，4 択から十分に絞り切れない状況で，末尾の選択肢「d」を機械的に除外する挙動が 7 問で観測された．この挙動に起因する誤答が 2 問確認された．また，除外担当エージェントが本来の役割を超えて，最終回答まで選出しようとするケースも見られ，役割分担（出力制約）の理解が十分でない可能性が示唆された．

4.3.4 方式 7（多数決方式）による誤答分析と考察

多数決方式について，誤答の傾向を整理する（括弧内は件数）．

(1) 全員が誤答：5 体すべてが誤答であった（7 件）．

(2) 回答が分散：各選択肢ラベルに票が分散し，正答が過半数に至らなかった（2 件）．ここで「分散」とは，票が拮抗し再投票を要したが，最終的に正答へ収束しなかった場合を指す．

(3) 少数は正答だが多数が誤答：一部は正答を選ぶが，誤答側が多数となった（3 件）．

(4) 二択に集中するが誤答が多数：2 候補に票が集中するが，誤答側が多数となった（2 件）．

(5) その他（2 件）．

多数決方式が誤答した 16 問について，図 5 のエージェントごとの正誤を確認すると，8 方式中 5 方式以上が正答している問題が 4 問存在した．この 4 問のうち 3 問では，少数の投票者が正答を選択しているにもかかわらず，多数決により誤答が採用されていた．したがって，投票者の役割設計や，票の重み付け（信頼度に基づく加重投票）などを導入することで，改善の余地がある可能性がある．一方で，全体として複数方式が誤答する問題も一定数存在することから，エージェントの基礎性能や RAG の検索性能による影響が大きいと考える．

4.3.5 マルチエージェントのみが正答を導けた問題の考察

図 5 より，単体エージェント群の全エージェントが誤答である一方で，マルチエージェント群では正答を導けた問題を 5 問確認した．特に図 6 に示す正答が c の問 26 では，単体エージェント群が主として a を選択する（または入力解釈に失敗する）一方で，マルチエージェント群は推論の早い段階で a を不適切として除外できていた．この差は，複数エージェント化により，候補の提示と反証，検証が役割として分離され，誤り候補を積極的に棄却する方向に探索が誘導された可能性がある．

question:
次のうち、機構による立入検査等の実施について正しいものを教えてください。
choices:
-a 特別区の区長は、機構の職員に、地域連携薬局等に立ち入り、その構造設備を検査させることができる。
-b 保健所を設置する市の市長は、機構の職員に、経理の状況に関し、登録認証機関の事務所に立入検査をさせることができる。
-c 厚生労働大臣は、機構の職員に、基準適合性認証の業務に関し、登録認証機関の事務所に立入検査をさせることができる。
-d 厚生労働大臣は、犯罪捜査のために必要であれば、機構の職員に、基準適合性認証の業務の状況に関し、登録認証機関の事務所に立入検査をさせることができる。

図 6 マルチエージェント群のみが正解した問題（抜粋）

5. 結論

本研究では，デジタル庁の法令 4 択問題データセットを対象に，軽量 LLM（Qwen3:8b）を用いて，単体エージェントのプロンプティング（ベースライン，法的三段論法，消去法）と，マルチエージェント推論（ディベート，分業，多数決）を比較する形で，複数エージェントによる推論の効果を評価した．40 問での実験の結果，単体エージェントに推論手順を付与するだけでは，正解率の改善は確認できなかった一方で，マルチエージェント方式はいずれも単体方式を上回り，特に多数決方式は 60.0%で最良となった．以上より，法令 4 択問題においては，単体のエージェントへの推論手順の付与よりも，マルチエージェント方式を用いた相互検証や集約が有効である可能性が示唆された．

誤答分析では，ディベート方式の誤答の多くが，「両ディベーターが正答に到達できない」ケースに起因していた．また，分業方式では，中間生成の失敗や役割逸脱が性能を制限する要因となることが確認された．これらは，マルチエージェント化のみでは解けない問題が残ること，ならびにエージェント間の受け渡し形式や制約設計が重要であることを示している．

今後は，評価規模の拡張に加え，合意度や一致度に基づく条件付き実行，分業方式における中間生成の検証と制約設計の改善について検討する．

参考文献

[1] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu, “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,” arXiv:2311.05232, 2023.

[2] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras, “LexGLUE: A Benchmark Dataset for Legal Language Understanding in English,” arXiv:2110.00976, 2021.

[3] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, et al., “LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models,” arXiv:2308.11462, 2023.

[4] デジタル庁，“日本の法令に関する多肢選択式 QA データセット（lawqa_jp），” GitHub repository, 2025. https://github.com/digital-go-jp/lawqa_jp

[5] 植松 幸生，大杉 直也，“複数の LLM を用いた法令 QA タスクの Ground Truth Curation,” 言語処理学会第 31 回年次大会（NLP2025）, 2025.

[6] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and Sebastian Riedel, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” NeurIPS, 2020. (arXiv:2005.11401)

[7] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih, “Dense Passage Retrieval for Open-Domain Question Answering,” EMNLP, 2020. (arXiv:2004.04906)

[8] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” arXiv:2201.11903, 2022.

[9] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou, “Self-Consistency Improves Chain of Thought Reasoning in Language Models,” arXiv:2203.11171, 2022.

[10] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al., “AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation,” arXiv:2308.08155, 2023.

[11] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem, “CAMEL: Communicative Agents for “Mind” Exploration of Large Language Model Society,” arXiv:2303.17760, 2023.

[12] Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li, “Debate or Vote: Which Yields Better Decisions in Multi-Agent Large Language Models?,” arXiv:2508.17536, 2025.

[13] 永島 賢也，“法的三段論法に関する一考察—その「理論」と，世界と世の中—,” Web 記事, 2022.

[14] 益川 弘如，白水 始，根本 紘志，一柳 智紀，北澤 武，河美保，“思考発話法を用いた多肢選択式問題の解決プロセスの解明—大学入試センター試験問題の国語既出問題を活用して—,” 2018.

[15] Luyu Gao, Xueguang Ma, Jimmy Lin, Jamie Callan, ”Precise Zero-Shot Dense Retrieval without Relevance Labels”, arXiv:2212.10496, 2022.

[16] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, Nan Duan, ”Query Rewriting for Retrieval-Augmented Large Language Models”, arXiv:2305.14283, 2023.