# マルチエージェント推論に基づく法的事例生成パイプラインの提案と評価

上野 優太[1,a)]　名越 洸翔[1,b)]　内山 光彩[1,c)]　高橋 哲朗[1,d)]　小野 智司[1,e)]

**概要**：条文の適用場面を学習するには，条文の要件に整合した事例問題を多数解くことが有効である．しかし，教材の供給には限りがあり，学習者の理解度や論点に応じて事例を用意することは困難である．近年，LLM による自動生成が期待される一方，法律分野では事実関係と条文要件の不一致（法的正確性）や，結論への推論の飛躍（論理的整合性）といった不整合が生じやすい．民法条文を入力として，正解肢が成立する事例および成立しない事例を生成し，その整合性を外部評価（LLM-as-a-Judge）で定量化する枠組みを提案した．具体的には，(i) 解析，生成，検証，修正を分担するマルチエージェント構成，(ii) 正解肢の根拠となる要件を抽出して生成を制御する要件抽出，(iii) 法令適用の推論過程を構造化する Chain-of-Thought（CoT），および (iv) Web 検索を用いた検索拡張生成（RAG）を組み合わせ，比較評価を行った．実験の結果，要件抽出による構造化は論理的整合性に，Web 検索による外部知識の導入は法的正確性に，それぞれ一定の寄与を持つことが示唆された．また，要件抽出と CoT を統合した手法は，評価スコアの標準偏差が低減する傾向が確認され，生成品質の安定化に寄与する可能性が示された．

# Design and Evaluation of a Multi-Agent-Based Pipeline for Legal Case Generation

***Abstract:*** Learners can effectively master the application of statutory provisions by solving many case-based problems consistent with legal requirements. However, the supply of teaching materials is insufficient, and preparing cases on demand and tailored to a learner's level is challenging. While automatic generation by Large Language Models (LLMs) is promising, issues such as mismatches between the case facts and statutory requirements (i.e., legal accuracy) and unjustified leaps in reasoning toward conclusions (i.e., logical consistency) often arise in the legal domain. This study investigates a multi-agent framework that takes Civil Code articles as input, generates case scenarios in which the articles are applicable or non-applicable, and analyzes their quality both quantitatively and qualitatively. Specifically, we combine and evaluate (i) a multi-agent architecture that decomposes the workflow into parsing, generation, verification, and refinement, (ii) Requirement Extraction (RE) to control generation based on key legal elements, (iii) Chain-of-Thought (CoT) prompting to structure the reasoning process, and (iv) Retrieval-Augmented Generation (RAG) using Web search to reference external knowledge. Experimental results suggested that structuring via RE may improve logical consistency, while introducing external knowledge via Web search may improve legal accuracy. Furthermore, the approach that integrated RE and CoT tended to reduce the variance of generation quality.

## 1. はじめに

### 1.1 研究の背景と課題

法学教育や実務訓練において，抽象的な法規範を具体的事実に適用する能力の習得は不可欠である．この能力を養うには，条文の要件や判例の論点を含む事例問題を多数解くことが有効である [1]．しかし，既存の教材や過去問の数は有限であり，学習者の理解度や特定の苦手分野に合わせ

[1] 鹿児島大学
Korimoto, Kagoshima, Kagoshima 8900065, Japan
[a)] k4720741@kadai.jp
[b)] k5451187@kadai.jp
[c)] k7719654@kadai.jp
[d)] takahashi@ibe.kagoshima-u.ac.jp
[e)] ono@ibe.kagoshima-u.ac.jp

て，法的整合性の取れた事例問題を人手で作成するには，多大な専門的知識と労力を要する．そのため，個別最適化された教材の提供は困難である [2,3]．以上より，法的整合性の取れた事例生成には，条文要件の充足を制御しつつ，事実から結論への推論を一貫させる仕組みが必要である．

### 1.2 LLM の可能性と現状の問題点

近年, 大規模言語モデル（Large Language Models; LLM）の発展により，多様なテキスト生成が可能となり，教育分野においても問題の自動生成への応用が期待されている．しかし，厳密性が求められる法律分野では，自然言語として流暢であるだけでは不十分であり，法的論理の整合性が求められる．汎用的な LLM に事例生成を行わせた場合，事実関係と適用条文の要件が不整合となる．あるいは結論に至る推論に矛盾が生じるなど，ハルシネーションが発生し得る [4]．とりわけ，特定条文の構成要件を満たす事実，あるいは意図的に満たさない事実を生成させる制御は，LLM 単体では依然として困難である．

### 1.3 本研究の目的とアプローチ

そこで本研究では，法令名，条文番号，条文テキスト，および当該条文において問題となる法的論点を入力として，条文が適用される事例および適用されない事例を生成し，その法的整合性を外部評価により定量化する枠組みを提案する．ここで外部評価とは，別の LLM に妥当性判定を行わせる LLM-as-a-Judge（外部判定）を指す．単一プロンプトで一括生成するのではなく，法的推論プロセスを模倣したマルチエージェント方式を採用する．具体的には，問題文とコンテキストから正解肢の根拠となる要件を抽出する工程と，Chain-of-Thought（CoT）により適用過程を段階的に明示させる工程を組み合わせ，さらに，生成・検証・修正の反復による品質改善 [5–7] と，Web 検索に基づく検索拡張生成（RAG）による外部知識参照 [8] を統合して，法的整合性の高い事例生成を試みる [5,6]．

### 1.4 本論文の構成

本稿では，提案枠組みの構築手法について述べるとともに，民法等を題材に生成された事例を用いて，法的妥当性と教育用教材としての有用性を検証する．また，要件抽出および CoT の有無が生成品質に与える影響を比較し，法的生成タスクにおける LLM の課題と可能性について議論する．

## 2. 関連研究

### 2.1 マルチエージェント方式

大規模言語モデルを単体で用いるのではなく，複数の役割に分解して協調させるマルチエージェント構成は，複雑な生成タスクの品質を安定化させる方法として提案されている [9,10]．また，生成結果を評価し，フィードバックに基づいて修正する反復設計は，自己フィードバックによる改善や失敗の言語化に基づく改善として整理されている [5,6]．本研究は，この系譜に沿って，法的選択問題に対する事例生成と整合性検証を役割分担し，反復により品質を調整する枠組みを採用する．

### 2.2 法的推論タスクへの LLM の適用事例

法律分野では，事実と結論の整合性に加え，規範と事実の対応付けが明示されることが求められる．IRAC（Issue, Rule, Application, Conclusion）等の枠組みにより推論過程を整える試みが報告されている [1] 一方で，ハルシネーションにより誤った前提知識に基づく推論が生じ得ることも指摘されている [4]．また，生成物を評価する方法として，LLM-as-a-Judge が体系化されつつある [11]．日本語法務タスクに関しては，日本の法令に関する多肢選択式 QA データセット lawqa_jp が公開され [2]，複数 LLM を用いた Ground Truth 整備も報告されている [3]．評価基盤としては，法情報抽出や含意判定を含む COLIEE が継続的に実施されている [12]．

### 2.3 反復的事例生成によるデータ拡張

教育や評価のためにデータが不足する状況では，LLM により指示，入力，出力の組を合成してデータを拡張する枠組みが提案されている [13]．ただし法律分野の事例生成では，論点に対応する事実を整合的に配置し，結論へ至る推論も破綻させない必要があるため，一回の生成では不整合が混入しやすい．推論と行動を往復する設計 [7] や，生成，検証，修正を反復する設計 [5,6] は，この制約下で品質を担保する上で有効な構成要素である．

### 2.4 条文要件抽出と推論過程の明示化（CoT）

要件抽出は，規範を満たすために必要な要素を明示し，生成および検証の観点を揃えるための操作である．要件事実論に基づく知識の構造化は，条文を原則と例外の階層として表現する立場として実装されており [14]，本研究の要件抽出は，この考え方を生成制御へ接続する．また CoT は，推論過程を段階的に記述させることで，結論への飛躍を抑制し得る手法として広く用いられる [15]．

## 3. 提案手法

### 3.1 パイプライン概要

本研究では，特定の法令条文（法令名，条番号，条文テキスト）と生成すべき事例の類型（適用事例／非適用事例）を入力とし，法的整合性の取れた事例を生成するマルチエージェント方式を提案する．単一プロンプトで一括生成するのではなく，(1) 条文構造の解析，(2) 事例生成，(3) 法的整合性の検証，の順に処理を進める．法的整合性

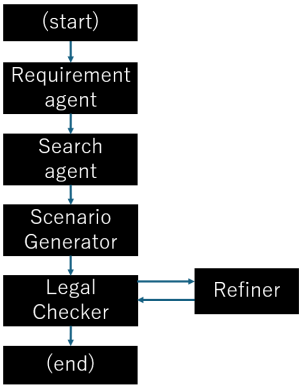


**図 1** パイプライン概要

の検証において，法的誤りや論理矛盾が指摘された場合は，指摘事項を修正指示として事例を再生成する反復により，品質を段階的に向上させる．実装では，これらの処理を LangGraph 上の状態遷移として定義し，自律的な修正ループを実現する．

### 3.2 エージェント構成

提案パイプラインは以下の 5 役から構成される．

(1) Requirement Extractor Agent ：条文テキストを解析し，法的効果を発生させるための要件と，その効果を妨げる例外を構造化データとして抽出する．

(2) Search Agent ：入力された条文テキストをクエリとして Web 検索を行い，関連する判例，法解釈，具体的な適用事例などの外部知識を取得する．

(3) Scenario Generator Agent ：要件および外部知識に基づき，指定された類型（適用／非適用）に合致する具体的事実関係（登場人物，状況，トラブル等）を含む事例案を生成する．

(4) Legal Checker Agent ：生成された事例を法的正確性および論理的整合性の観点から検証し，不合格の場合は具体的な修正指示を出力する．

(5) Refiner Agent ：修正指示に基づき，事例の再生成を行う．

### 3.3 Web 検索（RAG）の導入

本手法では, LLM の内部知識のみに依存することによるハルシネーションや知識の欠落を防ぐため，RAG（Retrieval-Augmented Generation: RAG）を統合する [8].

具体的には，Search Agent が，入力された条文情報に基づき，「{ 法令名 } { 条文番号 } { 要件キーワード } 判例 要件 解説」という形式の固定クエリを生成し，Web 検索を実行する．検索結果から得られた上位 3 件の Web ページについて，タイトル，本文，URL を連結し，最大 3,000 文字までコンテキストとしてプロンプトに動的に挿入する．

これにより，実際の判例や実務上の解釈指針を反映した，より具体的かつ法的正確性の高い事例生成を可能にする

### 3.4 要件抽出

要件抽出は Requirement Extractor Agent が行う．入力は法令名，条番号，および条文テキストである．本エージェントは条文を解析し，条文適用の要件と，条文適用を阻害または排除する例外を構造化して出力する．具体的には以下を出力する．

- 条文適用の要件：条文が適用されるために満たすべき条件のリスト（例：R1，R2，...）．
- 条文適用を阻害または排除する例外：但書や他項等で規定される適用除外条件のリスト（例：E1，E2，...）．

後段の事例生成では，事例類型（適用／非適用）に応じてどの要件を満たすか，どの要件を欠落させるか，および，どの例外に該当させるかを明示的に制御し，法的整合性の高い状況設定を目指す．

### 3.5 事例生成

事例生成は Scenario Generator Agent が行う．入力は条文テキスト，事例類型，要件構造，および検索された外部知識である．本エージェントは，(1) 生成方針の決定，(2) 事例本文の生成，の 2 段階で出力を構成する．まず，要件構造と外部知識を参照し，条文が適用される（または，されない）具体的な状況設定を立案する．特に検索結果に含まれる判例や具体例を参照することで，「典型的な紛争類型や権利関係」を取り入れた現実性の高い事実関係を構築する．次に，学習教材としての可読性を高めるため，以下の制約を設けて事例本文を出力する．

- 構造化された記述：状況，争点，当事者の主張，問題提起，結論，解説を区分して記述する．
- 事実の具体化：抽象的な法律用語を避け，「A が B に鍵を渡した」のように具体的な行動レベルで記述する．

### 3.6 段階的推論手順の導入

段階的推論手順は，事例生成において法令適用の過程を段階構造として記述させるための指示として組み込む [15]. 具体的には，背景整理 → 要件との対応付け → 判断，の流れを固定し，最終段落で判断根拠と要件充足（または不充足）を簡潔に整理させることで，結論への飛躍を抑制することを狙う．

### 3.7 反復条件と終了条件

検証および修正ループは，Legal Checker Agent が担う．

表 1　Legal Checker Agent による評価観点

| 事例タイプ | 評価観点 |
|---|---|
| 全タイプ共通 | 1. シナリオは具体的で現実的か<br>2. 法的分析は論理的に正しいか<br>3. 条文の引用や解釈に誤りはないか |
| 適用事例 | 1. シナリオは条文の要件を全て満たしているか<br>2. 法的分析は条文の適用を正確に説明しているか<br>3. 適用が明確であることが示されているか |
| 非適用事例 | 1. シナリオは条文の要件の少なくとも 1 つを欠いているか<br>2. 法的分析は非適用の理由を正確に説明しているか<br>3. なぜ適用されないかが明確に示されているか |

本エージェントは，生成された事例に対し，以下の 2 段階の判定を行う．

(1) 形式的要件：所定の構造（状況や結論等の見出し）および文字数（例：500–1,000 字）を満たしているか．

(2) 内容的要件：条文の適用要件や法解釈に誤りがないかを LLM が検証し，0.0–1.0 のスコア（検証スコア）で評価する．

本研究では，検証スコアが閾値（0.70）以上であることを合格条件とする．条件を満たさない場合，Legal Checker Agent は具体的な修正指示（法的誤りの指摘や不足要件の提示等）を出力し，これを入力として事例の再生成を行う．この反復プロセスは，合格判定が出るか，最大反復回数（デフォルト 6 回）に達するまで継続される．なお，検証スコアの算出に当たっては，事例のタイプに応じた評価観点をプロンプトに明示し，LLM に対して厳格な評価を指示している．具体的にはシナリオの具体性，法的正確性，条文解釈の正確性という共通の評価項目に加え，表 1 のように，事例タイプごとに適合度を確認させている．これら計 6 つの観点を総合的に判断し，0.0 から 1.0 の範囲でスコアを算出させる仕組みである．

## 4. 実験設定

本節では，提案パイプラインの有効性を検証するための実験設定，使用モデル，および比較手法について述べる．

### 4.1 実験データ

本実験では，特定の法令データセットを用いるのではなく，日本の民法から抽出した 40 種類の主要な条文を評価対象として採用した．具体的には，以下の条文を使用した．

- 総則分野: 第 3 条の 2（意思能力），第 5 条（未成年者の法律行為），第 9 条（成年被後見人の法律行為），第 93 条（心裡留保），第 94 条（通謀虚偽表示），第 95 条（錯誤），第 96 条（詐欺又は強迫），第 99 条（代理行為の要件及び効果），第 109 条（代理権授与の表示による表見代理），第 110 条（権限外の行為の表見代理），第 117 条（無権代理人の責任），第 162 条（所有権の取得時効），第 166 条（債権等の消滅時効）

- 物権分野: 第 177 条（不動産に関する物権の変動の対抗要件），第 192 条（即時取得），第 206 条（所有権の内容），第 233 条（竹木の枝の切除及び根の切取り），第 249 条（共有物の使用），第 295 条（留置権の内容），第 304 条（先取特権の物上代位），第 342 条（質権の内容），第 369 条（抵当権の内容）

- 債権分野: 第 412 条（履行期と履行遅滞），第 415 条（債務不履行による損害賠償），第 423 条（債権者代位権），第 424 条（詐害行為取消権），第 466 条（債権の譲渡性），第 474 条（第三者の弁済），第 533 条（同時履行の抗弁），第 536 条（債務者の危険負担等），第 541 条（催告による解除），第 555 条（売買），第 562 条（買主の追完請求権），第 587 条（消費貸借），第 601 条（賃貸借），第 632 条（請負），第 644 条（受任者の注意義務），第 709 条（不法行為による損害賠償），第 715 条（使用者責任），第 717 条（土地工作物等の占有者及び所有者の責任）

これらの条文を選定した理由は，主に以下の 3 点である．

(1) 法学教育における基礎的重要性：
法学教育において基本的かつ重要とされる論点（意思表示の瑕疵，不法行為など）を含んでおり，教育用事例生成システムの評価に適していると考えられること．

(2) 分野の網羅性：
総則，物権，債権の各編から条文を選定することで，特定の分野に偏ることなく，一定の法的文脈の多様性を確保しようと試みたこと．

(3) 要件構造の多様性:
主観的要件（善意，悪意など）や形式的要件（登記など），異なる論理構造を持つ条文を含めることで，様々なパターンの検証を意図したこと．

各設問に対し，正解肢が成立する事例と成立しない事例を 1 件ずつ生成し，合計 80 事例を評価対象とした．

### 4.2 使用モデル

#### 4.2.1 事例生成モデル (Generator)

事例生成モデルには，Qwen3-32B-Instruct を採用した．推論は Ollama を用いてローカル環境で実行し，temperature は 0.7 とした．

#### 4.2.2 評価モデル (Evaluator)

生成事例の評価には，実験時点で利用可能であった GPT-5.2 を用いた．評価は LLM-as-a-Judge（外部 LLM による採点）の枠組みに従い [11]，各事例に対して法的正確性と論理的整合性を定量化した．

### 4.3 比較条件

要件抽出および段階的推論手順の寄与を明らかにするため，以下の 4 条件で比較した．

- ベースライン：問題文と選択肢から直接事例を生成し，検証と修正を反復する．ただし，要件抽出結果を明示的に入力へ付与せず，法令適用の推論過程の段階化も指示しない．
- 要件抽出統合：ベースラインに対して要件抽出結果を入力として付与し，事例が論点や要件から逸脱しにくいよう制御する．
- CoT 統合：要件抽出統合に加え，法令適用の推論過程を段階構造として明示させる指示を付与し，結論への飛躍を抑制する．
- RAG 統合：CoT 統合に加え，Web 検索により関連する判例や法解釈等の外部知識を取得し，入力としてコンテキストに付与する．LLM の内部知識のみに依存せず，実際の判例や事実に即した具体性と法的正確性の高い事例生成を促す．

### 4.4 評価指標

Evaluator である GPT-5.2 に対し，生成された事例を入力し，法的正確性と論理的整合性の 2 観点から 5 段階評価（1: 不適切 – 5: 適切）をさせその理由を記述させた．

法的正確性は，想定される要件が事例内の事実により支持されているか，または意図した不充足が整合的に設計されているかを問う．論理的整合性は，事実から結論への推論に飛躍や矛盾がないかを問う．

なお，LLM の生成結果には確率的な揺らぎが含まれるため，各手法について同一のデータセット（80 事例）を用いて 3 回の試行を行った．本実験の結果として報告するスコアは，これら 3 回の試行における全事例の評価値の平均である．

### 4.5 LLM-as-a-judge の妥当性検証

本実験では評価者として GPT-5.2 を採用するが，その自動評価が法的な厳密性を有しているかを確認するため，予備実験として実際の事例に対する評価を行った．対象は，民法の主要論点（即時取得，債権者代位権など）を含む教科書レベルの正解事例 20 件である．

評価結果を表 3 に示す．正解事例に対するスコアは，Fact Score で 3.72，Logic Score で 3.50 であり，満点（5.0）には至らなかった．この結果から，本評価基準ではスコアの絶対値のみをもって手法の優劣を判断することは必ずしも適切でない可能性が示唆される．そこで本研究では，ベースラインに対する向上の程度，および正解事例の水準にどの程度近いかという，手法間の相対的な傾向に着目して評価を行う．

**表 2** 比較条件

| 手法 | 要件抽出 | 段階的推論 | Web 検索 |
|---|---|---|---|
| ベースライン | – | – | – |
| 要件抽出統合 | ✓ | – | – |
| CoT 統合 | ✓ | ✓ | – |
| RAG 統合 | ✓ | ✓ | ✓ |

**表 3** 実際の事例に対する評価スコア

| 対象 | Fact Score | Logic Score |
|---|---|---|
| 実際の事例 | 3.73 (SD: 0.715) | 3.33 (SD: 1.05) |

**表 4** 各手法の平均評価スコアおよび標準偏差 (SD)

| | Fact Score | | Logic Score | |
|---|---|---|---|---|
| 手法 | Mean | SD | Mean | SD |
| ベースライン | 3.20 | 0.744 | 2.68 | 1.004 |
| 要件抽出統合 | 3.24 | 0.864 | **2.78** | 1.161 |
| CoT 統合 | 3.35 | 0.700 | 2.56 | 0.923 |
| RAG 統合 | **3.41** | 0.769 | 2.71 | 0.942 |

## 5. 実験結果

本節では，40 問の設問に対して各手法を適用した結果を示す．で適用／非適用の 2 類型を生成した計 80 事例（40 設問 ×2）を対象に，4 手法により生成された合計 320 事例について GPT-5.2 を用いて評価を行った評価結果を示す．

### 5.1 定量的評価

表 4 に，各手法における法的正確性および論理的整合性の平均スコアおよび標準偏差（SD）を示す．なお，前節（表 3）で示した通り，教科書レベルの正解事例における平均スコアは Fact 3.73 / Logic 3.33 であった．本節では，この基準値との比較も踏まえて各手法の傾向を述べる．

まず Fact Score においては，Web 検索により外部知識を取り込む RAG 統合が全条件中で最も高い平均値（3.41）を示した．

一方，Logic Score においては，最も簡素な要件抽出統合が最高値（2.78）を記録した．特に，推論過程の明示化を意図した CoT 統合は，Logic においてベースラインを下回る結果（2.56）となった．このことは，非適用事例の説明において詳細な推論連鎖を要求すると，論理構成が複雑化し，かえって整合性を損ねる場合があることを示唆する．ただし，本結果のみから因果を断定することはできない．

次に出力のばらつきに注目すると，CoT 統合に特徴が見

**表 5** 手法別の指摘エラー件数集計

| | 法的正確性 | | | 論理整合性 | | |
|---|---|---|---|---|---|---|
| 手法 | L1 | L2 | L3 | C1 | C2 | C3 |
| ベースライン | 3 | 51 | 33 | 13 | 50 | 20 |
| 要件抽出 | 1 | 45 | 50 | 7 | 45 | 28 |
| CoT 統合 | 3 | 52 | 39 | 13 | 45 | 26 |
| RAG 統合 | 1 | 48 | 31 | 12 | 44 | 29 |

られた．表 4 の標準偏差（SD）に着目すると，CoT 統合は Fact（0.700）・Logic（0.923）の双方で，他手法と比べてばらつきが小さい傾向にあった．分散の差を F 検定で確認したところ，Logic では要件抽出統合（SD 1.16）との間に 5%水準で有意差（$p = 0.0417$）が認められ，Fact でも 10%水準で有意傾向（$p = 0.0618$）が確認された．もっとも，本評価は LLM による自動採点に基づくため，この結果のみをもって安定性の優劣を結論付けることは避けるべきである．一つの解釈として，CoT 統合は極端に低い品質の出力を抑制し，出力のばらつきを低減する方向に働いた可能性が示唆される．

### 5.2 生成事例による定性評価

定量評価の結果を補完するため，実際に出力された事例を用いた定性的な比較を行う．ここでは，民法第 5 条（未成年者の法律行為）を対象とした生成結果を例に，各手法の出力特性と，LLM による自動評価（Fact/Logic スコア）の妥当性について議論する．

#### 5.2.1 ベースライン手法の出力

以下にベースライン手法で出力した事例を示す．

【状況】15 歳の A（未成年者）が、スマートフォンを購入するために販売店 Y と契約を結んだ。A の法定代理人である父 X はこの契約について事前に同意していない。【トラブル】父 X は、A の契約行為が無効であるとして、販売店 Y に対して契約解除を求める。【Y の主張】販売店 Y は、契約は未成年者 A が自ら行ったものであり、父 X の同意がなくても有効であると主張する。【解説】A は未成年者であり、スマートフォンの購入（義務を負う法律行為）には法定代理人 X の同意が必要です（民法第 5 条 1 項）。同意が得られなかったため、この契約は父 X が取り消すことができます（民法第 5 条 2 項）。ただし、A が「権利を得るだけ」（例：贈与）の場合は同意不要です。

#### 5.2.2 RAG 統合手法による出力

以下に RAG 統合手法で出力した事例を示す．

【状況】15 歳の高校生・A は、スマートフォン販売店 B と契約し、新品のスマートフォンを 50 万円で購入した。A の法定代理人である父 C は、この契約を知らず、A は C に相談せずに契約を進めた。C は後日、契約を知り、B に契約取消を求めた。【トラブル】B は「未成年者が高額商品を購入した場合でも、単に権利を得る行為は例外」と主張し、契約の無効を拒否した。C は「A が義務を負う契約をしたため、第 5 条第 1 項に違反し、第 2 項により取消可能」と反論した。【Y の主張】B の主張：「A はスマートフォンの所有権を得る行為であり、『単に権利を得る』例外に該当する。」【解説】1. **要件満たし**（R1）：A（15 歳）は法定代理人 C の同意を得ずに B と契約を締結した（R1）。2. **例外の不適用**: 契約は「単に権利を得る」行為に限らず、A が 50 万円の支払い義務を負う双務契約であるため、第 1 項の適用を免れる例外（「単に権利を得、義務を免れる」）は成立しない。3. **法的効果**: 第 2 項により、C は契約の取消を請求できる。4. **三段論法**: - 大前提: 未成年者が義務を伴う法律行為をした場合、第 5 条第 1 項に違反する。- 小前提: A は義務を伴う契約をした（R1）。- 結論: A の契約は第 2 項により取消可能である。

#### 5.2.3 評価スコアと定性分析

本事例（民法第 5 条）における各手法の評価スコア（Fact/Logic）および，LLM-as-a-Judge による主な評価理由を以下に示す．

**ベースライン手法**

Score：Fact 3.67 / Logic 3.67
評価理由：

法的用語の不正確さが主な減点要因となった．具体的には，未成年者取消権の行使による遡及的な無効を説明すべきところを，無効や解除といった用語と混同しており，法定代理人による追認や取消権の法的性質に関する整理が不十分であると指摘された．

**RAG 統合手法**

Score：Fact 4.67 / Logic 4.00
評価理由：

事実関係の構成（年齢，同意の欠如，対価支払義務の発生）については，条文要件を完全に満たしており高く評価された．一方で，解説文において第 5 条 1 項に違反や契約は無効といった表現が散見され，取消しうる行為と当初から無効な行為の法的区別が厳密にはなされていない点が，Logic スコアの減点要因となった．

この結果から，RAG による知識拡張が条文要件の充足に寄与した可能性が示唆される．

### 5.3 評価コメントに基づく誤答要因の分析

各手法の改善要因を詳細に分析するため，GPT-5.2 が出力した評価コメントを分析し，指摘された誤りの内容を以下の 6 類型に分類，集計した（表 5）．

- 法的正確性: L1 (条文実在性)，L2 (法解釈，射程)，L3 (結論の妥当性)
- 論理整合性 (Consistency): C1 (自己矛盾)，C2 (構成要素の欠落)，C3 (論理の飛躍)

集計結果から，以下の傾向が確認された．第一に，要件抽出による自己矛盾の抑制である．自己矛盾（C1）の件数に着目すると，要件抽出のみを行う手法において，件数が比較的少なく（7 件）留まる傾向が見られた．これは，要件リストというガイドラインに従うことで，記述内容の一貫性が一定程度保たれたためと考えられる．ただし，CoT や RAG 統合など記述量が増加する手法では，ベースラインと同程度（12–13 件）の結果となっており，手法の複雑化に伴い，整合性の維持が再び困難になる傾向が確認された．

第二に，推論の詳細化に伴う論証の不足である．論理の飛躍（C3）に関しては，ベースライン（20 件）が最も少なく，提案手法群（26–29 件）の方が件数が多い．これは，ベースラインが簡潔な記述に留まるのに対し，提案手法は詳細な法的推論を展開しようと試みるため，その過程でより精緻な説明が求められ，結果として説明不足や飛躍と判定される箇所が増えたためと推察される．また，法解釈の誤り（L2）も依然として高い水準（45–52 件）にあり，生成モデルに対する法解釈の厳密な制御は，RAG を用いてもなお困難な課題であることが示唆される．

## 6. 考察

誤答分析（表 5）の結果は，論理と知識の不可分性を示唆している．知識を与えずに論理構造のみを強化した要件抽出統合において，結論の誤り（L3）がベースラインよりも悪化した傾向が見られた．これは，LLM が与えられた論理枠組みを埋めるために，欠落した知識をハルシネーションによって補完しようとした可能性がある．この点は，十分な専門知識を伴わない状態で，論理整合性のみを追求することの難しさを，示唆しているとも解釈できる．

一方で，RAG 統合において L3 の件数がベースラインと同等水準に留まったことは，Web 検索によって得られた外部知識が，推論過程における事実誤認を一定程度抑制する役割を果たした可能性を示唆している．すなわち，法学教育のような専門性の高いドメインにおいては，プロンプトによる推論誘導と，RAG による知識参照が，互いの欠点を補う関係として機能することが望ましいと考えられる．

なお，本実験では，CoT 統合による詳細な推論指示が，一部の事例において，かえって平均スコアを低下させる現象 [15] が確認された．これは，使用した LLM（Qwen3-32B）にとって，複雑な法的推論と事例生成の処理が過大な負荷となった可能性を示唆している．要件抽出で見られた多面的な記述の両立は，より大規模なモデルの活用により改善が見込まれると考えられ，今後の課題である．

## 7. 結論

本研究では，法学教育における事例作成支援を目的として，条文の法的要件に基づき，正解肢が成立する事例および成立しない事例を生成するマルチエージェント型パイプラインを提案した．民法事例を対象とした評価実験の結果，CoT 統合による品質の安定化，ならびに要件抽出および RAG 統合による性能向上の傾向が確認された．今後，より推論能力の高いモデルを採用することで，CoT で見られた論理スコアの低下の改善を試みる．また，法律の専門家による人手評価自動評価で確認された傾向が，実務家の厳しい基準と合致するかを検証することが重要である．

### 参考文献

[1] Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Yue Zhuo, Patrick Charles Emerton, and Genevieve Grant. Can chatgpt perform reasoning using the IRAC method in analyzing legal scenarios like a lawyer?, 2023.

[2] Digital Agency (Japan). lawqa_jp: 日本の法令に関する多肢選択式 QA データセット, 2025. デジタル庁公開の 4 択法令 QA データセット（README 記載の説明に基づく）.

[3] 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. lawqa_jp README に記載の利用文献（書誌詳細は要確認）.

[4] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Li. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023. Accepted by ACM Transactions on Information Systems (TOIS).

[5] Aman Madaan, Niket Tandon, Prakhar Gupta, Alexander Hall, Shuyin Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Peter Clark, et al. Self-refine: Iterative refinement with self-feedback, 2023.

[6] Noah Shinn, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023.

[7] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2022. ICLR camera-ready (arXiv comment より).

[8] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[9] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for “mind” exploration of large

scale language model society, 2023.

[10] Qingyun Wu, Shuyin Zhu, Linyi Li, Wei Yang, Yixuan Zhang, Zhiyuan Li, Jie Zhou, Yujian Lin, Jian Li, and Weizhu Liu. Autogen: Enabling next-gen LLM applications via multi-agent conversation, 2023.

[11] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on LLM-as-a-judge, 2024.

[12] Mi-Young Kim, Juliano Rabelo, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. COLIEE 2022 summary: Methods for legal document retrieval and entailment. In *New Frontiers in Artificial Intelligence - JSAI-isAI 2022 Workshop, JURISIN 2022, and JSAI 2022 International Session, Kyoto, Japan, June 12-17, 2022, Revised Selected Papers*, Vol. 13859 of *Lecture Notes in Computer Science*, pp. 51–67. Springer, 2022.

[13] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions, 2022. ACL 2023 camera-ready (arXiv comment より).

[14] Ken Satoh, Kento Asai, Takamune Kogawa, Masahiro Kubota, Megumi Nakamura, Yoshiaki Nishigai, Kei Shirakawa, and Chiaki Takano. PROLEG: an implementation of the presupposed ultimate fact theory of japanese civil code by PROLOG technology. In *New Frontiers in Artificial Intelligence - JSAI-isAI 2010 Workshops, LENLS, JURISIN, AMBN, ISS, Tokyo, Japan, November 18-19, 2010, Revised Selected Papers*, Vol. 6797 of *Lecture Notes in Computer Science*, pp. 153–164. Springer, 2010.

[15] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2022.