大規模言語モデルを用いた End-to-End の法令あてはめとその出力を利用した検索拡張手法の検証

榊原 豪太 [1] 李 吉屹 [2] 吉岡 真治 [2]
[1] 北海道大学 工学部 情報エレクトロニクス学科 情報理工学コース
[2] 北海道大学大学院情報科学院/研究院
sakakibara.gota.y6@elms.hokudai.ac.jp
{jyli, yoshioka}@ist.hokudai.ac.jp

概要

本研究では日本の民法を対象に，実際の法律業務を想定した End-to-End(法律に関するクエリから正誤を出力) での法令のあてはめタスクにおける大規模言語モデル (Large Language Model; LLM) の挙動を分析するとともに，条文検索が幻覚 (hallucination) の抑制にどれほど寄与するかを検証した．

その結果，検証対象とした LLM は内部知識として条文知識を完全には有しておらず，多くのケースで幻覚を発生させることが分かった．また，条文検索を組み合わせることで幻覚の抑制に寄与することが確認された．

文脈情報を含むべく LLM の出力した根拠条文をもとに検索を行い，さらに準用を含む条文を検出し関連条文を補完する検索拡張手法 (Retrieval Augmented Generation; RAG)[1] を利用し，その検索精度および法的推論精度の評価を行った結果，一定の効果が得られた．

1 はじめに

昨今の**大規模言語モデル (Large Language Model; LLM)** の発展に伴い，特に専門度の高く専門家が不足しがちな法律分野においても LLM を活用しようという動きが広まっている [2]．一方で実際の法律業務で LLM を利用する上ではその出力に**幻覚 (hallucination)** が含まれる可能性があり，[3, 4] では人間の専門家と同程度の精度を担保することが難しいことも示されている．

そのような幻覚に対処する手法として，**検索拡張手法 (Retrieval Augmented Generation; RAG)**[1] が知られている．LLM が推論をする際に信頼のできる外部知識を参照することで，幻覚の抑制が期待できる．

図 1 準用の模式図

また我が国の民法を含む大陸法 (Civil Law) ではしばしば図 1 に示すような**準用 (mutatis mutandis)** が用いられる．これは他の規定で示されている内容を重複して記載することを避けるために用いられるものであり，準用元と準用先の両方を参照する必要があることから条文検索を行う上で特に注意が必要である．

本研究では日本の民法を対象として，実際の法律業務を想定した End-to-End(法律に関するクエリから正誤を出力) での法的推論タスクにおいての LLM の挙動を分析するとともに，その際に LLM が参照した条文をクエリとした条文検索と問題文をクエリとした条文検索を組み合わせ，さらに抽出した条文に対応する準用先 (元) 条文を含めた RAG を提案し，それが幻覚の抑制にどれほど寄与するかを検証した．

2 関連研究

2.1 COLIEE

法令のあてはめタスクに関する著名な先行研究として，**COLIEE(Competition on Legal Information Extraction/Entailment)**[5] が知られている．COLIEE は毎年開催される国際的な法情報抽出・推論コンペティションであり，2025 年に開催されたものは 4 つの Task と Pilot Task からなるが [6]，このうち本研究で扱うのは Task3 (Statute Retrieval) と Task4 (Legal Textual Entailment) である．

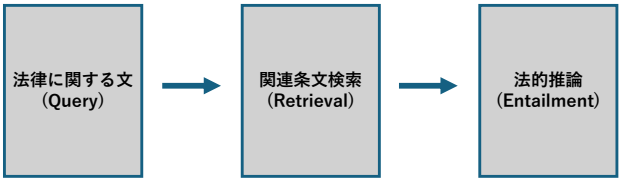

図 2　法令のあてはめタスクの模式図

法令のあてはめを行う過程を図 2 のように考えた場合，Task3 は関連条文検索 (Retrieval) に対応し，Task4 は法的推論 (Entailment) に対応するといえる．ただし COLIEE Task4 では，関連条文が過不足なくあらかじめ与えられる点で，本研究の End-to-End タスクとは異なる．

2.2　COLIEE Task3

COLIEE Task3 は，我が国の司法試験択一回答式問題の問題文に対して関連する条文を抽出するタスクである [5].

COLIEE 2025[6] では 8 チームから 22 の手法が提出され，大きく分けて以下の 2 つのアプローチが採用された．

- 検索方法を工夫した手法
  Okapi BM25[7] と Sentence Transformer Model を組み合わせた手法や，抽出から reranking までを段階的にそれぞれ LLM を用いて行い，最後に 3 つの LLM で多数決を行う手法などさまざまな検索方法の工夫が見られた．
- 民法の構造に着目した手法
  民法に見られる特有の入れ子構造や引用関係に着目し，それらを含んだ Graph Neural Network (GNN) を形成し，それをもとに関連条文を抽出する手法など，民法の構造に着目した手法が見られた．

2.3　COLIEE Task4

COLIEE Task4 は，司法試験問題文とそれに関連する条文が過不足なく与えられ，問題文に対する正誤判定を行うタスクである [5].

COLIEE 2025[6] では，10 チームから 29 の手法が提出され，大きく分けて以下の 2 つのアプローチが採用された．

- LLM を Fine Tuning した手法
  他のモデルを用いてデータ数を増大させたデータセットを用いて LLM を Fine Tuning することで精度向上を図った手法などが見られた．

表 1　実験で使用したモデル

	gpt-oss:120b[9]	**llama4:128x17b**[10]
Parameters	116.8B	401.6B
Quantization	MXFP4	Q4_K_M
Cutoff	2024.08	2024.06
Release Date	2025.08.05	2025.04.05

- プロンプトを工夫した手法
  Few-shot Prompting を用いた手法が主であり，与える例の正誤数を等しくすることでバイアスを抑制した手法などが見られた．

3　実験設定

3.1　データセット

本研究では COLIEE 2024[8] および COLIEE 2025[6] におけるテストデータセットを評価に用いるデータセットとして利用した．これらのデータセットは民法条文のうち，公式の英訳が提供されている 768 の条文に関連する司法試験の択一解答式問題から構成されている．COLIEE 2024 のデータセットは 2023 年の司法試験問題から抽出された 109 の問題 (以下，R05 とする) からなり，COLIEE 2025 のデータセットは 2024 年の司法試験問題から抽出された 73 の問題 (以下，R06 とする) からなる．データセットとしては問題文とそれに関連する条文とがペアで与えられている．なお，データセット内の問題は条文に関する知識のみで回答できるものに限定されている．

また情報検索におけるコーパスとしてはデジタル庁から公開されている **e-Gov 法令データセット**[1] を用いた．

3.2　モデル

実際の COLIEE のルールではモデルデータまたは訓練データセットが公開されているモデルのみの利用が認められており，Task3 および Task4 ではテストデータとなる司法試験が実施される以前に公開されたモデルのみ使用が認められている [5].

本研究では最新の LLM の能力の検証が目的であるため，前者にのみ従い表 1 に示す 2 モデルを利用した．なお，再現性の観点からモデルの Temperature はすべて 0 に設定した．

1)　e-Gov ポータル https://www.e-gov.go.jp

表 2　実験 1: End-to-End タスクの結果

Model	R05	R06
gpt-oss:120b	62.39% (68/109)	58.90% (43/73)
llama4:128x17b	66.06% (72/109)	53.42% (39/73)

3.3　評価手法

COLIEE Task3 および Task4 の評価指標 [5] に倣い，条文検索を評価する際には **Precision, Recall, F2-measure** を，法的推論を評価する際には **Accuracy** を用いた．

4　実験

4.1　実験 1: End-to-End タスク

実際の法律業務を想定したとき，COLIEE のタスク設定のように条文検索と法的推論が分離されていることは稀である．そのため本実験では法律に関するクエリから正誤を出力する End-to-End タスクを考えることで，LLM が条文に関する内部知識をどれほど有しており，法律業務でどれほど有効であるかを評価することを目的とする．

入力は司法試験問題文とし，出力は正誤を表す"Y"または"N"および根拠条文とした．評価指標としては Accuracy を用いた．

実験結果は表 2 に示す．正誤に合わせて表示させた問題条文を確認したところ，多くの条文で誤りがあり不適切な条文を用いていたことによる性能低下が確認された．

4.2　実験 2: BM25 条文検索併用

本実験では，LLM により正確な条文情報を与えることで回答精度が向上するかどうかを検証することを目的とする．

入力は司法試験問題文と Okapi BM25[7] で検索した上位 m 件とし，出力は正誤を表す"Y"または"N"とした．評価指標としては条文検索結果に対して Precision, Recall, F2-measure を，法的推論に対して Accuracy を用いた．また追加指標として，**Accuracy with Correct Retrieval** (条文検索において必要とする条文をすべて抽出できたとき，クエリに対して正答を導いた確率) も算出した．

実験結果は表 3 に示す．この結果から適切な条文情報を含む条文の情報を検索結果として与えることにより，LLM が誤った条文に基づく判断を減らす (幻覚を抑制する) ことに寄与することが確認された．またモデルの性質の違いもまた結果に現れている．gpt-oss では検索結果となる条文の数が増大し必要な条文の再現率が向上することによって，全体の性能が向上している．一方で llama4 では条文の再現率が向上しているにもかかわらず，関連しない条文が増大しているためか性能が低下している．システム設計の際には，このようなモデルの性質を考慮した再検討が必要であると考える．

4.3　実験 3: LLM 出力による条文検索併用

単語レベルのマッチングによる条文検索を行うことで適切な条文を提供できる可能性が示されたものの，質問中の具体的な情報をもとに条文へのあてはめが必要な質問に対しては単語レベルのマッチングでは適切な条文が得られないことが確認された．

このような問題に対応するためには意味的なマッチングを行う必要がある．この問題に対し LLM が生成した根拠条文を利用した追加検索の手法を提案する．具体的には実験 1 で行った LLM による根拠条文の出力を利用した条文の追加検索を行う．

LLM が提示する条文は記憶が不完全なために完全な条文を示すことはできなかったが，LLM のもつ意味的理解の機能により関連条文と単語レベルで共有している条文を生成していることが多いことが確認された．この分析結果に基づき，LLM が出力した根拠条文をクエリとした条文検索を行うことにより意味的に類似した条文を追加検索することが可能になると考えた．

入力は司法試験問題文と，実験 1 で LLM が出力した根拠条文をクエリとして Okapi BM25 で検索した上位 m 件および問題文をクエリとして Okapi BM25 で検索した上位 5 件とし，出力は正誤を表す"Y"または"N"およびその回答根拠となる条文とした．

評価指標としては条文検索に対して F2-measure, **BM25** (問題文をクエリとした検索結果が正解条文に含まれる割合), **LLMoutput** (LLM の出力をクエリとした検索結果が正解条文に含まれる割合), **Junyo** (準用元 (先) が正解条文に含まれる割合) を，法的推論に対して Accuracy と **Accuracy with Correct Retrieval** を用いた．

実験結果は表 4 に示す．実験 2 の m=5 の場合と比較すると抽出件数が増えたが，LLM の出力を利用した条文検索の併用により R5 年の gpt-oss を除き条文検索精度および法的推論精度の向上が見られた．

表 3 実験 2: BM25 条文検索併用の結果

Model	Data	m	F2 Score	Precision	Recall	Accuracy with Correct Retrieval	Accuracy
gpt-oss:120b	R05	5	0.4166	0.1688	0.6972	91.18% (62/68)	80.73%
		10	0.3026	0.0936	0.7523	92.00% (69/75)	82.57%
		15	0.2443	0.0679	0.7939	95.00% (76/80)	87.16%
	R06	5	0.4405	0.1945	0.7112	82.22% (37/45)	71.23%
		10	0.3346	0.1110	0.7913	84.62% (44/52)	79.45%
		15	0.2737	0.0795	0.8495	87.72% (50/57)	80.82%
llama4:128x17b	R05	5	0.4166	0.1688	0.6972	80.88% (55/68)	80.73%
		10	0.3026	0.0936	0.7523	81.33% (61/75)	78.90%
		15	0.2443	0.0679	0.7939	78.75% (63/80)	79.82%
	R06	5	0.4405	0.1945	0.7112	86.67% (39/45)	79.45%
		10	0.3346	0.1110	0.7913	82.69% (43/52)	76.71%
		15	0.2737	0.0795	0.8495	75.44% (43/57)	73.97%

表 4 実験 3: LLM 出力を利用した条文検索併用の結果

Model	Data	m	F2	BM25	LLMoutput	Junyo	Accuracy with Correct Retrieval	Accuracy
gpt-oss:120b	R05	1	0.4199	69.72%	1.83%	0.46%	88.57% (62/70)	79.82%
		2	0.4151	69.72%	3.21%	1.38%	90.28% (65/72)	80.73%
		3	0.3952	69.72%	4.13%	0.46%	88.89% (64/72)	79.82%
	R06	1	0.4386	71.12%	0.34%	0.46%	89.13% (41/46)	76.71%
		2	0.4200	71.12%	0.34%	0.00%	84.44% (38/45)	76.71%
		3	0.4147	71.12%	1.99%	0.00%	89.13% (41/46)	83.56%
llama4:128x17b	R05	1	0.4259	69.72%	1.83%	0.92%	81.69% (58/71)	79.82%
		2	0.4224	69.72%	3.52%	0.76%	79.45% (58/73)	80.73%
		3	0.3992	69.72%	4.43%	0.00%	81.94% (59/72)	81.65%
	R06	1	0.4497	71.12%	2.74%	0.46%	87.50% (42/48)	79.45%
		2	0.4436	71.12%	3.42%	0.00%	87.50% (42/48)	78.08%
		3	0.4231	71.12%	3.42%	0.00%	81.25% (39/48)	72.60%

5 おわりに

本研究では実際の法律業務を想定した End-to-End での法令あてはめタスクにおける LLM の挙動を分析するとともに，その際に LLM が参照した条文をクエリとした条文検索と問題文をクエリとした条文検索を組み合わせ，さらに抽出した条文に対応する準用先 (元) 条文を含めた RAG を提案し，それが幻覚の抑制にどれほど寄与するかを検証した．

その結果検証対象とした LLM において，その出力を利用することで条文検索および法的推論精度の向上が見られた．従来の文脈情報を含む手法に比べて簡易的な手法であるため，今後さらなる精度向上が期待できる．

参考文献

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 9459–9474. Curran Associates, Inc., 2020.

[2] Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. Can gpt-3 perform statutory reasoning? In **Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law**, ICAIL '23, p. 22–31, New York, NY, USA, 2023. Association for Computing Machinery.

[3] Jungmin Choi, Jungo Kasai, and Keisuke Sakaguchi. Evaluating GPT in Japanese Bar Examination. In **Proceedings of the Thirtieth Annual Meeting of the Association for Natural Language Processing**, pp. 2577–2582, Kobe, Japan, 2024. Association for Natural Language Processing.

[4] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. **Journal of Legal Analysis**, Vol. 16, No. 1, pp. 64–93, 06 2024.

[5] Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. Overview and discussion of the competition on legal information extraction/entailment (coliee) 2023. **The Review of Socionetwork Strategies**, Vol. 18, No. 1, pp. 27–47, 2024.

[6] Randy Goebel, Yoshinobu Kano, Calum Kawn, Mi-Young Kim, Juliano Rabelo, Ken Satoh, Hiroaki Yamada, and Masaharu Yoshioka. Proceedings of the workshop on the twelfth international competition on legal information extraction and entailment (coliee 2025), 2025.

[7] Stephen Robertson and Hugo Zaragoza. **The Probabilistic Relevance Framework: BM25 and Beyond**, Vol. 3. Now Publishers Inc., Hanover, MA, USA, 2009.

[8] 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.

[9] OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b Model Card, 2025.

[10] Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation, 2025.

A Okapi BM25

本研究で利用している Okapi BM25[7] を用いた reranking では以下の式に基づいてスコアリングをしている．

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\frac{|D|}{\text{avgdl}}\right)} \tag{1}$$

なお任意のパラメータである k_1 および b について，本研究ではそれぞれ $k_1 = 1.5$，$b = 0.75$ を用いた．

B プロンプト

各実験で用いたプロンプトの例を以下に示す．

B.1 実験 1: End-to-End タスク

System Prompt
あなたは日本の法律に詳しい法律専門家です．
出力は必ず JSON 形式のみで返し，他のテキストは一切含まないでください．

User Prompt
日本の司法試験の民法に関する択一回答式問題です．記述 (t2) が正しいかどうかを判定し，その回答根拠となる民法の該当部分を引用して示してください．
なお，あなたの回答はすべて JSON 形式のみで出力し，前後の説明文やコードフェンスなどは含めないでください．また JSON 形式は schema に従ってください．
法律に関する記述 (t2)
{t2_text}

B.2 実験 2: BM25 条文検索併用

System Prompt
あなたは日本の法律に詳しい法律専門家です．
出力は必ず JSON 形式のみで返し，他のテキストは一切含まないでください．

User Prompt
日本の司法試験の民法に関する択一回答式問題です．与えられた条文情報のみを参照して記述 (t2) が正しいかどうかを判定してください．
なお，あなたの回答はすべて JSON 形式のみで出力し，前後の説明文やコードフェンスなどは含めないでください．また JSON 形式は schema に従ってください．
参照する条文情報
{context}
法律に関する記述 (t2)
{t2_text}

B.3 実験 3: LLM 出力を利用した条文検索併用

System Prompt
あなたは日本の法律に詳しい法律専門家です．
出力は必ず JSON 形式のみで返し，他のテキストは一切含まないでください．

User Prompt
日本の司法試験の民法に関する択一回答式問題です．与えられた条文情報のみを参照して記述 (t2) が正しいかどうかを判定してください．
なお，あなたの回答はすべて JSON 形式のみで出力し，前後の説明文やコードフェンスなどは含めないでください．また JSON 形式は schema に従ってください．
参照する条文情報
{context}
法律に関する記述 (t2)
{t2_text}

C 評価指標について

本研究で扱った評価指標に関して以下に説明を示す．

C.1 F2-measure

COLIEE Task3 では Precision, Recall から算出される F2-measure が評価指標として採用されている [5]．

条文検索プロセスは法令のあてはめタスクにおいて，法的推論に先んじて行われるものであり，適切な法令を選ぶための候補を割り出すものであるため，F2-measure として Recall を重視している．

$$F2 = \frac{5\cdot\text{Precision}\cdot\text{Recall}}{4\cdot\text{Precision}+\text{Recall}} \tag{2}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive}+\text{False Positive}} \tag{3}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive}+\text{False Negative}} \tag{4}$$

C.2 Accuracy

COLIEE Task4 では法律に関するクエリに対して Y/N で回答したものに対する Accuracy が評価スコアとなる [5]．