Chain-of-Thought 過程の誘導による LLM の性能改善と推論過程および性能の説明性向上

藤田　真伎[1]　狩野　芳伸[1]
[1] 静岡大学
{mfujita,kano}@kanolab.net

概要

近年，大規模言語モデルの性能上は著しいものがあるが，その大規模さゆえにモデル構造や推論過程がブラックボックス化している問題がある．そこで，Chain-of-Thought の過程を誘導するフォーマットを提案し，推論過程のどの部分が特に推論結果に影響を与えているかを特定できるようにした．また，Fine Tuning により任意のモデルで同様のフォーマットでの出力ができること，この手法を用いることで複数モデル間で推論能力の差がある部分を明確にできることも示した．司法試験の民法分野の含意関係推論タスクを題材として，大規模言語モデルを用いたタスク性能の向上を達成した．

1　はじめに

近年，大量の学習データやパラメータで訓練された大規模言語モデル (Large Language Model，以下 LLM) を用いることでさまざまなタスクの性能向上が示されている．一方で，LLM にはそのパラメータ数の多さゆえに推論過程がブラックボックス化しているという問題があり，機械学習モデルの予測根拠を説明する XAI(eXplainable AI)[1] や説明可能性 [2] について研究されている．また，Chain-of-Thought[3](以下 CoT) などの手法を用いることで段階的な推論が可能になり，性能向上が報告されている．一方で，どの部分の推論が解答の助けになっているかは明らかでなく，また LLM の出力には幻覚 [4] が発生する場合もあり，段階的な推論を含めた出力の方法には改善の余地がある．

LLM の中でも Open AI 社の提供する GPT-4 はその性能が高い. しかし，GPT-4 はその学習データや内部構造が公開されておらず，動作の解釈や再現，カスタマイズがより難しい．

踏まえて，本研究の貢献は主に以下の二つである．

一つ目に，LLM が推論過程を示しながら出力を行うのに適したフォーマットを提案し，それにより CoT の過程を分析することで，モデルの推論に貢献している推論箇所を特定することや，モデルが苦手としている推論箇所を特定できることを示した．さらにこの提案フォーマットにより，LLM のタスク性能を向上させた．

二つ目に，提案フォーマットに沿って内部構造が公開されているモデルの Fine Tuning を行うことでフォーマットに沿った出力を可能にすることを確認した．さらに，段階的な推論が必要な法律分野のデータセットを用い，一般的な性能では劣る公開モデルでも，提案手法により推論段階によっては GPT-4 に匹敵する性能を達成した．

2　関連研究

2.1　Chain-of-Thought(CoT)

CoT とは LLM に問いに対する答えのみを出力させるのでなく，答えを推論するのに必要な中間ステップを出力させた後に答えを出力させる手法で，段階的な推論が必要となるタスクを中心に精度の向上が報告されている．段階的に推論する旨を指示に書く ZeroShot での CoT や，実際に段階的に推論を行う例を与える Few Shot[5] での CoT などがある．

2.2　OpenAI GPT-4

OpenAI が提供する GPT-4 [6] は Transformer [7] の Decoder 部分を用いた LLM である．多言語を扱うことができ，かつ様々なタスクで高精度な回答ができる．たとえばアメリカの司法試験の模擬試験においては上位 10%に相当するスコアが示されている．一方で，具体的なモデルの内部構造や訓練データなどは公開されておらず，内部構造がブラックボックスであるために分析や改善が難しい面がある．

2.3 日本語 LLM

本研究では，我々自身でモデルの Fine Tuning 等を行い比較するために，内部構造が公開されており訓練可能な形で利用できる LLM が必要である．日本語専用に事前学習された LLM [8] [9] や，多言語が扱えるモデル [10] [11] など非常に多くのモデルが公開されている．

2.4 法自然言語処理タスク COLIEE

法律分野の自然言語処理技術に関するワークショップとして，我が国司法試験の民法分野短答式問題の自動解答を行い性能を競う Competition on Legal Information Extraction and Entailment(以下 COLIEE) [12][13][14][15][16][17][18][19][20][21] が毎年開催されてきた．COLIEE の Task4 は民法条文一覧と過去問題が与えられ，それをもとに自動で解答を行うシステムを作成し，テスト問題の解答の精度を競うものである．このタスクは，問題文とその問題を解くのに必要な関連民法条文の二つが与えられ，問題文が正しいか否かを Yes か No の二択で答え，解答の精度を競うものである．我々は，段階的な推論が有効な分野の一つと考えられる法律分野を対象に，COLIEE データセットを用いて実験を行った．

3 提案手法

本論文では，CoT の出力フォーマットを固定することで推論過程を比較可能にし，タスクに応じてそれぞれの推論ステップごとの有効性とモデルの改善点を示す手法を提案する．

1. 対象タスクの特徴や課題をもとに，タスク固有のプロンプトと，タスクの回答にあたる CoT の出力フォーマットを設定する．
2. 任意の生成モデルに Few Shot Prompting や Fine Tuning を行い，設定したフォーマットに沿った出力を行わせる．
3. どの推論ステップで最終的な出力が大きく変化するか，また他のモデルと比較しどのステップに違いがあるかを分析することで，推論ステップごとの有効性とモデルの改善点を明確にする．

4 実験

COLIEE を題材に提案手法の実験と分析を行う．COLIEE 配布の司法試験年度に対応したデータのうち評価には R01〜R03 の三年分の問題 255 問を使用し，後述の公開モデルの Fine Tuning には H18〜H30 の 629 問を使用した．

生成モデル一般に，個別タスクにおいて具体的にどのような能力が優れているかを特定することが難しい．例えば COLIEE の場合，民法条文から適切な部分を見つける能力や，論理的な推論能力など，様々な可能性がある．そこで，共通した出力フォーマットと CoT を用い，GPT-4 と公開モデルを比較することで，二つのモデルの性能差を定量化する．本論文では数種類のモデルを比較実験したうえで，手元で学習可能なモデルサイズのうち公開モデルとして正答率が高かった stabilityai/StableBeluga13B [10] を用い，GPT-4 と比較することとした．

4.1 プロンプトと出力フォーマット

異なるプロンプトや Few Shot を比較した結果，図 1 のプロンプトおよびフォーマットで出力することとした．すなわち，フォーマットを「問題要点・民法要約・推論箇所・回答」の四つの要素として設定する．この四要素を定めた理由は以下の通りである．まず，司法試験の民法問題では含意か否かを判定する際の要点が存在することが多い．また，与えられる民法条文は必ずしも必要最小限ではなく，現実の裁判でも膨大な条文から必要箇所を抽出しなければならない．最後に，問題の要点と民法条文の要約から論理的な推論を行うステップを経て，最終的な解答を出力するフォーマットを設定した．

4.2 公開モデルの Fine Tuning と GPT-4 の Few Shot Prompting

stabilityai/StableBeluga13B は，Few Shot Prompting のみでは出力フォーマットが固定されず CoT での段階的な推論ができなかったため，Fine Tuning を行った．このモデルでフォーマットの学習ができれば他の多くのモデルでも提案手法が実現可能と考える．

学習データは図 1 のフォーマットに沿うよう GPT-4 を用いて作成した．COLIEE 配布の H18〜R03 までのデータ中で GPT-4 が正解を出力した 629 例の推論過程つきの出力をもとに，Lora[22][23] を用いて学習する．

GPT-4 は執筆時点で Fine Tuning が利用できなかっ

たため，Few Shot で正例と負例を毎回固定した同じ一例ずつを与えた．なお，Few Shot で与える例は GPT-4 の出力をもとに人手で調整したものである．

指示
以下は司法試験の民法問題です．民法条文部分に記載されている関連民法条文を参照し問題が正しいか正しくないか，問題の要点を述べた後に正しい場合は「True」を正しくない場合は「False」を出力しなさい．
民法条文
第五条
未成年者が法律行為をするには，その法定代理人の同意を得なければならない．ただし，単に権利を得，又は義務を免れる法律行為については，この限りでない．(略)
問題文
未成年者がした売買契約は，親権者の同意を得ないでした場合であっても，その契約が日常生活に関するものであるときは，取り消すことができない．
回答
問題の要点：未成年者がした売買契約が親権者の同意を得ないでした場合，その契約が日常生活に関するものであるとき，その契約を取り消すことができるかどうか．
民法条文によれば，未成年者が法律行為をするには，その法定代理人の同意を得なければならない．しかし，単に権利を得，又は義務を免れる法律行為については，この限りでないとされています．また，法定代理人が目的を定めて処分を許した財産は，その目的の範囲内において，未成年者が自由に処分することができるとされています．
したがって，未成年者がした売買契約が日常生活に関するものである場合，その契約は法定代理人の同意を得ないでしたにせよ，未成年者が自由に処分することができる範囲内にあると解釈できます．そのため，その契約を取り消すことはできないと考えられます．
回答：True

図 1　プロンプト・出力フォーマットの例

4.3　GPT-4 による有効な推論ステップ特定

我々は GPT-4 に CoT を用いることで COLIEE タスクの正解率向上を確認した (A.1)．生成した推論過程のどの部分がモデルの回答に影響を与えているか特定するため．一度 CoT を用いて段階的な推論を行った後に，4.1 節の 4 つの要素に基づいて分割を行った．その後，分割後の途中までの推論過程をもとにモデルにタスクの回答をさせることで，どの要素が推論に大きな影響を与えるか分析した (図 2)．

問題要点と条文要約の出力は，元の条文や問題文が長いほど効果的に作用すると考えられる．そこで評価データを，条文および問題文の長さでそれぞれ二分割し，問題と条文の長さが正答率の変化に影響

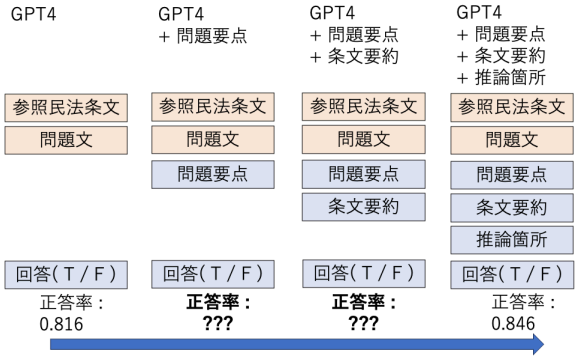

図 2　GPT-4 を用いた推論過程ごとの正答率の定量化

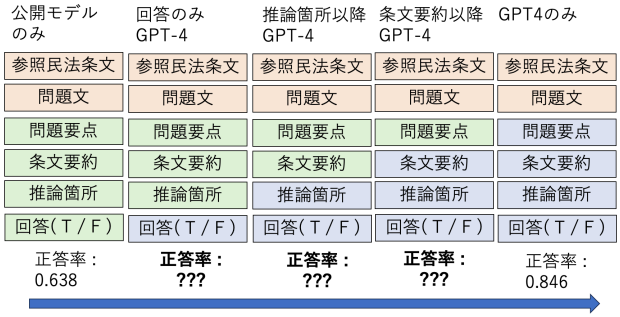

図 3　モデル性能の比較手法

を与えるか確認した．

4.4　二種モデル間の性能比較

先の 4 要素に沿った段階的な出力 (図 3) を，公開モデルに CoT を用いて行わせた．その後，要素ごとに分割を行い，GPT-4 に以降を推論させてから正答率を算出することで，CoT のどの段階においてモデルの性能の差が生まれているかを分析した．

5　実験結果

5.1　モデルの学習

4.2 節で記載した，公開モデルの学習結果を表 1 に示す．Fine Tuning により問題の解答率精度の向上とともに，出力フォーマットに沿った出力を行えるようになったことが分かる．

5.2　GPT-4 による有効な推論ステップ特定

4.3 節の CoT の有効な推論ステップの特定の実験結果を表 2 に示す．評価データは COLIEE の R01〜R03 の問題 255 問中，公開モデルと GPT-4 がともにフォーマットに沿った出力を行えた 240 問に限定した．問題文文字数が評価データにおける中央値 (70 文字) 以上の場合を「問題長文」とし，該当する 121 問に絞った正答率を記載した．同様に問題短文 119

表 1 StableBeluga-13B の学習前後の結果比較

	訓練前	訓練後
正答率	0.597 (151/253)	0.641 (161/251)
フォーマットを満たす出力	0 (0/255)	0.953 (243/255)

表 2 推論過程ごとの正答率の変化結果

	GPT-4	GPT-4+ 問題要点	GPT-4+ 問題要点+ 条文要約	GPT-4+ 問題要点+ 条文要約+ 推論箇所
正答率	0.817 (196/240)	0.808 (194/240)	0.838 (201/240)	0.846 (203/240)
問題長文	0.793	0.802	0.843	0.851
問題短文	0.840	0.815	0.832	0.840
条文長文	0.842	0.808	0.858	0.875
条文短文	0.792	0.808	0.817	0.817

問，民法条文でも中央値である 170 文字を基準に条文長文 120 問と条文短文 120 問の正答率を記載した．いずれも長文なほど正答率が向上しており，要約の効果があったと考えられる．

5.3 二種モデル間の性能比較

4.4 節に沿って各ステップごとの精度を求めたものが表 3 である．各要素は図 3 に対応している．すなわち，すべてを公開モデルで実行する場合に始まり，図表の右にいくほど GPT-4 が分担する要素が増え，一番右では GPT-4 がすべてを実行した場合であり，右に行くほど性能が向上しているが，「推論箇所以降」より先はあまり変化がない．

6 分析と考察

6.1 公開モデルの Fine Tuning

これにより，任意のモデルでフォーマットに適した段階的な出力をすることで，提案手法の一つである CoT を用いた性能比較を行えるようになる可能性を示唆した．

6.2 GPT-4 による有効な推論ステップ特定

表 2 の結果を分析する．推論箇所を出力することで，文字数に関わらず全体的に正答率が向上しており，問題と条文の要約をした後にも一度テキスト形式で論理的な推論を挟むことで適切な推論が行えた，有効な推論ステップと考えられる．条文要約を用いることで条文長文，条文短文の場合ともに正答率が向上し．特に条文が長文のときに大きく正答率が向上していることから長文を扱いやすくするという要約の意図通りに作用した有効な推論ステップと考えられる．問題要点を用いることで問題長文の正答率が上がっており，こちらも要点化の有効性を示唆している．問題短文の場合の正答率は低下していることから，問題の要点のみを記載するフォーマットは正答率の低下を招く可能性を示唆している．これは入力部分に書かれている民法条文と出力部分に書かれている問題の要点の比較が難しく，回答に必要な情報の一部のみを出力させると精度が低下するる可能性を示唆している．総合すると，要素ごとに分割したことで，それぞれの要素が正答率の向上に寄与したことを定量的に分析できたといえる．

表 3 モデル性能の比較結果

	正答率
公開モデルのみ	0.638 (153/240)
回答のみ GPT-4	0.771 (185/240)
推論箇所以降 GPT-4	0.854 (203/240)
条文要約以降 GPT-4	0.862 (207/240)
GPT-4 のみ	0.846 (201/240)

6.3 二種モデル間の性能比較

表 3 の結果から，GPT-4 と公開モデルを比較した時に，どちらがどの点で優っているか定量的に分析できることを示す．推論箇所以降 GPT-4，条文要約以降 GPT-4，GPT-4 のみの結果に大きな差がない（表 3 右側）ことから，公開モデルと GPT-4 とで問題の要点の抽出と民法条文の要約性能に大差がないと考えらえる．一方で推論箇所や回答までを公開モデルが行うと精度が大きく低下しており（表 3 左側），含意関係認識の部分で大きな性能差があることを示唆している．

7 おわりに

CoT において出力フォーマットや推論過程を固定することで性能向上させるとともに，任意のモデル間で性能差を段階的かつ定量的に分析する手法を提案した．COLIEE 司法試験の自動解答を題材に，実際に性能が向上し，現在の司法試験自動解答の課題である，推論過程とモデル構造のブラックボックス部分を解消できる可能性を示唆した．

今後は，分析により特定された不足する性能について，データの拡張やアンサンブルにより改善を試みたい．また，テンプレートの生成，モデルの性能調査，改善の流れを用いることで，定量的な分析に基づいたモデル改善を行うことができ，これを他のタスクにも応用できることを確認したい．

謝辞

本研究は JSPS 科研費 JP22H00804， JP21K18115，JP20K20509， JST AIP 加速課題 JPMJCR22U4， およびセコム科学技術財団特定領域研究助成の支援をうけたものです．

参考文献

[1] 恵木正史. Xai(explainable ai) 技術の研究動向. 日本セキュリティ・マネジメント学会誌, Vol. 34, No. 1, pp. 20–27, 2020.

[2] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017. https://arxiv.org/abs/1702.08608.

[3] 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, 2023. https://arxiv.org/abs/2201.11903.

[4] 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, 2023. https://arxiv.org/abs/2311.05232.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[6] OpenAI. Gpt-4 technical report, 2023. https://arxiv.org/abs/2303.08774.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[8] Tianyu Zhao, Akio Kaga, and Kei Sawada. rinna/youri-7b. https://huggingface.co/rinna/youri-7b.

[9] Jun Suzuki Namgi Han Saku Sugawara Shota Sasaki Shuhei Kurita Taishi Nakamura Takumi Okamoto. Hirokazu Kiyomaru, Hiroshi Matsuda. llm-jp/llm-jp-13b-instruct-full-jaster-dolly-oasst-v1.0. https://huggingface.co/llm-jp/llm-jp-13b-instruct-full-jaster-dolly-oasst-v1.0.

[10] Stability AI. stabilityai/stablebeluga-13b, 2023. https://huggingface.co/stabilityai/StableBeluga-13B.

[11] Xwin-LM Team. Xwin-lm, 9 2023. https://github.com/Xwin-LM/Xwin-LM.

[12] Competition on legal information extraction/entailment (COLIEE-14) workshop on juris-informatics (jurisin) 2014, 2014. http://webdocs.cs.ualberta.ca/~miyoung2/jurisin_task/index.html.

[13] Mi-Young Kim, Randy Goebel, and Satoh Ken. COLIEE-2015: evaluation of legal question answering. In **Ninth International Workshop on Juris-informatics (JURISIN 2015)**, 2015.

[14] Mi-Young Kim, Randy Goebel, Yoshinobu Kano, and Ken Satoh. COLIEE-2016: evaluation of the competition on legal information extraction and entailment. In **International Workshop on Juris-informatics (JURISIN 2016)**, 11 2016.

[15] Yoshinobu Kano, Mi-Young Kim, Randy Goebel, and Ken Satoh. Overview of COLIEE 2017. In Ken Satoh, Mi-Young Kim, Yoshinobu Kano, Randy Goebel, and Tiago Oliveira, editors, **COLIEE 2017. 4th Competition on Legal Information Extraction and Entailment**, Vol. 47 of **EPiC Series in Computing**, pp. 1–8. EasyChair, 2017.

[16] Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, Yao Lu, Juliano Rabelo, Naoki Kiyota, Randy Goebel, and Ken Satoh. COLIEE-2018: Evaluation of the competition on legal information extraction and entailment. In Kazuhiro Kojima, Maki Sakamoto, Koji Mineshima, and Ken Satoh, editors, **New Frontiers in Artificial Intelligence**, pp. 177–192, Cham, 2019. Springer International Publishing.

[17] Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. A summary of the COLIEE 2019 competition. In Maki Sakamoto, Naoaki Okazaki, Koji Mineshima, and Ken Satoh, editors, **New Frontiers in Artificial Intelligence**, pp. 34–49, Cham, 2020. Springer International Publishing.

[18] Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. COLIEE 2020: Methods for legal document retrieval and entailment. In **New Frontiers in Artificial Intelligence: JSAI-IsAI 2020 Workshops, JURISIN, LENLS 2020 Workshops, Virtual Event, November 15–17, 2020, Revised Selected Papers**, p. 196–210, Berlin, Heidelberg, 2020. Springer-Verlag.

[19] 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) 2021. **The Review of Socionetwork Strategies**, Vol. 16, No. 1, pp. 111–133, 2022.

[20] 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**, p. 51–67, Berlin, Heidelberg, 2023. Springer-Verlag.

[21] Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. Summary of the competition on legal information, extraction/entailment (COLIEE) 2023. In **Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law**, pp. 472–480, 2023.

[22] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In **International Conference on Learning Representations**, 2022. https://openreview.net/forum?id=nZeVKeeFYf9.

[23] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

A　参考情報

A.1　GPT-4 を用いた結果比較

表 4 は，FewShot と Chin-of-Thought(表内 CoT) をそれぞれ用いた場合，用いなかった場合の計四通りの正答率を示したものである．CoT を用いない場合は True/False のみを出力している．

表 4　GPT-4 の推論結果

	CoT + FewShot	CoT + ZeroShot	Few Shot	ZeroShot
正答率	0.846 (215/255)	0.8549 (218/255)	0.815 (208/255)	0.792 (202/255)
要点出力数	1 (255/255)	0.823 (210/255)	0.004 (1/255)	0 (0/255)

A.2　関連民法条文が長文の例

COLIEE で与えられる問題には民法条文が長文の場合があり，図 4 はその一例である．これは文字数にすると 654 文字に相当し，モデルによっては文字列長が推論に悪影響を与える場合も多い．

第四百六十五条の三　個人根保証契約であってその主たる債務の範囲に金銭の貸渡し又は手形の割引を受けることによって負担する債務（以下「貸金等債務」という。）が含まれるもの（以下「個人貸金等根保証契約」という。）において主たる債務の元本の確定すべき期日（以下「元本確定期日」という。）の定めがある場合において、その元本確定期日がその個人貸金等根保証契約の締結の日から五年を経過する日より後の日と定められているときは、その元本確定期日の定めは、その効力を生じない。
２　個人貸金等根保証契約において元本確定期日の定めがない場合（前項の規定により元本確定期日の定めがその効力を生じない場合を含む。）には、その元本確定期日は、その個人貸金等根保証契約の締結の日から三年を経過する日とする。
３　個人貸金等根保証契約における元本確定期日の変更をする場合において、変更後の元本確定期日がその変更をした日から五年を経過する日より後の日となるときは、その元本確定期日の変更は、その効力を生じない。ただし、元本確定期日の前二箇月以内に元本確定期日の変更をする場合において、変更後の元本確定期日が変更前の元本確定期日から五年以内の日となるときは、この限りでない。
４　第四百四十六条第二項及び第三項の規定は、個人貸金等根保証契約における元本確定期日の定め及びその変更（その個人貸金等根保証契約の締結の日から三年以内の日を元本確定期日とする旨の定め及び元本確定期日より前の日を変更後の元本確定期日とする変更を除く。）について準用する。

図 4　民法条文が長文の例 (R1-17-U)

A.3　モデル別の評価比較

openAI の提供するモデル，公開モデルなどについて正答率をまとめたものを表 5 に示す．

パラメータなどはモデルの推奨値を使っているが，temperature や top_p の値は最低限の再現性を確保するため低く設定した．与えるプロンプトは図 1 をもとに，各モデルの入力フォーマットに合わせ調整した．正例と負例を一例ずつ与える Few Shot と，CoT を組み合わせている．

正答率は解答できたものに限定しているため，True/False を出力に含まない問題は除外している．そのため，モデルによっては分母が 255 問にならないものもある．

表 5　各モデルの推論結果

モデル名	正答率	解答数
stabilityai/StableBeluga-13B	0.597 (151/253)	0.992 (253/255)
matsuo-lab/weblab-10b-instruction-sft	0.523 (126/241)	0.945 (241/255)
gpt-3.5-turbo-0613	0.651 (164/252)	0.988 (252/255)
TheBloke/Xwin-LM-70B-V0.1-GPTQ	0.608 (155/255)	1.0 (255/255)
Xwin-LM-7B-V0.1-GPTQ	0.55 (131/238)	0.9333 (238/255)
ELYZA-japanese-Llama-2-7b-instruct	0.525 (105/200)	0.784 (200/255)
llm-jp/llm-jp-13b-instruct-full-jaster-v1.0	0.629 (158/251)	0.984 (251/255)
llm-jp/llm-jp-13b-instruct-full-jaster-dolly-oasst-v1.0	0.603 (152/252)	0.988 (252/255)
rinna/youri-7b-instruction	1.0 (2/2)	0.008 (2/255)
rinna/japanese-gpt-neox-3.6b-instruction-ppo	0.519 (83/160)	0.627 (160/255)