日本の司法試験を題材とした GPT モデルの評価

Evaluating GPT in Japanese Bar Examination: Insights and Limitations

チェ ジョンミン [1]　笠井淳吾 [2]　坂口慶祐 †[1,3]
[1] 理化学研究所 AIP　[2]Kotoba Technologies, Inc.　[3] 東北大学大学院情報科学研究科
jungmin.choi@riken.jp, jkasai@kotoba.tech, keisuke.sakaguchi@tohoku.ac.jp †: 責任著者

Abstract

Large-scale language models like ChatGPT have been reported to exceed the accuracy of human experts in a wide range of tasks. Recent research reports that ChatGPT passed the Japanese National Medical Examination, confirming its high performance in Japanese. We evaluated the accuracy of GPT-3, GPT-4, and ChatGPT in the Japanese Bar Examination (the multiple-choice format section), focusing on Constitutional Law, Civil Law, and Criminal Law over the past five years. The results revealed that the current correct answer rate for these exams is only 30-40% (compared to the average pass rate of 70%), which is significantly low. This study went beyond just the correct answer rate, dissecting the necessary reasoning and knowledge for the responses, and examining the performance of large-scale language models from each perspective. The findings show that 1) large-scale language models possess extensive knowledge of many statutes, 2) they have a high correct answer rate for questions that require understanding of legal theories but not specific knowledge of law, and 3) they have a low correct answer rate for questions requiring knowledge of case law. The primary reason for their lower performance compared to the American Bar Examination is thought to be a lack of knowledge in Japanese law, especially in case law.

Keywords: Natural Language Processing, Large Language Models, Bar Examinations

概要

ChatGPT などの大規模言語モデルが，多岐にわたるタスクにおいて人間の専門家の精度を上回ると報告されている．とくに日本の医師国家試験に ChatGPT が合格したという最近の研究報告からも，日本語についての高い性能が確認されている．本研究では，日本の司法試験（短答式）の憲法，民法，刑法それぞれ過去 5 年分を対象に，GPT-3, GPT-4 および ChatGPT の精度を評価した．結果として，現段階では日本の司法試験に対する正答率が 3〜4 割と，合格水準に比べ非常に低いことが明らかになった．本研究では，単なる正解率にとどまらず，回答に必要な知識，能力を分解し，それぞれの観点での大規模言語モデルの性能を検証した．その結果，1) 大規模言語モデルは多くの条文の知識を有していること，2) 特定の条文や判例の知識を必要としないが学説の理解を必要とする問題に関しては正解率が高いこと，3) 判例の知識を必要とする問題に関しては正解率が低いこと，が示された．アメリカの司法試験と比較して性能が低い原因の大部分は，日本法の知識，とくに判例の知識の乏しさにあると考えられる．

キーワード：自然言語処理，大規模言語モデル，司法試験

1　はじめに

大規模言語モデルはさまざまな一般的なタスクだけでなく，高度に専門的な他タスクでも人間の水準に匹敵することから大きな注目を集めている (Brown 他, 2020; Wei 他, 2022)．Kung 他 (2023) は，米国医師免許試験 (United States Medical Licensing Examination) において，ChatGPT が合格水準の正解率を得たことを示す．Choi 他 (2023) は，法科大学院の論述試験で，単位認定の水準を満たす答案を作成したと報告する．また，GPT-4 はアメリカの司法試験である Uniform Bar Exam (択一式試験) において合格水準を大きく超え，受験者の上位 10%と同等の成績を修めたとされる (OpenAI, 2023).

日本語でも，GPT-4 等の大規模言語モデルは高い性能を示している．日本語の多様な言語処理タスクを収録した評価用データセットである JGLUE (Kurihara 他, 2022) において，GPT-4 はゼロショットで人間の回答に肉薄する優れた成績を

得た．また，日本の医師国家試験においては，数ショットの設定で合格水準の成績を得た (Kasai 他, 2023).

本研究では，日本の司法試験（短答式試験）を取り上げる．司法試験は，法律という領域の特性上，極めて地域性が高いため，英語で英米法の司法試験問題に回答できるモデルが日本法の司法試験問題に対応できるかは自明ではない．

実験の結果，この試験に関して現状の大規模言語モデルは人間の合格水準にははるかに及ばなかった．GPT-4 はランダム選択によるベースラインを上回るものの，正解率は 30-40%程度に留まっており，GPT-3，ChatGPT についてはランダム選択と変わらない正答率であった．また，問題類型ごとの性能を検証したところ，1) 大規模言語モデルは多くの条文の知識を有していること，2) 特定の条文や判例の知識を必要としないが学説の理解を必要とする問題[1)]に関しては正解率が相対的に高いこと，3) 判例の知識を必要とする問題に関しては正解率が低いこと，が示された．つまりアメリカの司法試験と比較して性能が低い原因の大部分は，日本法の知識，とくに判例の知識の乏しさにあると考えられる．

大規模言語モデルによる日本の司法試験の正解率が低い点を踏まえて，より詳細な検証をするため，(1) 正誤判定タスク，(2) 条文補完タスクという 2 種類の緩和問題を準備した．正誤判定タスクでは，司法試験の問題形式の複雑さを緩和して正誤判定の二値分類に置き換え，条文補完タスクでは，条文を途中まで入力して続きを出力させた．実験の結果，(1) に関しては正解率はほぼ変わらず，司法試験の正誤問題自体の難しさが示された．(2) に関しては，GPT-4 は条文の知識を相当程度有していることが観察された．

専門的な法的サービスは，その社会的な意義，需要に比して供給が限定されており，大規模言語モデルによる補助や自動化が期待されることは言を俟たない．本研究では大規模言語モデルによる日本の司法試験の精度を評価し，これまでに他のタスク等で報告されているような正解率には至っていないことを確認した．本研究に関するデータを公開し[2)]，今後の精度向上や言語モデルを用いた法的サービス応用に対する手がかりとなることを期待する．

1) 法の解釈をめぐる主要な学説を下敷きにして作成された問題である．学説そのものを暗記している必要はないが，議論の構成や批判の方法を理解している必要がある．

2) https://github.com/keisks/j_bar_exam

2 背景

ここでは，データの背景である日本の法曹養成制度と司法試験について概説する．

2.1 日本の法曹養成制度

司法試験に合格し，司法修習を修了することによって法曹資格が与えられる．司法試験受験資格は，法科大学院を修了することのほか，法科大学院修了相当の学識を有するかどうかを判定する試験（「予備試験」）に合格することによって与えられる．

2.2 司法試験の概要

司法試験は，短答式と論文式による筆記の方法により行われる．短答式試験は，**憲法，民法，刑法**の 3 科目について行われ，論文式試験は公法系科目（憲法，行政法関連），民事系科目（民法，商法，民事訴訟法関連），刑事系科目（刑法，刑事訴訟法関連）そして選択科目（専門的な法律分野に関する科目から受験者が選択する 1 科目）について行われる．本研究では，評価が容易な短答式試験に絞って大規模言語モデルの性能を検証する．

短答式試験

短答式試験は，「裁判官，検察官又は弁護士となろうとする者に必要な専門的な法律知識及び法的な推論の能力を有するかどうかを判定すること」[3)]が目的とされている．問題は多肢選択式で与えられ，法務省が定める短答式試験の合格基準は，各科目において，満点の 4 割（憲法 20 点，民法 30 点，刑法 20 点）以上の成績を得，各科目の合計得点が一定以上の成績[4)]を得ることである．

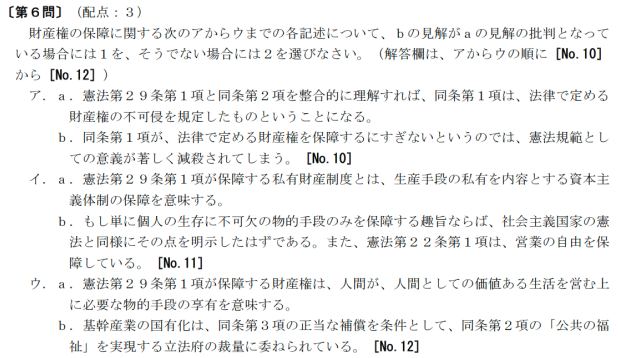
内容

各問は，回答に必要な知識または能力によって分類できる．すなわち，憲法: 条文の知識，判例の知識，学説の理解；民法：条文の知識，判例の知識；刑法：判例の知識，学説の理解である[5)]．図 1，2，3 にそれぞれの類型の代表的な問題を示す．

3) https://www.moj.go.jp/jinji/shihoushiken/shiken_shinshihou_shikenqa.html

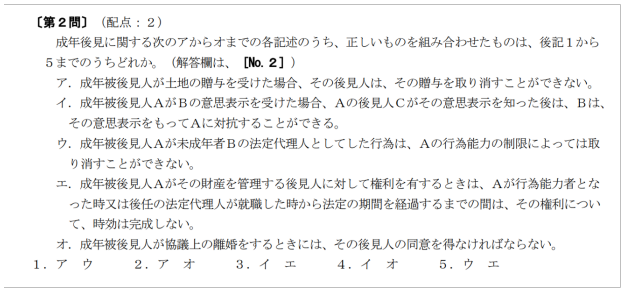
4) 年度によって異なり，2018 年度から 2023 年度では，満点 175 点中 96 点から 108 点

5) 厳密には，判例の知識は条文の知識を包含しているが，ここでは明示的に判例の知識を要求する問題を「判例の知識」に分類する．



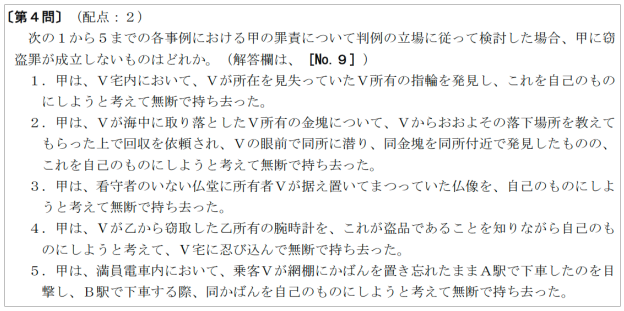
〔第６問〕（配点：３）
財産権の保障に関する次のアからウまでの各記述について、ｂの見解がａの見解の批判となっている場合には１を、そうでない場合には２を選びなさい。（解答欄は、アからウの順に［No. 10］から［No. 12］）
ア．ａ．憲法第２９条第１項と同条第２項を整合的に理解すれば、同条第１項は、法律で定める財産権の不可侵を規定したものということになる。
ｂ．同条第１項が、法律で定める財産権を保障するにすぎないというのでは、憲法規範としての意義が著しく減殺されてしまう。［No. 10］
イ．ａ．憲法第２９条第１項が保障する私有財産制度とは、生産手段の私有を内容とする資本主義体制の保障を意味する。
ｂ．もし単に個人の生存に不可欠の物的手段のみを保障する趣旨ならば、社会主義国家の憲法と同様にその点を明示したはずである。また、憲法第２２条第１項は、営業の自由を保障している。［No. 11］
ウ．ａ．憲法第２９条第１項が保障する財産権は、人間が、人間としての価値ある生活を営む上に必要な物的手段の享有を意味する。
ｂ．基幹産業の国有化は、同条第３項の正当な補償を条件として、同条第２項の「公共の福祉」を実現する立法府の裁量に委ねられている。［No. 12］

図 1: 憲法，学説の理解を問う問題の例．たとえばアは，a が，当該条文を「法律で定める財産権の不可侵を規定したもの」と解釈している一方，同 b は，そう解釈することで不合理が生じると主張している．したがって，b は a の批判となっている．



〔第２問〕（配点：２）
成年後見に関する次のアからオまでの各記述のうち、正しいものを組み合わせたものは、後記１から５までのうちどれか。（解答欄は、［No. 2］）
ア．成年被後見人が土地の贈与を受けた場合、その後見人は、その贈与を取り消すことができない。
イ．成年被後見人ＡがＢの意思表示を受けた場合、Ａの後見人Ｃがその意思表示を知った後は、Ｂは、その意思表示をもってＡに対抗することができる。
ウ．成年被後見人Ａが未成年者Ｂの法定代理人としてした行為は、Ａの行為能力の制限によっては取り消すことができない。
エ．成年被後見人Ａがその財産を管理する後見人に対して権利を有するときは、Ａが行為能力者となった時又は後任の法定代理人が就職した時から法定の期間を経過するまでの間は、その権利について、時効は完成しない。
オ．成年被後見人が協議上の離婚をするときには、その後見人の同意を得なければならない。
１．ア　ウ　　２．ア　オ　　３．イ　エ　　４．イ　オ　　５．ウ　エ

図 2: 民法，条文の知識を問う問題の例．たとえばアは，民法第 9 条「成年被後見人の法律行為は，取り消すことができる． ただし，日用品の購入その他日常生活に関する行為については，この限りでない.」に反しており，誤っている．



〔第４問〕（配点：２）
次の１から５までの各事例における甲の罪責について判例の立場に従って検討した場合、甲に窃盗罪が成立しないものはどれか。（解答欄は、［No. 9］）
１．甲は、Ｖ宅内において、Ｖが所在を見失っていたＶ所有の指輪を発見し、これを自己のものにしようと考えて無断で持ち去った。
２．甲は、Ｖが海中に取り落としたＶ所有の金塊について、Ｖからおおよその落下場所を教えてもらった上で回収を依頼され、Ｖの眼前で同所に潜り、同金塊を同所付近で発見したものの、これを自己のものにしようと考えて無断で持ち去った。
３．甲は、看守者のいない仏堂に所有者Ｖが据え置いてまつっていた仏像を、自己のものにしようと考えて無断で持ち去った。
４．甲は、Ｖが乙から窃取した乙所有の腕時計を、これが盗品であることを知りながら自己のものにしようと考えて、Ｖ宅に忍び込んで無断で持ち去った。
５．甲は、満員電車内において、乗客Ｖが網棚にかばんを置き忘れたままＡ駅で下車したのを目撃し、Ｂ駅で下車する際、同かばんを自己のものにしようと考えて無断で持ち去った。

図 3: 刑法，判例の知識を問う問題の例．たとえば選択肢 1 は，大審院大正 15 年 10 月 8 日判決，刑集 5 巻 440 頁の判旨に反しており，誤っている．

2.3　データの作成

データの作成は，法務省が公開している過去の問題と解答を，JSON 形式のファイルに整理することによって行なった．2018 年度のデータを，大規模言語モデルに問題形式を学習させるためのプロンプトに使用し，2019 年度から 2023 年度までの問題を検証の対象とした．問題数は各年度，憲法 20 問，民法 36-37 問，刑法 20 問である．

3　実験と分析

司法試験回答タスク

実際の受験者の環境と同様，短答式試験の各科目各問に対して，図 1， 2， 3 のように，問題文と全選択肢を結合して入力とし，正解の選択肢を出力させることを目指した．モデルがこの問題形式に従って出力するよう，文脈内学習 (in-context learning) のプロンプトとして，2018 年度の問題からランダムに 5 問選択し，入力の先頭に追加した．2018 年度の問題はプロンプトにのみ使用し，評価の対象外とした．

3.1　モデルと評価

3.1.1　モデル

本実験では OpenAI 社が提供する 3 つの大規模言語モデル，GPT-3， ChatGPT (gpt-3.5-turbo) と GPT-4 の API を使用した．これらのモデルの訓練データおよび訓練方法については公開されていないが，Transformer を基にした自己回帰言語モデルと考えられている．

3.1.2　プロンプト

プロンプトとは，期待される出力の形式を規定するテキストであり，これによって出力が大きく左右されることが知られている．本実験では，2018 年の問題からランダムに選択した 5 問と正解をプロンプトとした．

3.2　結果

図 4 に各年度におけるモデルの正解率を示す．スコアは，年度で平均すれば GPT-4 が全ての科目において最も高く，ランダム選択による期待値を有意に上回っているが，正答率は 30-40%にとどまっており合格者平均（約 70%）には程遠い結果となった．ChatGPT と GPT-3 は，ランダム選択のベースラインの精度と変わらなかった．

3.3　分析

大規模言語モデルの性能をより詳細に調べるため，以下のような問題の類型ごとの分析を行っ

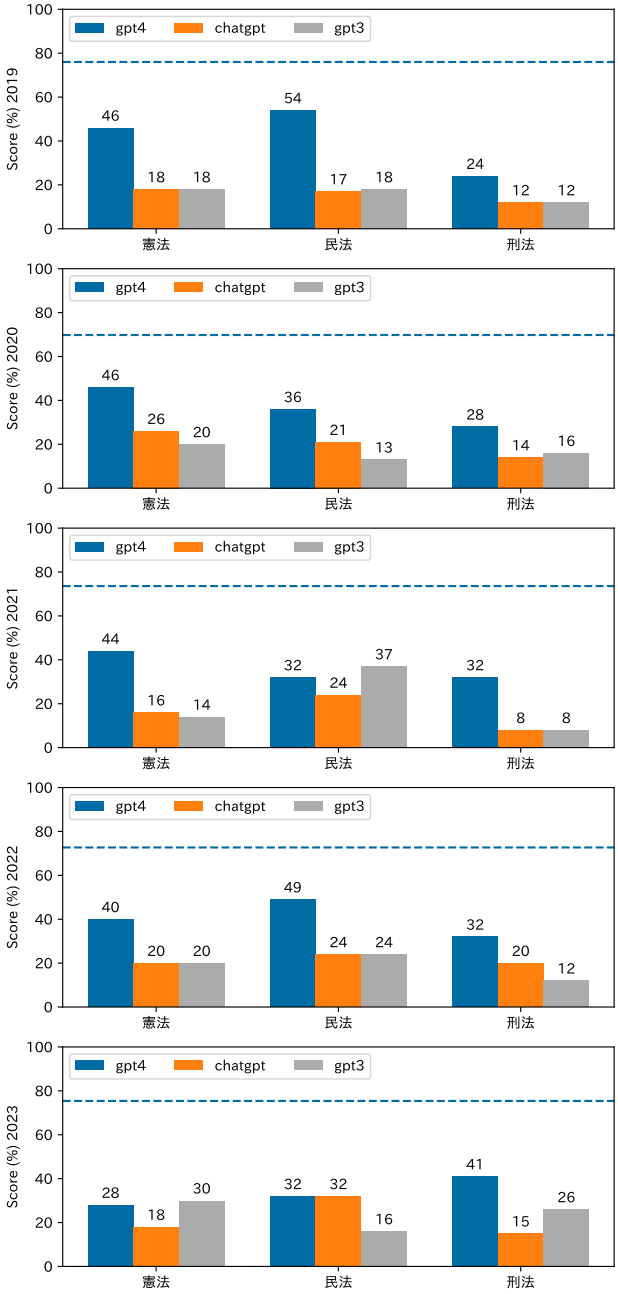


図 4: 大規模言語モデルによる 2019 から 2023 年度の各科目ごとの正解率．点線は各年ごとの合格者平均点を示している．

た．司法試験の問題は，主に学説の理解を問うもの (theory)，条文の知識を問うもの (statute)，判例の知識を問うもの（case) という 3 類型に分けることができる．2019 年度から 2023 年度の全科目の問題を類型ごとに採点した結果が表 5 である．GPT-4 の結果に注目すると，theory 型のスコアが最も高く，statute 型，case 型，とスコアが下がっていく傾向が見られる．すなわち，スコアは必要とされる法律の知識の量と反比例しており[6]，法律知識の不足が正解率を引き下げていることを示唆している．

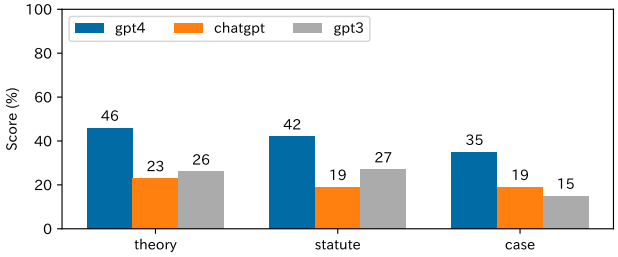


図 5: 問題類型ごとの正解率

表 1: 正誤判定タスクの結果

model	2023	2022	2021	2020	2019
GPT-4	**66.7**	**64.3**	**64.3**	**62.8**	**63.2**
ChatGPT	57.2	50.8	52.4	53.3	56.8
GPT-3	56.1	52.4	56.8	54.4	56.2

4 緩和タスク

正誤判定タスク

司法試験回答タスクでは，問題の形式が，「正しいもの/誤っているものの組合わせを選べ」など，複雑なものになっている．こうした形式の複雑さの影響を除去してモデルの性能を調べるため，問われる内容はそのままに，入力と出力を次のように修正したタスクで実験を行った．すなわち，問題文を，各文の正誤を問う形に変更し，正しいものには 1，誤っているものには 0 とラベルを付与したものを 915 問準備した[7]．正誤判定タスクの例は Appendix A（表 4）に示す．正誤判定タスクにおける各モデルの成績は，表 1 に示す通りである．本タスクと同様に，GPT-4 が最も高く，ランダム回答の正解率を優位に上回っていた一方，GPT-3 と ChatGPT はランダム回答の正解率を大きく上回ることはなかった．

条文補完タスク

大規模言語モデルがそもそもどれだけ日本の法律（条文）の知識を保有しているかを調べるため，各

6) theory 型は，特定の法律の知識ではなく，学説の構成や批判の方法についての理解を問うものであり，論理操作で解ける部分も大きい．

7) なお，短答式試験から作成された二値分類タスクという点で類似するものとして COLIEE Task 4 (Kim 他, 2023) があるが，COLIEE Task 4 は条文の知識のみで解ける問題に限定されており，各記述と条文と合わせた含意関係認識問題である点，本稿で提案する正誤判定タスクは，大規模言語モデルへの指示を含む形に書き換えられている点などが異なる．

表 2: 条文補完タスクの結果 (BERT score)

model	憲法	民法	刑法
GPT-4	**80.3**	**78.5**	**79.4**
ChatGPT	70.3	71.9	71.5
GPT-3	53.1	53.4	53.5

表 3: 条文補完タスクの結果 (BLEU score)

model	憲法	民法	刑法
GPT-4	**26.2**	**16.8**	**20.1**
ChatGPT	5.4	6.3	8.6
GPT-3	1.5	1.6	1.4

条文を文長で半分に分割し，前半を入力として後半を出力させることができるかというタスクを設定した．一つの条文が複数の文から成る場合は各文を独立に扱い，憲法: 475 文，民法: 2,355 文，刑法: 195 文を得た．なお，各法典の最初の 4 条をプロンプトとして与えている．条文補完タスクは生成問題であるため，厳密な評価は難しい．大まかな傾向を把握するため，本実験では自然言語処理で広く使われている 2 種類の自動評価指標を採用する．目標とする文との意味的な類似度を表す BERTScore[8] と，表層的な語彙の一致度を表す BLEU[9] を用いた．その結果を評価指標別に，それぞれ表 2 と表 3 に示す．いずれの評価指標においても，GPT-4 のスコアが最も高く，ChatGPT，GPT-3 が続く結果となった．GPT-4 に関していえば，大半の例で，元の条文に近い文を出力できているといえる．法律分野ごとの得手不得手があるかどうかを調べるため，民法典の各編の平均スコアを算出したが，どの編でもほぼ同じスコアであることがわかった．詳細については Appendix A（図 6）を参照されたい．なお，これらのスコアは生成された文と目標とする文との一般的な類似性を表すものではあるが，法律分野における厳密な意味類似性を評価することは難しく，正しく生成できているにもかかわらずスコアが低い例も多数ある．Appendix A（表 5, 6）に，具体的な生成例とスコアを示す．

5 関連研究

本節では法律タスクにおける大規模言語モデルの最近の応用を概観する．大規模言語モデルが有望視されている分野のひとつは，判決予測および法的推論である．Nguyen 他 (2023) は，日本の民法の条文に基づく二値分類の含意関係認識タスクにおいて，GPT-4 が 8 割を超える高い正解率を得たことを報告する．Trautmann 他 (2022) は，判決予測タスクにおける性能を向上させるために，法律分野に対応したプロンプトエンジニアリングを導入している．すなわち，最終的な判決を予測させる前に「違法性は認められるか」といった中間的な質問を挟む手法である．この手法は 3 つの多言語データセットにおいて有効であることが示されている．Blair-Stanek 他 (2023) は，GPT-3 の法的推論能力を調査し，数段階のプロンプトにより，このタスクでモデルが高い精度と信頼性を達成できることを明らかにしている．同様に Yu 他 (2022) は，法律家の分析的アプローチを模倣した Chain-of-Thought (CoT) プロンプトを導入し，大規模言語モデルが論理的に首尾一貫した関連性のある文章を生成するよう導く．この研究では，Nguyen 他 (2023) と同様の含意関係認識タスクにおいて，CoT プロンプトによるゼロショット回答が数ショット回答を上回ったことを報告している．Choi 他 (2023) は，ChatGPT が米国の法科大学院の論述試験に対して生成した答案は及第点を上回るレベルであると評価している．

大規模言語モデルは法に関する補助的なタスクへの応用も検討されている．Nay (2023) は，米国の裁判官が作成した法廷意見書をプロンプトとして特定の法の立法趣旨を説明させるタスクを提案し，GPT-4 が一定の理解を示したと述べている．Oltz (2023) は，大規模言語モデルが法学教員の周辺的な業務（推薦書や研究者紹介，学術会議におけるスピーチの執筆，試験問題の作成）において，実際の教員と比べても遜色のない性能を見せたという．さらに，Macey-Dare (2023) は，大規模言語モデルが弁護士による法律相談を代替する可能性を検討し，実際にあった判決を題材に，有罪判決を受けた被告人に対する助言を ChatGPT に生成させたところ，実際の弁護士には及ばないものの，上訴の可能性やその方法について妥当な助言を生成したと評価する．Iu and Wong (2023) は，ChatGPT は訴状や答弁書の骨組みを書くことや簡単な法的助言をすることができる，と分析する．

8） https://huggingface.co/spaces/evaluate-metric/bertscore

9） https://huggingface.co/spaces/evaluate-metric/sacrebleu

6 おわりに

本研究では，大規模言語モデルが日本の司法試験にどの程度対応できるかを評価した．その結果，現段階では合格水準とは大きく離れており，特に判例に関連した問題に対する対処能力が低いことが明らかになった．また，緩和実験からは大規模言語モデルが条文の知識を一定程度には保有しているものの，具体的な事例にあてはめ結論を導くことは難しいことがわかった．日本の司法試験は大規模言語モデルのベンチマークとしてはまだ難易度が高いことから，今後は緩和問題で行ったように単純な二値分類等のタスクから性能向上のベンチマークを準備したり，大規模言語モデルに日本の法律知識をどう学習させるのかあるいは抽出するのか，また日本の法律分野における厳密な意味類似性を評価できるような自動評価尺度の構築といった研究を進めることが考えられる．

謝辞

本稿の執筆にあたっては，吉永一行教授 (東北大学大学院法学研究科)，ならびに佐々木将也弁護士 (西村あさひ法律事務所・外国法共同事業) にご助言を賜りました．深くお礼申し上げます．

参考文献

[1] Blair-Stanek, Andrew, Nils Holzenberger, and Benjamin Van Durme (2023) 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: Association for Computing Machinery.

[2] Brown, Tom, 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 (2020) Language Models are Few-Shot Learners, in Larochelle, H., M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin eds. *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877–1901: Curran Associates, Inc.

[3] Choi, Jonathan H., Kristin E Hickman, Amy Monahan, and Daniel Schwarcz (2023) ChatGPT goes to law school., *Journal of Legal Education*.

[4] Iu, Kwansai and Vanessa Man-Yi Wong (2023) ChatGPT by OpenAI: The End of Litigation Lawyers? *SSRN Electronic Journal*.

[5] Kasai, Jungo, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir R. Radev (2023) Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations, *ArXiv*, Vol. abs/2303.18027.

[6] Kim, Mi-Young, Juliano Rabelo, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh (2023) COLIEE 2022 Summary: Methods for Legal Document Retrieval and Entailment, in Takama, Yasufumi, Katsutoshi Yada, Ken Satoh, and Sachiyo Arai eds. *New Frontiers in Artificial Intelligence*, pp. 51–67, Cham: Springer Nature Switzerland.

[7] Kung, Tiffany H., Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng (2023) Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models, *PLOS Digital Health*, Vol. 2, No. 2, pp. 1–12, 02.

[8] Kurihara, Kentaro, Daisuke Kawahara, and Tomohide Shibata (2022) JGLUE: Japanese General Language Understanding Evaluation, in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2957–2966, Marseille, France: European Language Resources Association, June.

[9] Macey-Dare, Rupert (2023) ChatGPT & Generative AI Systems as Quasi-Expert Legal Advice Lawyers - Case Study Considering Potential Appeal Against Conviction of Tom Hayes, *SSRN Electronic Journal*.

[10] Nay, John J. (2023) Large Language Models as Fiduciaries: A Case Study Toward Robustly Communicating With Artificial Intelligence Through Legal Standards, *ArXiv*, Vol. abs/2301.10095.

[11] Nguyen, Ha-Thanh, Randy Goebel, Francesca Toni, Kostas Stathis, and Ken Satoh (2023) Black-Box Analysis: GPTs Across Time in Legal Textual Entailment Task.

[12] Oltz, Tammy Pettinato (2023) ChatGPT, Professor of Law, *SSRN Electronic Journal*.

[13] OpenAI (2023) GPT-4 Technical Report, *ArXiv*, Vol. abs/2303.08774.

[14] Trautmann, Dietrich, Alina Petrova, and Frank Schilder (2022) Legal Prompt Engineering for Multilingual Legal Judgement Prediction, *ArXiv*, Vol. abs/2212.02199.

[15] Wei, Jason, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le (2022) Finetuned Language Models are Zero-Shot Learners, in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*: OpenReview.net.

[16] Yu, Fang, Lee Quartey, and Frank Schilder (2022) Legal Prompting: Teaching a Language Model to Think Like a Lawyer, *ArXiv*, Vol. abs/2212.01326.

A Appendix

表 4: 短答式問題とそれをもとに作成した正誤判定問題の例．問題文本文と，アの記述を組み合わせ，大規模言語モデルへの指示「正しければ〜」を含む形に書き換えている．残りのイからオの記述についても同様の方法で正誤判定問題を作成するが，ここでは省略している．

元の短答式問題	正誤判定問題（緩和）
制限行為能力者の行為であることを理由とする取消しに関する次のアからオまでの各記述のうち、誤っているものを組み合わせたものは、後記１から５までのうちどれか。 ア。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。 イ。成年被後見人がした売買契約は、成年後見人の同意を得て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を除き、取り消すことができる。 ウ。被保佐人がした保証契約は、保佐人の同意を得てした場合には、取り消すことができない。 エ。被補助人が、補助人の同意を得なければならない行為を、その同意又はこれに代わる家庭裁判所の許可を得ないでしたときは、その行為は取り消すことができる。 オ。成年被後見人の行為であることを理由とする取消権の消滅時効の起算点は、成年被後見人が行為能力者となった時である。 1: ア イ 2: ア オ 3: イ ウ 4: ウ エ 5: エ オ **(正解は 2)**	制限行為能力者の行為であることを理由とする取消しに関する次の記述につき、正しければ [1]、誤っていれば [0] と答えなさい。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。**(正解は 0)**

表 5: GPT-4 による条文補完の結果と BERTScore (BS)，BLEU スコア (BL)．法律上異なる効果である，「取り消すことができる」と「無効である」を混同する誤りが見受けられる一方，「この限りでない」の内容を正しく書き下すなど，実際の条文の文言とは隔たっているためスコアは低いが，条文の知識を正しく保有していることを示す例がある．

条文前半	条文後半	モデル出力	BS	BL
第百四十六条 （時効の利益の放棄） １ 時効の利益は、あらかじめ	放棄することができない。	放棄することができない。	1.0	1.0
第九十条 （公序良俗） １ 公の秩序又は善良の風俗に反	する法律行為は、無効とする。	する行為は、法律効果を生じない。	0.83	0.19
第五条 （未成年者の法律行為） １ 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。	る法律行為は、未成年者が行うことができる。	0.78	0.14
第五条 （未成年者の法律行為） ２ 前項の規定に反する法律行為	は、取り消すことができる。	は、無効である。	0.86	0.15

表 6: 各モデルで，最も BERTscore が高かった条文（a: GPT-4, b: ChatGPT, c: GPT-3) と，それに対する正解，及び全てのモデルの出力結果.

	条文前半	正解とモデル出力
(a)	第百四十六条　（時効の利益の放棄）　１　時効の利益は、あらかじめ	**(正解)** 放棄することができない。 ✓**(GPT-4)** 放棄することができない。 ✗**(ChatGPT)** 放棄されることを規定し、回答することが望ましい。 ✗**(GPT-3)** 放、棄、し、な、け、れ、ば、な、ら、な、い、。
(b)	第二百五十六条　（共有物の分割請求）　１　各共有者は、いつでも共有物の	**(正解)** 分割を請求することができる。 ✓**(GPT-4)** 分割を請求することができる。 ✓**(ChatGPT)** 分割を請求することができる。 ✗**(GPT-3)** 分、割、を、請、求、し、て、も、か、ま、わ、な、い、。
(c)	第千二十二条　（遺言の撤回）　１　遺言者は、いつでも、遺言の方式に従って、そ	**(正解)** の遺言の全部又は一部を撤回することができる。 ✓**(GPT-4)** の遺言を撤回することができる。 ✗**(ChatGPT)** うものをも含め，有効に撤回しうことが可能であるようにするべきである. ✓**(GPT-3)** の遺言を撤回することができる。

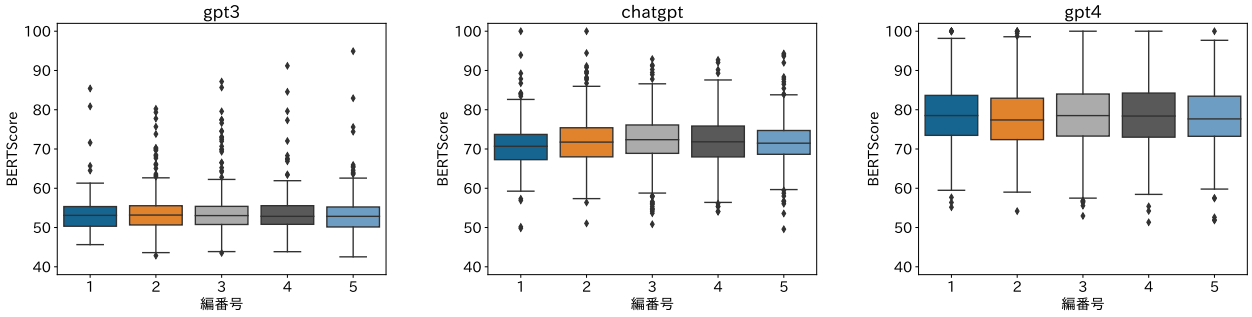


図 6: 条文補完タスクにおける，民法各編ごとの BERTScore の分布．GPT-4 が最も高く，GPT-3 が最も低いという傾向は変わらず，編ごとの差異はどのモデルにおいても見られなかった.