JBE-QA: 日本法知識を評価する司法試験質問応答データセット

Cao Zhihan[1], 西野文人 [2], 山田寛章 [1], Nguyen Ha Thanh[2,3], 宮尾祐介 [4], 佐藤健 [2]
[1] 東京科学大学
[2] 情報・システム研究機構データサイエンス共同利用基盤施設人工知能法学研究センター
[3] 国立情報学研究所　[4] 東京大学
cao.z.c8a7@m.isct.ac.jp　nishino@nii.ac.jp　yamada@comp.isct.ac.jp
nguyenhathanh@nii.ac.jp　yusuke@is.s.u-tokyo.ac.jp　ksatoh@nii.ac.jp

概要

本研究では，大規模言語モデルの日本法ドメイン知識を評価する質問応答データセット JBE-QA を構築した．JBE-QA は，2015 年から 2024 年までの司法試験の短答式試験をもとに作成され，民法・刑法・憲法の三分野を網羅する全 3,464 件の事例からなるベンチマークである．JBE-QA を用いて計 26 種類の大規模言語モデルを評価した結果，推論機能を有する商用モデルが最も高い性能を示した．

1　はじめに

法分野タスクは，人間の専門家の知識と経験に大きく依存する．専門家の負荷を軽減するため，法文書の効率的な検索や解析を可能にする法ドメイン言語処理の必要性が高まっている [1, 2].

大規模言語モデル（large language model, LLM）が法ドメイン言語処理に広く応用されている [3]．法ドメインでは高い信頼性が要求されるため，LLM にも法律分野に関する知識の正確さが求められる．このため，LLM の法律分野に関する知識の質を評価するためのリソースが開発されてきた．LegalBench [4] は，米国の憲法，刑法，民法を含む複数の分野を対象とした，法的推論に関するマルチタスクベンチマークである．MultiEURLEX [5] は，農業や運輸に関する法令の文書分類を目的としている，欧州連合の多言語法律コーパスである．また，JEC-QA [6] は，中国の司法試験から構築された大規模な質問応答データセットであり，憲法，刑法，民法など多様な分野の問題を含んでいる．

しかし，日本においては公開されているリソースは少なく，多くの場合で対象分野が民法に限られている．COLIEE データセット [7] は，民法に関する情報検索および含意関係認識タスクを含む．Choi らは，日本法ドメインにおける LLM の性能を評価するために司法試験短答式を元に二択形式の設問集を構築したが [8]，民法の一部の問題のみが公開されている．Japanese Tort-case Dataset [9] は，同じく民法に基づくデータセットであり，モデルが不法行為の判決を予測できるかを評価するためのもので，法ドメイン知識評価とは異なる．

そこで，本研究では日本法ドメイン向けに新たなデータセット「Japanese Bar Exam QA（JBE-QA）」を構築した．JBE-QA は，六法のうちでの重要な民法・刑法・憲法に関する全 3,464 件の問題を収録しており，既存リソースと比較してより網羅性高く，規模も大きい．さらに，本研究では JBE-QA を用いて，商用・オープンの別を問わず計 26 種類の LLM の性能を評価した．このベースライン評価は，現在の LLM が日本法に関してどの程度の知識を有しているかを示し，今後のモデル改善のための基準となる．JBE-QA の実験に用いたコード[1)] とデータセット[2)] はオンラインで公開している．

2　データセット構築

2.1　司法試験

司法試験は法曹資格を取得するための国家試験であり，短答式試験と論文式試験から構成される．短答式試験は，法曹として必要な法的知識および法的推論能力を測定することを目的としている．一方，論文式試験は，法的分析，構成，論述能力など，実務において必要とされる応用能力を評価することを目的としている．いずれの試験も，民法・刑法・憲法など複数の分野にまたがる多様な問題を含んで

1） https://github.com/hancules/JBE-QA
2） https://huggingface.co/datasets/nguyenthanhasia/japanese-bar-exam-qa-v2

いる．

2.2 問題設計

JBE-QA は，短答式試験を基に構築する．短答式試験は，法的知識の応用的能力を評価する論文式試験と異なり，法的知識により重点があるため，LLM の法的知識の理解を網羅的に評価するという目的に合致する．また，短答式試験は選択肢形式であるため，自動評価に適しているという利点がある．実際の短答式試験では，1 つの設問に対して複数の命題文が提示され，各命題は正か誤かのいずれかに分類される[3)]．各選択肢は，それら命題に対する正誤の組み合わせとして与えられる．受験者はすべての命題の正誤を正しく反映している選択肢を選ぶことを要求される．

JBE-QA では，実際の短答式試験を簡略化し，各設問を複数の二値分類問題に分割している．評価対象のモデルは，全命題の正誤の組み合わせではなく，個々の命題について一つずつその正誤を判定するタスクを解くことを要求される．この方式は短答式問題を取り扱った先行研究 [8, 7] に倣ったものであり，モデルの知識の正誤を命題個別に評価できる利点がある．

2.3 データ収集と処理

司法試験過去問は，法務省のウェブサイト[4)]で公開されている．本研究では，2015 年から 2024 年までの過去問の PDF をダウンロードし，XML ファイルへと変換したのち，事前処理を施した．形式の正規化および一般的な抽出エラーの検出には自動スクリプトを適用し，人手による確認と修正も行った．元の試験に含まれる一部の設問は，二値の正誤形式に変換できなかったため，除外した．最終的に除外された設問は，刑法から 46 問，憲法から 5 問，民法から 1 問の合計 52 問であった．

2.4 データセット構造

元の多肢選択問題を二値分類問題に変換するために，元問題を下記の通りのフィールドを持つように構造化した．JBE-QA 中の各事例は以下のフィールドを持つ．

- `id`: 基となる設問を一意的に定める識別子.
- `year`: 該当設問を含む試験の実施年号.
- `subject`: 法分野 （英語）.
- `subject_jp`: 法分野 （日本語）.
- `instruction`: 指示文.
- `question`: 命題文.
- `label`: ラベル (正なら Y で, 誤なら N).
- `answer`: 標準化されたラベル (正なら True で, 誤なら False).
- `theme`: 主題またはテーマ. 必要な場合のみ.
- `lead_in`: 指示文や命題文とは別の背景情報. 必要な場合のみ.
- `remark`: 注釈等. 必要な場合のみ.

元問題の全ての設問には，問題を解くための指示を明示する文章がある．この文章を抽出し，`instruction` というフィールドにしている．正誤の判断の対象である命題文は，`question` として抽出している．設問にスコープや主題を提示する文章を，`theme` として抽出している．

一部の設問にのみ特有のフィールドがある．設問の解答にあたって文脈や補足情報が必要とされる場合があり，そのための背景情報が提示されることがある．たとえば，別途提示された事例を参照する設問があり，その事例に関する記述が該当する．JBE-QA では，このような背景情報を `lead_in` フィールドとして抽出している．また，設問には，解答を容易にしたり，内容を単純化したりするための注釈や前提条件が明示されていることもある．たとえば，不動産に関する金銭分配の問題において，「損害賠償や強制執行費用は考慮しないこと」といった指示が該当する．このような記述は `remark` フィールドに格納している．

なお，データセットの品質と整合性を確保するための事後処理として，自動抽出過程において発生した設問と正解ラベルの不一致を検出・修正するために，人手による精査を実施した．

2.5 統計

JBE-QA は，合計 3,464 件の事例から成る．各事例は質問応答ペアである．そのうち，1,998 件は民法で，811 件は刑法で，655 件が憲法である．また，解答の正誤に関しては，「False」ラベルが 52.4%，「True」ラベルが 47.6%となっている．

3 ベースライン構築実験

現行の LLM を用いて JBE-QA を解かせ，ベースラインを確立する実験を行う．

3） 実際には，より複雑な構造を持つ設問も一部存在する．
4） https://www.moj.go.jp/barexam.html

JBE-QA の各事例は二値分類問題である．実験中，評価対象の LLM を，各事例に対して二値の真偽値を出力するよう指示するが (3.2 節)，一部のモデルは指示に従わずに二値以外の出力を返す場合がある．そのよう場合には，出力を 0（False）として扱う．性能の評価尺度には F1 値を用いる．また，指示に従う性能を定量化するため，回答率という尺度を使う．回答率は，あるモデルが指示通りに二値の真偽値を出力として返す事例の割合として定義される．各モデルは，0-shot と 4-shot の 2 つの設定で評価される．評価結果の比較可能性を担保する観点から，4-shot に使った四つの事例は各尺度の計算時に除外する．このため，各モデルの回答率と F1 値は 3460 件の事例で計算された値となる．

3.1 モデル

商用モデルとオープンモデルの両方に対して評価実験を行う．商用モデルは OpenAI [10] と Anthropic [11] が提供しているモデルを取り上げる．OpenAI のモデルは GPT-4.1，GPT-4o，GPT-5，o3 と o4-mini を用いる．Anthropic のモデルは，Claude の Opus-3，Opus-4.1，Sonnet-4，Sonnet-4.5 と Haiku-3.5 を用いる．Claude 系中，Claude Opus-4.1 と Sonnet-4 & 4.5 は extended thinking（拡張思考モード，以下 ER）[12] が可能なため，これらの三つのモデルに対して，ER 使用と不使用の両設定で評価する．

オープンモデルでは，多言語モデルに加えて日本語特化モデルも対象とする．対象オープンモデルは全て指示学習を経ており，15B 以上のパラメタを持つ．多言語モデルは，Qwen 2.5-32B & 72B [13]，GPT-OSS-20B & 120B [14]，Gemma 3-12B & 27B [15]，Llama 3.1 [16] & 3.3 [17] を用いる．日本語特化モデルは ABEJA-V2 [18]，Swallow-3.1 & 3.3 [19, 20, 21] 及び LLM-jp-3.1 の 13B の dense モデルと 8x13B の mixture-of-experts モデル [22] を用いる．Opus-4.1，Sonnet-4 と Sonnet-4.5 の ER 機能の使用の有無を考慮して，合計 26 のモデルを評価対象とした．

3.2 実験設定

モデルには「以下の法律に関する問題を解答せよ。理由や説明は不要。「正しい」と判断した時に 1 を、「誤り」と判断したときに 0 を出力せよ。出力は必ず 1 または 0 のいずれかの整数値のみとせよ。」というシステムプロンプトを構成した．ただし，LLM-jp-3.1 系モデルは，デフォルトではシステムプロンプトの指定ができないため，各入力（命題文）の前に，上記プロンプトを挿入する．

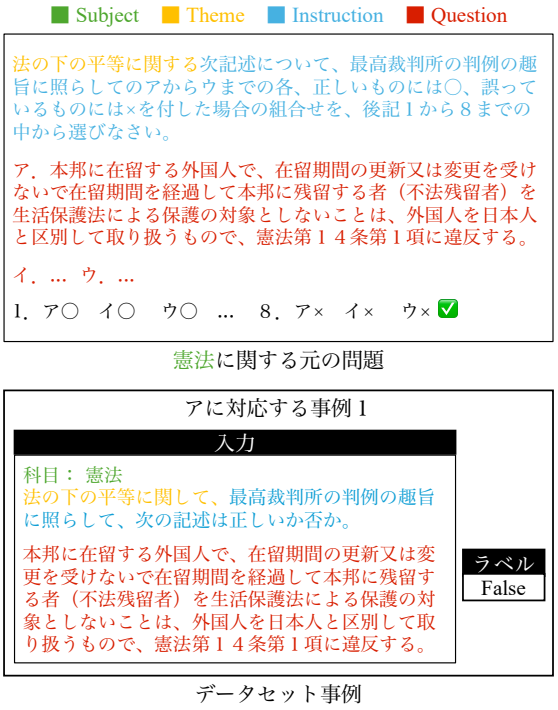

図 1 入力事例の例

4-shot に用いる 4 つの事例は，訓練集合から正例・負例をそれぞれ 2 件ずつサンプリングしたものであり，全モデルの評価に共通して使用される．

JBE-QA の構造を利用して，以下のように入力を構成する[5]．1 行目には，設問の法分野（`subject_jp`）が指定される．2 行目には，設問のテーマ（`theme`）が存在する場合に記述される．背景情報（`lead_in`）が存在する場合はテーマの直後に挿入される．その後，指示文（`instruction`）および命題文（`question`）が順に提示される．最後に，設問に注釈（`remark` フィールド）がある場合は，末尾に付加される．図 1 では入力事例の一例が示されている．

非推論モデルの評価では，オープンか商用かを問わず，温度を 0 に固定する．非推論モデルの最大出力トークン数は 1,000 に固定している．オープン推論モデルの評価では温度を 1 に設定し，商用推論モデルの評価では温度をデフォルト値のままにする[6]．十分な推論生成用トークンを確保するため，推論モデルの最大出力トークン数は，各モデルにおける出力上限の 4 分の 1 に設定する．全モデルとも，評価は 1 回のみ実施される．

5） フォーマッタ用のスクリプトはオンラインで入手可能．
6） 多くの場合は 1 であるが，GPT-5 や o3 など一部のモデルでは温度を変更できない．

4 結果

O	J	R	Model	F1		回答率	
				0-shot	4-shot	0-shot	4-shot
×	×	×	GPT-4.1	0.724	0.731	0.989	0.995
			GPT-4o-11	0.708	0.731	0.997	0.995
			Sonnet 4	0.694	0.738	0.895	1.000
			Opus 3	0.686	0.708	1.000	1.000
			Opus 4.1	0.602	0.799	0.648	1.000
			Haiku 3.5	0.148	0.666	0.259	1.000
			Sonnet 4.5	0.055	0.751	0.027	1.000
×	×	✓	Opus 4.1 (w/ ER)	**0.814**	**0.861**	0.873	1.000
			GPT-5	0.794	0.783	1.000	1.000
			Sonnet 4 (w/ ER)	0.780	0.776	0.992	1.000
			o3	0.769	0.762	1.000	1.000
			o4-mini	0.672	0.673	1.000	1.000
			Sonnet 4.5 (w/ ER)	0.077	0.828	0.040	0.999
✓	×	×	Qwen2.5-72B	0.707	0.716	1.000	1.000
			Llama-3.3-70B	0.678	0.620	1.000	0.999
			Llama-3.1-70B	0.674	0.672	1.000	1.000
			Gemma-3-27B	0.652	0.627	1.000	1.000
			Gemma-3-12B	0.637	0.585	1.000	1.000
			Qwen2.5-32B	0.622	0.645	0.999	1.000
✓	×	✓	GPT-OSS-120B	0.632	0.617	0.998	1.000
			GPT-OSS-20B	0.601	0.590	1.000	0.994
✓	✓	×	Swallow-3.1-70B	0.698	0.698	1.000	1.000
			Swallow-3.3-70B	0.692	0.731	1.000	1.000
			ABEJA-V2-32B	0.687	0.694	0.992	1.000
			LLM-jp-3.1-13B	0.582	0.244	1.000	0.202
			LLM-jp-3.1-8x13B	0.495	0.320	1.000	0.836

表 1 0-shot・4-shot の各モデルに対する F1 値および回答率スコア．O はオープンモデルかどうか，J は日本語特化モデルかどうか，R は推論モデルかどうかを表す．モデルは，O，J，R の値が同じグループ内で，0-shot 設定における F1 値の高い順にソート済み．太字は前モデル中の最高値を示す．

表 1 は 0-shot と 4-shot 別に F1 値と回答率スコアの結果を示す．全体的に，商用モデルはオープンモデルより良い性能を示している．商用モデルの F1 値は，Haiku-3.5 および Sonnet-4.5 を除けば，0.602 から 0.861 の間に分布している．一方で，オープンモデルでは，0.320 から 0.731 と，商用モデルに比べて全体的に低い．Haiku-3.5 および Sonnet-4.5 の低い F1 値はその低い回答率スコアに由来すると考えられる．大部分のモデルは指示に従っており，0.990 より高い回答率スコアをあげている．しかし，Haiku-3.5 および Sonnet-4.5 の回答率スコアは 0.259 と 0.027 である．実際に，Haiku-3.5 および Sonnet-4.5 は正誤判定出力の後に理由を出力してしまう場合があり，結果的に指定の回答形式違反となったため，低い回答率と低い F1 値となった．

文脈内学習の有効性に関して，商用モデルは 4-shot においてより良い性能を示した．特に Haiku-3.5 と Sonnet-4.5 でこの傾向は顕著であり，4-shot では 0.500 以上に向上する．一方，オープンモデルは 13 モデル中 8 モデルで，むしろ 0-shot でより良い性能を示した．例えば，Llama-3.3 は 4-shot では 0.620 の F1 値を示すが, 0-shot から 0.058 低下している．本実験で確認できる範囲では，文脈内学習は商用モデルにとって有用だった一方，オープンウェイトモデルにとっては妨げとなった．

推論モデルと非推論モデル間を比較すると，商用モデルにおいては，推論モデルがより高い性能を示した．推論モデルの F1 値は，0-shot で 0.077 から 0.814，4-shot で 0.673 から 0.861 の範囲に分布している．一方，非推論モデルの F1 値は，0-shot で 0.055 から 0.724，4-shot で 0.666 から 0.799 となり，全体的に推論モデルの方が優れている．しかし，オープンモデルにおいては，推論による性能向上は商用モデルほど顕著ではない．オープン推論モデルの GPT-OSS 系モデルは，オープン非推論モデルを必ずしも上回らない結果となった．

日本語特化モデルの有効性は，0-shot においてわずかに確認できた．元の Llama-3.1 と 3.3 と比べて，日本語特化させた Swallow-3.1 と 3.3 は約 0.03 高い F1 値を示す．同様に，ABEJA-V2 はそのベースモデルである Qwen2.5-32B を 0.065 と上回る．4-shot 設定では，日本語特化がもたらす性能向上は Llama-3.3 において最も顕著であった．Swallow-3.3 は Llama-3.3 を 0.111 上回る性能を示すが，他のモデルにおける改善幅はいずれも 0.050 未満に留まった．

5 おわりに

本研究では大規模言語モデルの日本法ドメイン知識を評価するためのデータセット，JBE-QA を構築した．JBE-QA では，既存リソースが対象とする民法に加えて刑法と憲法を取り込むことで，より包括的な法律知識の評価を可能とした．また，計 26 の LLM に対して，0-shot と 4-shot の設定で評価実験を行い，現時点における大規模言語モデルの日本法ドメイン知識の質に関するベースラインを確立した．結果として，商用の推論モデルである Anthoropic Opus 4.1 (w/ ER) が最良の性能を示した．また，日本語特化モデルは few-shot において性能向上をもたらすことが確認できた．

謝辞

本研究は文部科学省「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」及び，JST さきがけ JPMJPR236B の支援を受けたものです．

参考文献

[1]Ha-Thanh Nguyen, et al. Attentive deep neural networks for legal document retrieval. **Artificial Intelligence and Law**, Vol. 32, No. 1, pp. 57–86, 2024.

[2]Yen Thi-Hai Vuong, et al. Sm-bert-cr: a deep learning approach for case law retrieval with supporting model. **Artificial Intelligence and Law**, Vol. 31, No. 3, pp. 601–628, 2023.

[3]Ha Thanh Nguyen, et al. Llms for legal reasoning: A unified framework and future perspectives. **Computer Law & Security Review**, Vol. 58, p. 106165, 2025.

[4]Neel Guha, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **Advances in Neural Information Processing Systems**, Vol. 36, pp. 44123–44279. Curran Associates, Inc., 2023.

[5]Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6974–6996, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[6]Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. JEC-QA: A legal-domain question answering dataset. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, pp. 9701–9708, 2020.

[7]Randy Goebel, et al. 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.

[8]Jungmin Choi, Jungo Kasai, and Keisuke Sakaguchi. Evaluating gpt in japanese bar examination: Insights and limitations. **Jxiv preprint**, December 2023.

[9]Hiroaki Yamada, et al. Japanese tort-case dataset for rationale-supported legal judgment prediction. **Artificial Intelligence and Law**, pp. 1–25, 2024.

[10]OpenAI. Gpt-4 technical report, 2023.

[11]Anthropic. Claude 3: Opus, sonnet, and haiku, mar 2024.

[12]Anthropic. Building with extended thinking, 2025.

[13]Qwen, et al. Qwen2.5 technical report. **arXiv preprint arXiv:2412.15115**, 2025.

[14]OpenAI. gpt-oss-120b & gpt-oss-20b model card. **arXiv preprint arXiv:2508.10925**, 2025.

[15]Gemma Team. Gemma 3 technical report. **arXiv preprint arXiv:2503.19786**, 2025.

[16]Meta. Introducing llama 3.1: Our most capable models to date, jul 2024.

[17]Meta. Llama 3.3 70b instruct: A multilingual instruction-tuned large language model, dec 2024.

[18]ABEJA. Abeja-qwen2.5-32b-japanese-v0.1: A japanese language model based on qwen2.5-32b-instruct, 2025.

[19]Kazuki Fujii, et al. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In **Proceedings of the First Conference on Language Modeling**, COLM, p. (to appear), University of Pennsylvania, USA, October 2024.

[20]Naoaki Okazaki, et al. Building a large japanese web corpus for large language models. In **Proceedings of the First Conference on Language Modeling**, COLM, p. (to appear), University of Pennsylvania, USA, October 2024.

[21]Youmi Ma, et al. Building instruction-tuning datasets from human-written instructions with open-weight large language models. **arXiv preprint arXiv:2503.23714**, 2025.

[22]National Institute of Informatics. Llm-jp 3.1 series: Open japanese language models, 2025.

A 科目別スコア

O	J	R	Model	0-shot		
				民法	憲法	刑法
×	×	×	GPT-4.1	0.722	0.739	0.719
			GPT-4o-11	0.719	0.720	0.668
			Haiku 3.5	0.060	0.238	0.282
			Opus 3	0.664	0.783	0.656
			Opus 4.1	0.617	0.733	0.405
			Sonnet 4	0.668	0.826	0.636
			Sonnet 4.5	0.028	0.184	0.011
×	×	✓	GPT-5	0.779	0.845	0.791
			Opus 4.1 (w/ ER)	0.816	0.905	0.715
			Sonnet 4 (w/ ER)	0.757	0.847	0.784
			Sonnet 4.5 (w/ ER)	0.016	0.284	0.045
			o3	0.757	0.831	0.746
			o4-mini	0.661	0.718	0.663
✓	×	×	Gemma-3-12B	0.632	0.676	0.613
			Gemma-3-27B	0.632	0.712	0.647
			Llama-3.1-70B	0.680	0.680	0.651
			Llama-3.3-70B	0.678	0.697	0.660
			Qwen2.5-32B	0.581	0.710	0.650
			Qwen2.5-72B	0.695	0.746	0.707
✓	×	✓	GPT-OSS-120B	0.612	0.701	0.627
			GPT-OSS-20B	0.590	0.655	0.583
✓	✓	×	ABEJA-V2-32B	0.671	0.748	0.673
			LLM-jp-3.1-13B	0.577	0.619	0.563
			LLM-jp-3.1-8x13B	0.452	0.595	0.513
			Swallow-3.1-70B	0.700	0.711	0.681
			Swallow-3.3-70B	0.699	0.698	0.666

表 2 0-shot 各モデルに対する科目別 F1 値.

O	J	R	Model	4-shot		
				民法	憲法	刑法
×	×	×	GPT-4.1	0.729	0.758	0.715
			GPT-4o-11	0.740	0.750	0.687
			Haiku 3.5	0.645	0.724	0.675
			Opus 3	0.698	0.790	0.658
			Opus 4.1	0.795	0.876	0.741
			Sonnet 4	0.712	0.861	0.692
			Sonnet 4.5	0.733	0.872	0.685
×	×	✓	GPT-5	0.767	0.846	0.771
			Opus 4.1 (w/ ER)	0.853	0.918	0.831
			Sonnet 4 (w/ ER)	0.762	0.835	0.766
			Sonnet 4.5 (w/ ER)	0.821	0.889	0.796
			o3	0.731	0.843	0.776
			o4-mini	0.648	0.740	0.686
✓	×	×	Gemma-3-12B	0.561	0.675	0.560
			Gemma-3-27B	0.599	0.698	0.633
			Llama-3.1-70B	0.667	0.710	0.651
			Llama-3.3-70B	0.565	0.735	0.650
			Qwen2.5-32B	0.608	0.727	0.667
			Qwen2.5-72B	0.707	0.750	0.710
✓	×	✓	GPT-OSS-120B	0.611	0.666	0.594
			GPT-OSS-20B	0.569	0.636	0.606
✓	✓	×	ABEJA-V2-32B	0.673	0.773	0.682
			LLM-jp-3.1-13B	0.096	0.471	0.343
			LLM-jp-3.1-8x13B	0.254	0.512	0.307
			Swallow-3.1-70B	0.677	0.751	0.710
			Swallow-3.3-70B	0.725	0.772	0.711

表 3 4-shot の各モデルに対する科目別 F1 値.

B ベースライン実験使用モデル

B.1 商用モデル

- OpenAI GPT-4.1: `gpt-4.1-2025-04-14`
- OpenAI GPT-4o: `gpt-4o-2024-11-20`
- OpenAI GPT-5: `gpt-5-2025-08-07`
- OpenAI o3: `o3-2025-04-16`
- OpenAI o4-mini: `o4-mini-2025-04-16`
- Claude Opus-3: `claude-3-opus-20240229`
- Claude Opus-4.1: `claude-opus-4-1-20250805`
- Claude Sonnet-4: `claude-sonnet-4-20250514`
- Claude Sonnet-4.5: `claude-sonnet-4-5-20250929`
- Claude Haiku-3.5: `claude-3-5-haiku-20241022`

B.2 オープンモデル

- Alibaba Qwen 2.5-32B: `Qwen2.5-32B-Instruct`
- Alibaba Qwen 2.5-72B: `Qwen2.5-72B-Instruct`
- OpenAI GPT-OSS-20B: `gpt-oss-20b`
- OpenAI GPT-OSS-120B: `gpt-oss-120b`
- Google Gemma 3-12B: `gemma-3-12b-it`
- Google Gemma 3-27B: `gemma-3-27b-it`
- Meta Llama 3.1: `Llama-3.1-70B-Instruct`
- Meta Llama 3.3: `Llama-3.3-70B-Instruct`
- ABEJA-V2: `ABEJA-Qwen2.5-32b-Japanese-v0.1`
- Swallow-3.1: `Llama-3.1-Swallow-70B-Instruct-v0.3`
- Swallow-3.3: `Llama-3.3-Swallow-70B-Instruct-v0.4`
- LLM-jp-3.1-13B: `llm-jp-3.1-13b-instruct4`
- LLM-jp-3.1-8x13B: `llm-jp-3.1-8x13b-instruct4`