大規模言語モデルの法廷通訳への導入可能性の検証

山岸聖子[1] 神藤駿介[1] 宮尾祐介[1,2]
[1] 東京大学 [2] 国立情報学研究所大規模言語モデル研究開発センター
{shoko,skando,yusuke}@is.s.u-tokyo.ac.jp

概要

外国人被告人が公正な裁判を受ける権利を保障するためには法廷通訳人の存在が必要不可欠であるが、その人数は年々減少傾向にあり、また、国家資格が存在しないため通訳の質の保証の問題も抱えている。本研究では、大規模言語モデルを用いた法廷通訳並びにチェック・インタープリターの実現可能性を検証する。法廷通訳能力の検証のためのデータセットの作成および評価基準の策定を行い、3 種類の機械翻訳システムの検証を行った。実験の結果、特に GPT-4o による訳文は法的等価性を高い水準で維持できる一方で、法律用語や疑問文の訳出エラー、日本の司法制度に関する知識の不足といった課題もあることが分かった。

1 はじめに

外国人の裁判を受ける権利を保障し [1, 2, 3]、適正な裁判を実現するためには法廷通訳人の存在が必要不可欠である。法廷通訳人になるためには通訳人候補者名簿に登録されることが必要だが、登録通訳人の人数は平成 30 年時点では 3,788 人であったところ、令和 5 年時点では 3,208 人と減少傾向にある [4]。また、国家資格が存在しないため通訳人の能力が十分に担保されているとは言い難く、通訳内容を検証する制度も確立されていないため、通訳内容の正確さも担保されていない状況である [5, 6]。実際に、誤訳が判決に影響を与えたケース [7] や、誤訳による冤罪事件も発生している [8]。このように、今の体制は**人数確保**、**質の保証**の問題を抱えている。

これらの問題点につき、大規模言語モデル（LLM）が法廷通訳人の役割を担えれば、法廷通訳人の人数不足を補うことができる。また、LLM がチェック・インタープリター（法廷通訳人の通訳内容を検証する立場の通訳人 [9]）の役割を担えれば、誤訳による裁判結果のミスリードや後日の紛争を防ぐことができ、通訳の質の保証に繋がる。

本研究では LLM による法廷通訳、並びにチェック・インタープリターの実現可能性を検証する。法廷通訳においては「法的等価性」の維持、すなわち、日本の法律概念における用語の意味や効果を別の言語に正確に反映させることが重視される。その実現手段として、日本の裁判所は逐語訳を採用している [10]。逐語訳は通訳者の解釈や補足を加えて訳出することが許されない点で一般的な通訳と大きく異なる。また、法律用語や付加疑問文の多用など、法廷特有の言い回しが存在するため訳出そのものの難易度も高い。本研究ではこれらの特色を勘案し、法定通訳に特化した評価基準を定めた。また、法廷通訳能力を検証するデータセット（法廷での具体的なやり取りや、訳出難易度の高い疑問文の用例からなる）を新たに構築し、これらを用いて検証を行った。

実験の結果、GPT-4o による訳出文は通訳内容の法的等価性を高い水準で維持できる一方、法律用語や疑問文の訳出エラー、日本の司法制度に関する知識の不足などといった課題も見つかった。これらの課題を解決することで、将来的に LLM が法廷での通訳機能を担う可能性も十分あると考えられる。

2 データセット構築

本研究では、機械翻訳システムの法廷通訳能力を検証するための二種類のデータセットを構築した。一方は「法廷通訳ハンドブック [10]」から抽出した対訳付きデータセットである[1)]。当書籍には法廷での模擬的なやり取りの具体例の日本語とその対訳が掲載されている。もう一方は、法廷で頻繁に使用される疑問文（否定疑問文や付加疑問文など）の用例からなるデータセットである。疑問文は法廷通訳人にとって訳しにくいものとなっており検証対象とすべきだが [11, 12, 13]、法廷通訳ハンドブックには数例しか掲載されていないため、LLM を用いて別途生成した。作成したデータセットは公開する[2)]。

1） 最高裁判所広報課の許可を得て使用
2） https://github.com/mynlp/court_interpreter

表 1　法定通訳の評価のために策定した評価基準

評価項目	評価	評価基準
省略	2 段階 (0/1)	原文にある情報が欠落しているか否か（単語レベルで評価） 原文：勾留される期間は、原則として 10 日間です。 訳文：The period of detention is 10 days.（「原則として」が欠落）
付加	2 段階 (0/1)	原文にない情報が付加されているか否か（単語レベルで評価） 原文：申し訳ないことをしたと思います。 訳文： I think I have done something wrong, and I deeply regret it.（下線部が付加されている）
単語の意味	2 段階 (0/1)	省略や付加がされていない各単語やフレーズが、それぞれ同じ意味で訳出されているか 原文：刃長 10cm のナイフ 訳文：10cm knife（「全長 10cm のナイフ」に意味が変わっている）
流暢性	5 段階 (1-5)	文法的に正しく、自然な言葉で表現され、読みやすい文章であるか 訳出内容の正しさは考慮せず、純粋に文自体の流暢性を評価
疑問文の訳出	2 段階 (0/1)	付加・修辞・否定疑問文のニュアンスが適切に訳出されているか 原文：法廷での宣誓を理解していますね？ 訳文：您明白在法庭上的宣誓吗？（「ね？」は「吗？」ではなく「吧？」と訳すべき）

2.1　「法廷通訳ハンドブック」からの抽出

中国語版、ベトナム語版、英語版からデータの抽出を行った。まず書籍に対して OCR 処理を行い、章や節のタイトルを除去し発言内容だけを抽出した。次に原文（日本語）と対訳で文単位の対応を取るために、原文を基準として各対訳に対して複数文を 1 文にまとめる・1 文を複数文に分ける作業を適宜行なった。この際、内容の改変を避けるため、必要最小限の修正（接続詞の省略・追加など）に留めた。最後に重複する文と簡単な文を取り除き、各言語 245 の文からなるデータセットを構築した。

2.2　法廷特有の疑問文の用例生成

GPT-4o に付録 A.1 に示すプロンプトを与えて付加・修辞・否定疑問文の例を作成し、重複する用例を取り除き、最終的に 50 例作成した。実際に生成されたデータの例を表 2 に示す。

2.3　機械翻訳文の生成

GPT-4o, llama-3.1-70b-versatile, Azure AI Translator の 3 つのシステムを用いて機械翻訳文を生成した（以下それぞれ “GPT”,“Llama”, “Azure”）。GPT および Llama については付録 A.2 に示すプロンプトを使用して訳出を行なった。法廷通訳ハンドブックから抽出したデータセット（2.1 節）には対訳が含まれているため、上記 3 システムの出力と合わせて、各日本語文に対し計 4 つの翻訳文が存在する。疑問文のデータセット（2.2 節）には日本語の原文しか無いため、上記 3 システムの機械翻訳文のみが存在する。

表 2　法廷通訳における疑問文の生成例

種類	文例
修辞疑問文	もし被告人が本当に有罪なら、なぜこれほど多くの疑問点が残るのでしょうか？
否定疑問文	被告人はその現場にいなかったのではありませんか？
付加疑問文	被告人の証言は信じがたいですよね？

3　法廷通訳のための評価基準の策定

本研究では、日本の裁判所の逐語訳を必須とする立場 [10] に則して評価基準を策定する。策定した評価基準を表 1 に示す。逐語訳されているか否かの判断基準として「省略」「付加」「単語の意味」の 3 つの基準を策定した。これに加えて、「逐語訳した結果、流暢性が失われていないか」を検証するために「流暢性」の基準も策定した。疑問文のデータセットに対しては「疑問文の訳出」も評価基準に追加した。逐語訳に特化して設定したこれらの評価基準のスコアが高い場合に「通訳内容の法的等価性が保たれている」と解することができ、LLM による法廷通訳の実現可能性が高いと言える。

4　実験設定

4.1　LLM による法廷通訳の実現可能性の検証

LLM による法廷通訳の実現可能性を検証するため、2 節で作成したデータセットを用いて 3 節で策定した評価基準をもとに人手評価を行った。

2.1 節で説明した法廷通訳ハンドブックのデータセットについては、中国語・ベトナム語・英語の各

対訳、および日本語原文の機械翻訳文（3 文）の合計 4 文に対する評価を行った。評価者の能力を加味して、中国語の訳文は日本語を原文とし、ベトナム語の訳文は英語の対訳を原文とし、英語の訳文は中国語の対訳を原文として評価を行った。原文を日本語に揃えた評価は今後の課題とする。表 1 の基準のうち、「疑問文の訳出」を除く 4 つの基準について評価を行った。2.2 節で説明した疑問文の用例データについては中国語でのみ評価を行った。これは、日本語を原文として評価できる評価者を中国語でしか確保できなかったことによる。機械翻訳文 3 文を提示し、表 1 の 5 つの基準全てについて評価を行った。

また、人手評価に加え、汎用的な翻訳評価指標である COMET [14] による評価も行った。人手評価の設定に合わせ、原文と各訳出文のみを提示する reference-free 設定で評価を行った。

4.2 LLM によるチェック・インタープリターの実現可能性の検証

1 節で述べたように、法廷においては法廷通訳に加え、訳文の質の保証（チェック・インタープリター）も重要である。そこで、LLM-as-a-judge [15] によって表 1 の各指標のスコア付けを行うことで、訳文評価の自動化の可能性を検証する。付録 A.3 に示すプロンプトを GPT-4o に与え、各訳出文の人手評価基準のスコアを算出し、実際の人手評価スコアとの相関を見て分析を行う。

5 結果

5.1 LLM に法廷通訳は可能か

表 3 に各評価項目の平均スコアを示す。人手評価の結果を見ると、中国語・英語・中国語（疑問文）は GPT、ベトナム語はハンドブックの対訳文のスコアが高いことが分かった。一方、各評価指標ごとの結果（図 1）を見ると、疑問文の訳出においては GPT よりも Azure の方が高評価であり、複数の翻訳システムを組み合わせることの有用性が示唆された。また、いずれの設定においても流暢性のスコアは高く、単語の意味のスコアは低い傾向にある。なお、ベトナム語と英語の評価結果は全体的に低いが、これは評価時の原文が日本語ではないことによると考えられる。

以降、人手評価（法廷通訳の指標）と COMET のスコア（汎用的な翻訳評価指標）を比較することで法廷通訳の特徴を明らかにし、法廷通訳特有の難しさについて議論する。表 3 から、両者の評価結果の傾向は似ているものの、全体的に COMET の評価結果が高くなっており、一般的な翻訳としてはどの機械翻訳も高い性能を持っていることが分かった。以下、人手評価スコアが低い例について、中国語の GPT による訳出文に絞ってエラー分析を行う。

表 3　ハンドブックに記載された対訳文と機械翻訳文の各評価項目の平均スコア。流暢性は 5 段階のスコアを 0-1 にスケーリングした。LLM-as-a-Judge は LLM-J と略記。

	評価方法	対訳	GPT	Llama	Azure
日本語 ↓ 中国語	人手評価	0.77	**0.82**	0.61	0.52
	LLM-J	0.89	**0.96**	0.75	0.63
	COMET	0.84	**0.87**	0.80	0.77
英語 ↓ ベトナム語	人手評価	**0.45**	0.35	0.28	0.29
	LLM-J	**0.70**	0.66	0.44	0.48
	COMET	**0.76**	0.74	0.69	0.68
中国語 ↓ 英語	人手評価	0.35	**0.46**	0.38	0.35
	LLM-J	0.65	**0.90**	0.76	0.68
	COMET	0.80	**0.87**	0.86	0.84
中国語（疑問文）	人手評価	-	**0.84**	0.76	0.75
	LLM-J	-	**0.95**	0.81	0.85
	COMET	-	**0.94**	0.92	0.92

ハンドブックのデータのエラー分析　人手評価スコアが 0.6 未満である 24 文を対象に分析を行う。COMET は 21 文に対して 0.8 以上の高い評価をしており、一般的な通訳としては正しいと判断されても、法廷通訳として適切とは限らないことが分かる。人手評価においては、特に「付加」と「単語の意味」のスコアが低い（それぞれ 0.14, 0.05）。付加の例としては、原文にはない単語が補足的に付加されているケースが多く、結果として文全体の意味が原文から乖離するケースもあった[3)]。単語の意味が異なる例としては法律用語の誤訳[4)]が多く、「10 日経つ前に釈放されることもある」という原文が「10 日前に釈放された」と訳されているなど、日本の司法制度の仕組みに関わる誤訳もあった。LLM に法廷通訳人の役割を担わせる場合は、単語の付加を避け、法律用語や日本の司法制度に関する知識を学習させることが重要であると考えられる。

疑問文データのエラー分析　「疑問文の訳出」の人手評価が 0 であった 14 文を対象に分析を行ったところ、否定疑問文と付加疑問文の語尾のニュアンスを正確に訳出できない傾向があった。例えば

3)　「これからあなたを勾留するかどうかを決めるために」という文に「継続」という単語が付与され、「継続して勾留するかどうかを決めるために」という意味の文になっていた。
4)　「論告」を「控诉意见陈述」に誤訳。正しくは「论罪求刑」。

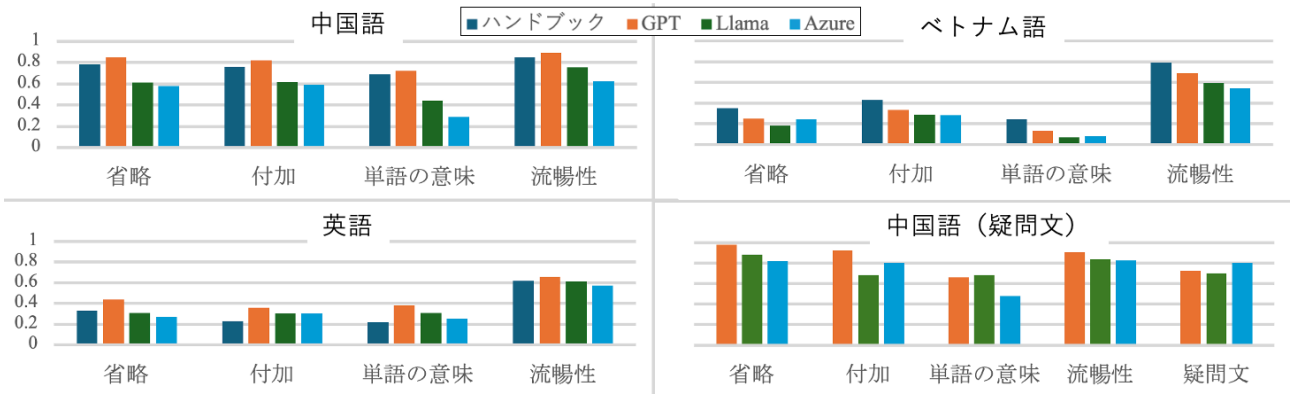

図 1　各評価項目ごとの人手評価スコア

表 4　中国語の人手評価結果と LLM-as-a-Judge の混同行列。各行列の左上から時計回りに FN, TP, FP, TN を表す。

省略	
9	248
5	33

付加	
2	244
4	45

単語の意味	
9	201
14	71

疑問文	
3	33
3	11

「ではないですか？」といった否定疑問文や「ですよね？／ですね？」といった付加疑問文が、一般的な疑問文「ですか？」と訳出されていた。また、過去の状況を説明する文が現在形に訳されているケースや、単語の意味を誤って訳しているケースもあった。否定疑問文や付加疑問文は一般的な疑問文に比べ文全体の訳出難易度が上がり、その結果、時制や単語の意味の訳出に影響を与えている可能性がある。単語の意味の正確性、並びに疑問文のニュアンスの訳出の精度を高めることで、法廷通訳人の役割を LLM が担える可能性が高まるだろう。

5.2　LLM にチェック・インタープリターは可能か

表 3 から、人手評価と LLM-as-a-Judge の評価結果の全体的な傾向は似ていることが分かる。一方、表 4 から分かるように LLM-as-a-Judge は False Positive が多く、特に「単語の意味」「疑問文」においてその傾向が顕著である。法廷で使用頻度の高い法律用語や疑問文の評価の精度を高めることで、LLM がチェック・インタープリターとしての役割を担うことが見込める。

6　関連研究

法廷通訳で最も重要なことは「発話者の意図を正確に伝えること」である。そのためには、どのように通訳すべきか（逐語訳か意訳か）、訳された内容の正確性をどのように担保すべきか（通訳の質の保証の仕組み）が重要であり、この点については翻訳・通訳分野で多くの先行研究で取り上げられている。

「逐語訳か意訳か」という点について、日本の裁判所は逐語訳を必須とする「導管モデル [16]」を採用している。「導管モデル」は法廷通訳人の役割を起点言語（日本語）での発言を目標言語（外国語）に逐語的に訳出する行為にあるとする立場であり、省略や編集、余分な要素の付加は許されず、話し手の言語レベルや語調、ニュアンスまでそのまま訳すことが必要とされている [17]。このような裁判所のあり方に対し、逐語訳は文化や慣習の違いを反映できず、外国人が法廷でのやり取りを正しく理解出来ないとして批判する意見もある [18, 19]。

このような「導管モデル」の問題点は、LLM が法廷で通訳を行う際にも大きな課題となるが、現状では法廷において、文化、慣習、法制度の違いを加味して通訳することは推奨されておらず [20]、また、通訳内容に反映させるべき「違い」の選定には時間をかけて慎重に検討する必要性があるため、この点については今後の研究課題とする。

7　おわりに

本研究では、LLM の法廷通訳への導入可能性を検証するためにデータセットの作成および評価指標の策定を行った。検証の結果、一定の条件下では LLM の訳出内容が法的等価性を維持していることを示した。法律用語や司法制度を学習させ、単語の意味の正確性・疑問文のニュアンスの訳出の精度を高めることで、将来的には法廷通訳の役割も担える可能性が十分にある。実運用に向けては、今後は音声入力に対する訳出精度の検証が必要である。中国では 2019 年に「法廷尋問同時通訳システム」が運用開始されており [21]、日本でもこのような通訳システムが導入されることが望ましい。

謝辞

本研究は、文部科学省補助事業「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」の支援を受けたものです。訳文の人手評価に協力頂いた Wei Yu 氏、Anh Tu Tran 氏、Wu Minchao 氏に感謝の意を表します。

参考文献

[1] 日本国憲法第 32 条.
[2] 国際人権 b 規約第 14 条 3 項.
[3] 刑事訴訟法第 175 条.
[4] 裁判所. ごぞんじですか 法廷通訳 -あなたも法廷通訳を-. 裁判所, 2024.
[5] 明木茂夫. 法廷通訳における二人合議制について – その翻訳論的考察 – . pp. 1–12.
[6] 石田美智代. 法廷通訳に求められる正確性と現場での実践. 静岡大学教育研究, Vol. 11, pp. 175–183, 2015.
[7] 裁判員裁判で通訳ミス多数　専門家鑑定　長文は６割以上. http://www.asahi.com/special/080201/OSK201003210091.html.
[8] 捜査で誤訳、冤罪生む　司法通訳の質向上急務　タガログ語やりとり、誤ったまま証拠に. https://www.nikkei.com/article/DGKKZO83941320X01C24A0CE0000/.
[9] 児玉晃一. 裁判員裁判とチェック・インタープリターについて. **LIBRA**, Vol. 9, p. 28, 2009.
[10] 最高裁判所事務総局刑事局監修. 法廷通訳ハンドブック実践編【中国語】【英語】【ベトナム語】改訂版. 法曹会, 2010.
[11] 水野かほる. 法廷通訳における訳出上の課題について – 否定疑問文を対象とした通訳調査からの考察 – . 通訳翻訳研究, Vol. 16, pp. 63–84, 2016.
[12] 水野かほる, 津田守. 裁判員裁判時代の法廷通訳人. 大阪大学出版会, 2016.
[13] 高畑幸, 坂巻静佳, 森直香, 水野かほる. 2022 法廷通訳の仕事に関する調査報告書. Technical report, 静岡県立大学法廷通訳研究会, 2023.
[14] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.
[15] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A survey on llm-as-a-judge, 2025.
[16] 吉田理加. 法廷通訳と言語イデオロギー. 通訳翻訳研究, Vol. 12, pp. 31–50, 2012.
[17] 水野かほる. 外国人事件における司法通訳の正確性 ―要通訳事件の事例からの考察―. 言語政策, Vol. 4, pp. 1–24, 2008.
[18] 水野真木子, 中村幸子, 吉田理加, 河原清志. 日本の司法通訳研究の流れ―方法論を中心に. 通訳翻訳研究, Vol. 12, pp. 133–154, 2012.
[19] 毛利雅子. 司法通訳人の役割　―法廷通訳における言語等価性との関連において―. 日本大学大学院総合社会情報研究科紀要, Vol. 8, pp. 315–323, 2007.
[20] 灘光洋子. 法廷通訳人が直面する問題点 -文化的差異をどう捉えるか-. 異文化コミュニケーション研究, Vol. 13, pp. 59–82, 2001.
[21] 中国初の法廷尋問同時通訳システムが上海金融裁判所で運用スタート. https://spc.jst.go.jp/news/191202/topic_2_01.html.

A　プロンプト

A.1　法廷特有の問文の用例生成のためのプロンプト

System Prompt:
あなたは沢山の刑事裁判を担当した経験のある裁判官/弁護人/検察官です。
User Prompt:
裁判中に裁判官/弁護人/検察官がよく使用する修辞/否定/付加疑問文の文章を 10 文考えてください。

A.2　LLM による翻訳のためのプロンプト

System Prompt:
あなたは法廷通訳を行う通訳士です。法廷におけるやり取りを正確に翻訳してください。
User Prompt:
以下の日本語の文を中国語（簡体字）/ベトナム語/英語に翻訳してください。回答は必ず一行になるようにして下さい。
＜日本語の文＞

A.3　LLM-as-a-judge のプロンプト

「法廷通訳ハンドブック」に対するプロンプトを示す。各評価基準の説明は省略した（表 1 参照）。

System Prompt:
あなたは法廷通訳を行う通訳士です。＜翻訳元言語＞の文と、それを＜翻訳先言語＞に翻訳した 4 つの文章 A,B,C,D を与えるので、それぞれの翻訳性能を評価してください。評価は以下の四つの基準に従って行ってください。
1. 省略 (Omission)
翻訳された文中において、原文にある情報が欠落しているかどうかを評価します。単語レベルで訳の省略があるか否かについて評価を行って下さい。
評価基準:（略）
2. 付加 (Addition)
翻訳された文中において、原文にない情報が付加されているかどうかを評価します。単語レベルでの訳の付加があるか否かについて評価を行って下さい。
評価基準:（略）
3. 単語の意味 (Word Meaning)
省略や付加がされていない各単語やフレーズについて、それぞれ適切に翻訳されているか、意味が変わっていないかを評価します。ただし、文全体として誤訳になっている場合も「0」と評価して下さい。
評価基準:（略）
4. 流暢性 (Fluency)
翻訳された文が文法的および自然な言葉で表現され、読みやすいかどうかを評価します。ただし、翻訳内容の正しさは考慮せず、純粋に文自体の流暢性を評価して下さい。他の言語の文字が含まれている場合は、流暢性が低いとみなして減点して下さい。
評価基準:（略）
回答は以下のフォーマットに従って一行で出力してください
id, A の omission スコア,A の addition スコア,A の word_meaning スコア,A の fluency スコア, ..., D の fluency スコア
例：handbook001,1,0,1,5,1,1,1,4,0,0,1,5,1,1,1,5
User Prompt:
id: id
＜翻訳元言語＞の文: ＜翻訳元の文＞
＜翻訳先言語＞の文 A: ＜翻訳先の文 A ＞
...
＜翻訳先言語＞の文 D: ＜翻訳先の文 D ＞