専門家評価と相関分析に基づく大規模言語モデルによる法廷通訳の実現可能性の検証

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が不可欠であるが，近年は通訳人の人数減少や質の保証が十分でない点が課題となっている．本研究は，大規模言語モデル（LLM）が法廷通訳および検証通訳として機能し得るかを検証する．先行研究で策定した評価指標をプロの法廷通訳人の知見に基づいて改訂し，人手評価も各言語の法廷通訳人に依頼した．さらに，LLM-as-a-Judge による評価結果との相関分析やエラー分析を取り入れ，より精緻な分析を行った．その結果，GPT は他の LLM と比較して良好な性能を示した一方，法律用語や疑問文の訳出，司法制度に関する知識不足といった課題が明確になった．

1　はじめに

外国人が日本で裁判を受ける際，その権利保障 [1, 2, 3] は司法手続の公正性に不可欠であり，法廷通訳は重要な役割を担っている．しかし，日本の法廷通訳制度は，**通訳人の人数不足**，**質の保証の不十分さ**という課題を抱えている．通訳人候補者数は近年減少しており [4]，国家資格や通訳内容を検証する制度も整備されていない [5, 6]．その結果，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]．また，通訳結果を事後的に検証する「検証通訳 [9]」も重要だが，専門家による人手検証には高いコストが伴う．近年は大規模言語モデル（LLM）の発展により機械翻訳の品質は向上しているが，逐語性や言語等価性を重視する法廷通訳にそのまま適用できるかは明らかではない．

本研究は，LLM による「法廷通訳」と「検証通訳」の実現可能性を検証することを目的とする．具体的には，法廷通訳に求められる言語等価性を LLM がどの程度維持できるか，また LLM が検証通訳としてプロの法廷通訳人による人手評価を再現できるかを検討する．そのために，法廷通訳経験者の知見に基づいて評価指標を改訂し，ベトナム語・中国語・英語の三言語について法廷通訳経験者による人手評価を行った．その結果，LLM は法廷通訳として一定水準で言語等価性を維持した翻訳文を生成できる可能性を示した一方で，法律用語や疑問文の訳出に課題が残り，また検証通訳としては LLM-as-a-Judge と人手評価との相関が低く，現段階では実用が困難であることが示唆された．

2　関連研究

法廷通訳では発話意図を忠実に伝えることが最重要とされ，通訳方針や訳出内容の正確性をいかに確保するかが長年議論されてきた．日本の裁判所は，通訳人の解釈や補足を認めない逐語訳原則 [10] の「導管モデル」[11] を採用し，司法の公平性を確保している [12]．一方で，逐語訳中心の運用は文化的・語用的差異を反映できず，当事者の理解を妨げる可能性があるとの批判もある [13, 14]．これらの理論的課題に加え，法廷通訳の質を制度的に保証することも重要である．しかし日本では公的資格認定制度がなく [15, 16, 17]，改善策が提案されてきたものの [5, 18]，十分な制度改革には至っていない．

法廷通訳人の不足や通訳の質の検証の不十分さを解決するため，AI による自動化を目指す研究も進められている．山岸ら [19] は，逐語訳の原則に基づき，LLM が法廷通訳および検証通訳として機能し得るかを検証した．しかし同研究では評価指標の設計や人手評価の過程に法廷通訳人が関与していないという課題があり，本研究でその解決を図る．

3　検証手順

LLM による法廷通訳および検証通訳の実現可能性を検証するため，法廷通訳場面を想定した日本語

文からなるデータセットを作成し，ベトナム語・中国語・英語への翻訳文を作成した．次に，逐語性を反映した評価指標を策定し，各言語の法廷通訳人による人手評価を行った．さらに，同一指標を用いて LLM-as-a-Judge [20] による評価を実施し，人手評価結果との比較を通じて，検証通訳としての LLM の有効性を検討した．

3.1 データセット構築

本研究では，既存研究 [19] と同様に「法廷通訳ハンドブック」[10] 由来の対訳データと法廷特有の疑問文データの二種を用いた．疑問文は用例を補うため ChatGPT で自動生成した．データセット構築の詳細は前回発表した論文を参照されたい[1)]．既存研究では法廷通訳経験の無い一般の多言語話者が評価したが，法廷通訳人は非常に多忙であることから，同じ量の文を評価することは現実的ではない．本研究では法廷通訳人による人手評価のため文を精選し，発話者や意図が明確で法廷特有の定型表現を含む文に限定した結果，ハンドブックは 45 文，疑問文は 25 文を評価対象とした．

次に，機械翻訳システムを用いて外国語訳（ベトナム語，中国語，英語）を生成した．使用したシステムは `GPT-4o`（以下 GPT），`llama-3.3-70b-versatile`（以下 Llama），`Azure AI Translator`（以下 Azure）の 3 種である．訳出のためのプロンプトは先行研究 [19] から若干修正した（付録 A）．法廷通訳ハンドブックのデータセットでは，日本語原文 1 文に対して既存対訳 1 文と機械翻訳文 3 文を合わせた計 4 文が存在する．疑問文データセットには日本語原文のみが存在するため，機械翻訳文 3 文のみ存在する．

3.2 法廷通訳のための評価指標の策定

山岸ら [19] は，各翻訳文に対して人手評価を行うため，日本の裁判所の逐語訳を原則とする立場 [10] に則した評価指標を策定した．評価項目には，逐語訳遵守の「省略」「付加」，意味の正確性を評価する「単語の意味」，翻訳文の自然さを評価する「流暢性」，および疑問文に対して「疑問文の訳出」がある．本研究では，法廷通訳人の意見を参考に「単語の意味」を 2 段階評価から 5 段階評価に，「疑問文の訳出」を 2 段階評価から 3 段階評価に変更した．これにより，概ね適切だが改善の余地がある翻訳文や，疑問文のニュアンスの違いをより細かく区別できる．改訂した評価指標を付録 B に示す．

3.3 LLM による法廷通訳と検証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，3.2 節の評価指標に基づき人手評価を実施した．既存研究 [19] では一般の多言語話者が評価を行っていたのに対し，本研究では法廷通訳経験の豊富な通訳人に依頼した点が特徴である．検証通訳については，LLM-as-a-Judge [20] により各指標を自動評価し，その有効性を人手評価との混同行列および Spearman 順位相関係数によって定量化する．

4 結果 1: LLM に法廷通訳は可能か

LLM が法廷通訳をどの程度代替し得るのかを検討するために実施した人手評価の結果を概観する．結果を表 1 に示す．各言語の平均スコアの有意差をブートストラップ法によって検定したところ，ハンドブックデータセットにおいては全ての言語において GPT と既存対訳が有意水準 0.05 で最も高い翻訳品質を示した．一方，ベトナム語においては既存対訳の評価スコアが GPT および Llama よりも有意に低いという結果が得られた．このことは，公式の既存対訳が必ずしも高品質であるとは限らず，対訳コーパス自体に品質のばらつきが存在する可能性を示唆している．したがって，LLM 翻訳の評価や実務利用を検討する際には，既存対訳を絶対的な基準とみなすのではなく，対訳コーパスの質の向上および継続的な更新・精査が必要であると考えられる．

疑問文データセットにおいては言語ごとに異なる傾向が確認された．有意に GPT が高いスコアを示したのはベトナム語のみであった．中国語では GPT が Azure よりも有意に高いスコアを示した一方で，GPT と Llama の間には有意差は認められなかった．また，英語では GPT が Llama よりも有意に高いスコアを示した一方で，GPT と Azure の間には有意差は見られなかった．これらの結果は，疑問文の訳出においては GPT が言語に依らず最良の選択肢とは限らないことを示している．

5 結果 2: LLM に検証通訳は可能か

5.1 スコアと相関係数の分析

表 1 に LLM-as-a-Judge（以下 LLM-J）の評価結果を示す．人手評価において有意に優れている翻訳

1) データセットはこちらで公開している: https://github.com/mynlp/court_interpreter

表 1: 人手評価および LLM-as-a-Judge の評価結果．5 段階の指標（単語の意味，流暢性）および 3 段階の指標（疑問文の訳出）のスコアは 0 から 1 に正規化した．

	ハンドブックデータセット								疑問文データセット					
	人手評価				LLM-as-a-Judge				人手評価			LLM-as-a-Judge		
	既存対訳	GPT	Llama	Azure	既存対訳	GPT	Llama	Azure	GPT	Llama	Azure	GPT	Llama	Azure
ベトナム語														
省略	0.29	**0.76**	0.62	0.53	0.82	**0.98**	0.91	0.91	**0.88**	0.84	**0.88**	**1.00**	**1.00**	**1.00**
付加	0.44	**0.96**	0.73	0.73	0.71	**0.93**	0.89	0.84	**0.92**	0.80	0.84	**1.00**	0.96	**1.00**
単語の意味	0.48	**0.58**	0.38	0.35	0.61	**0.88**	0.69	0.71	**0.71**	0.59	0.59	**0.90**	0.75	0.88
疑問文の訳出	-	-	-	-	-	-	-	-	**0.60**	0.34	0.38	**0.94**	0.86	**0.94**
流暢性	**0.80**	0.78	0.66	0.65	0.78	**0.93**	0.80	0.86	**0.82**	0.65	0.77	0.95	0.86	**0.96**
平均	0.50	**0.77**	0.60	0.57	0.73	**0.93**	0.82	0.83	**0.79**	0.64	0.69	**0.96**	0.89	**0.96**
中国語														
省略	0.76	**0.78**	0.76	0.60	**1.00**	**1.00**	0.98	0.91	0.84	**0.88**	0.76	**1.00**	**1.00**	**1.00**
付加	0.71	**0.84**	0.64	0.58	0.82	**0.89**	0.87	0.64	**1.00**	0.72	0.76	**1.00**	**1.00**	**1.00**
単語の意味	**0.73**	0.72	0.46	0.28	0.82	**0.96**	0.81	0.66	0.72	**0.78**	0.57	**0.91**	0.89	0.86
疑問文の訳出	-	-	-	-	-	-	-	-	**0.76**	0.62	0.50	**1.00**	0.94	0.90
流暢性	**0.86**	0.83	0.72	0.57	0.88	**0.97**	0.86	0.81	**0.87**	0.80	0.77	**0.98**	0.94	0.96
平均	0.77	**0.79**	0.65	0.51	0.88	**0.96**	0.88	0.76	**0.84**	0.76	0.67	**0.98**	0.95	0.94
英語														
省略	0.93	**1.00**	0.89	0.96	0.91	**1.00**	0.98	0.96	**1.00**	**1.00**	0.96	**1.00**	0.96	**1.00**
付加	0.96	**1.00**	**1.00**	**1.00**	0.69	0.93	0.93	**0.96**	**0.96**	0.92	**0.96**	**1.00**	0.96	0.96
単語の意味	**0.83**	0.73	0.57	0.59	0.68	**0.95**	0.83	0.78	**0.82**	0.71	0.76	**0.95**	0.82	0.94
疑問文の訳出	-	-	-	-	-	-	-	-	**0.76**	0.58	0.68	**0.98**	0.86	0.94
流暢性	**0.82**	0.72	0.56	0.58	0.87	**0.96**	0.90	0.91	**0.84**	0.81	0.81	0.98	0.94	**0.99**
平均	**0.89**	0.86	0.76	0.78	0.79	**0.96**	0.91	0.90	**0.88**	0.80	0.83	**0.98**	0.91	0.97

表 2: 人手評価と LLM-J との間の Spearman 順位相関係数．上はハンドブックデータセットの結果を，下は疑問文データセットの結果を示す．*, ** はそれぞれ $p < 0.05, 0.01$ を表す．LLM-J が全ての文で「正しい」と判断したケースは “-” で示す．

	ベトナム語	中国語	英語
省略	0.2043**	0.0461	-0.0488
付加	0.2744**	0.1922**	0.2998**
単語の意味	**0.4931****	0.3478**	0.2216**
流暢性	0.2397**	0.2679**	0.1563*

	ベトナム語	中国語	英語
省略	-	-	-0.0135
付加	0.2804*	-	-0.0393
単語の意味	0.3275**	0.2394*	0.1561
流暢性	0.3517**	0.3219**	0.2153
疑問文の訳出	0.2842*	0.3369**	0.2825*

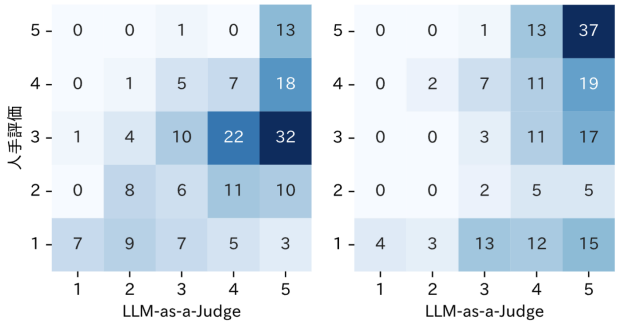
が LLM-J においても同様に評価されるか，という観点で分析を行う．

まずベトナム語について，ハンドブックデータセットでは人手評価・LLM-J のいずれにおいても GPT の翻訳が最も高く評価されており，両者の結果は一致している．このことは，少なくとも当該データセットにおいて，LLM-J が人手評価者の判断基準を適切に反映している可能性を示唆する．中国語および英語については，人手評価と LLM-J の結果の間に差異が見られた．いずれの言語においても，ハンドブックデータセットでは人手評価において GPT と既存対訳との間に有意差は認められなかったにもかかわらず，LLM-J による評価では GPT が既存対訳よりも有意に高く評価される結果となった．この差異は，LLM-J が GPT による訳文を相対的に高く評価しやすい傾向，すなわち自己生成文に対するバイアスの可能性を示唆している．なお，疑問文データセットにおいてはいずれの言語も GPT が最も良いという結果は得られなかった．これは疑問文データセットにおける LLM-J の過剰肯定傾向に起因すると考えられる．すなわち，いずれのモデルもスコアが高く出るがためにスコア差も現れづらく，有意差も生じづらいと考えられる．

表 2 に相関係数の値を示す．明確な相関（0.4 以上）が確認されたのは，ハンドブックデータセットのベトナム語「単語の意味」のみであった．その他は弱い相関にとどまり，相関がほぼ認められないケースもあった．相関が見られる場合においても，混同行列を見ると LLM-J は全般的に過剰肯定する傾向が強い（図 1）．このため，LLM-J は検証通訳の手段として信頼できるとは言い難い．

表 3: ハンドブック（ベトナム語）において人手評価と LLM-J との間で結果に乖離が見られた事例.

対象文	原文 / 訳文
エラー例 1 (Llama)	原文：しかし，場合によっては，10 日たつ前に釈放されることもありますし，更に 10 日以内の日数勾留が延長されることもあります. 訳文：Tuy nhiên, trong một số trường hợp, họ có thể được phóng thích trước khi 10 ngày trôi qua, hoặc họ cũng có thể bị gia hạn giam giữ thêm 10 ngày nữa.
エラー例 2 (GPT)	原文：あなたは，〇〇国国民として，領事関係に関するウィーン条約第 36 条第 1 項 (b) の規定により，勾留の事実を〇〇国領事官に通報することを要求しますか 訳文：Anh/chị có yêu cầu thông báo cho viên chức lãnh sự của nước 〇〇 về việc bị giam giữ theo quy định tại Điều 36 khoản 1(b) của Công ước Vienna về quan hệ lãnh sự không?
エラー例 3 (Llama)	原文：その場合には，明日から 14 日以内に〇〇高等裁判所<最高裁判所>あての控訴<上告>申立書をこの裁判所に差し出してください. 訳文：Trong trường hợp đó, ông cần nộp đơn kháng cáo đến Tòa án nhân dân tối cao trong vòng 14 ngày kể từ ngày mai tại tòa án này.
エラー例 4 (Llama)	原文：証拠品があなたのものであると確認したのではないですか？ 訳文：Anh/chị có xác nhận rằng vật chứng là của anh/chị không?



(a) ベトナム語「単語の意味」 (b) 中国語「単語の意味」

図 1: ハンドブックにおいて，人手評価と LLM-J の間で比較的高い相関が見られた例の混同行列.

5.2 エラー分析

人手評価と LLM-J の評価結果の乖離が顕著なケースについて分析を行う．具体的には，ハンドブックデータセットにおけるベトナム語の「省略」，および疑問文データセットにおけるベトナム語の「疑問文の訳出」を取り上げる．前者は人手評価において「省略あり（誤り）」となっている 81 件のうち，LLM-J は「省略なし（正しい）」と判断した例が 68 件ある．後者は人手評価において最低スコア「1」となっている 27 件のうち，LLM-J が最高スコア「3」を付けた例が 20 件ある．

表 3 に実例を示す．エラー例 1〜3 は「省略」の例であり，エラー例 4 は「疑問文の訳出」の例である．**エラー例 1** は「10 日以内」を「10 日」として解釈しており，勾留延長制度を正確に伝えていない．また，三人称複数「họ」の使用は法廷で被告人に直接説明する文脈では語用論的に不適切である．**エラー例 2** は「〇〇国国民として」という国籍要件が欠落しており，通知請求権の法的根拠を不明確にしている．LLM-J が権利主体や制度的前提条件の欠落を十分に検出できない可能性を示す例である．**エラー例 3** では，審級構造のうち「高等裁判所」に関する情報が欠落しており，司法制度の階層構造を歪める意味的誤りとなっている．また，疑問文においては質問の方向性・発話意図が変質してしまう事例が多く，**エラー例 4** のように「〜したのではないですか？」という否定的推量を含む追及的質問が，「〜しますか？」と肯定的推量に転換される例が見られた．

以上の結果から，LLM-J は人手評価と共通する傾向はあるものの，相関は高いとは言えず，過剰肯定の傾向が強いことが分かる．特に情報保持（省略）や語用的機能（疑問文の訳出）の精度は十分でなく，現時点で LLM-J を検証通訳として用いることは困難である．

6 終わりに

本研究では，LLM による法廷通訳の実現可能性を検証するため，専用データセットを構築し，改訂した評価指標に基づいてプロの法廷通訳人による人手評価および LLM-as-a-Judge による検証通訳の有効性を検証した．その結果，現行 LLM は語義の忠実性や語用的機能を安定的に満たさず，LLM-as-a-Judge は人手評価との相関が限定的で，情報保持や語用的判断に課題があることが明らかになった．今後は，法廷通訳特有の用語や語用論的要素の反映，対象言語・データセットの拡充，評価手法の改善が求められる．

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．本研究では，最高裁判所広報課の許可を得て「法廷通訳ハンドブック」の記載内容を使用しました．また，本研究の評価作業にご協力頂いた三名の通訳人の方々に深く感謝致します．皆様の専門的知見とご助力が，本研究の進行に不可欠でした．

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 – . 文化科学研究, 13 巻 1 号, pp. 1–12, 2001.
[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] 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.
[12] 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.
[13] 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.
[14] 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.
[15] 毛利雅子. 法廷通訳翻訳における言語等価性維持の可能性. 丸善プラネット, 2022.
[16] 水野かほる. 近年の司法通訳をめぐる状況と課題. 国際関係・比較文化研究, Vol. 11, No. 1, pp. 21–36, 2012.
[17] 渡辺修, 長尾ひろみ. 外国人と刑事手続. 成文堂, 1998.
[18] 水野真木子. 司法通訳資格認定制度の可能性について. 有斐閣, 1995.
[19] 山岸聖子, 神藤駿介, 宮尾祐介. 大規模言語モデルの法廷通訳への導入可能性の検証. 言語処理学会第 31 回年次大会予稿集, pp. 1351–1356, 3 2025.
[20] 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.

A　LLM による機械翻訳のプロンプト

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．意訳は避け，単語の省略や付加をしない逐語訳で訳すようにしてください．否定疑問文，付加疑問文，修辞疑問文はニュアンスが変わらないように注意して訳してください．回答は必ず一行になるようにして下さい．
＜日本語の文＞

B　策定した評価指標

評価項目	評価基準
省略 (2 段階)	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） ＜エラー例＞ 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)	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） ＜エラー例＞ 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 (5 段階)	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 1. 間違っている，文全体として誤訳となっている（意味が正しく伝わらない） 2. 原文とは異なる意味に解釈される可能性があり不適切（意味の拡大，縮小） 3. 正しく意味が伝わる可能性が高いが，より適切な単語を使うべきである 4. 正しく意味が伝わる可能性が高いが，自分ならこのようには訳さない（間違ってはいない） 5. 問題を感じない（自分が訳す時もこのように訳す） ＜スコアが 1 である例＞ 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)	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 1. 著しく不自然で理解しにくい 2. 不自然であり，全体的にぎこちない 3. 理解は可能だが，どこか不自然 4. ほぼ自然である 5. 非常に自然で流暢である
疑問文の訳出 (3 段階)	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 1. 間違っている（発言者の意図が正しく伝わらない） 2. 正しく意味が伝わる可能性が高いが，より適切な表現がある 3. 問題を感じない（自分が訳す時もこのように訳す） ＜スコアが 1 である例＞ 原文：法廷での宣誓を理解していますね？ 訳文：你明白在法庭上的宣誓 吗？（「ね？」は「吗？」ではなく「吧？」と訳すべき）