中世・近世日本語資料の現代語機械翻訳：評価用対訳データセットの構築と LLM の性能評価

東山 翔平 [1]　大内 啓樹 [2]　橋本 雄太 [3]　藤田 篤 [1]
[1] 情報通信研究機構　[2] 奈良先端科学技術大学院大学　[3] 国立歴史民俗博物館
{shohei.higashiyama,atsushi.fujita}@nict.go.jp
hiroki.ouchi@is.naist.jp yhashimoto@rekihaku.ac.jp

概要

近代以前の日本の歴史資料は大量に存在するものの，多くの現代人にとって，資料に書かれた内容を読み解くことは容易ではない．本研究では，古典語テキストの読解支援の技術として，古典日本語から現代日本語への機械翻訳に取り組む．本言語対での機械翻訳研究を推進するため，中世・近世資料からなる評価用対訳データセットの構築を行い，さらに同データセットを用いた商用・オープン LLM の翻訳精度自動評価を行い，モデル別・ドメイン別の性能差を確認した．

1　はじめに

近代以前に筆記・刊行された日本の歴史資料は，膨大な量が現存している．国書データベース [1] などのデータアーカイブ事業によって，数十万点規模の資料のデジタル画像化が進められており，クラウドソーシング翻刻 [2] や自動翻刻技術 [3, 4] の普及・発展に伴い，資料画像のテキスト化も進んでいる．このように，多くの資料について画像・文字情報へのアクセスが容易になってきた．しかし，資料に記された内容を読み解き，過去の時代の「知識」にアクセスするには，歴史的背景や古典日本語の専門知識がないと難しい．専門知識を持たない読み手でも資料の内容を理解できるようにし，知識へのアクセス性を高めるためには，古典語テキストの読解支援の技術が重要となる．本研究では，その中でも，古典語から現代語への機械翻訳（「現代語機械翻訳」と呼ぶ）に焦点を当てる．

現代語機械翻訳における大きな課題として，(i) 学習資源の不足と (ii) 評価資源の不足がある．学習資源とは，翻訳モデルの学習に必要なデータ，特に大規模な対訳テキスト・単言語テキストを指す．評価資源の中核は，翻訳モデルの性能評価に利用可能な対訳テキストと，深層学習指標の学習済みモデルである．2 種類の資源 (i)，(ii) とも，機械翻訳技術の発展のために欠かせないものであるが，特に，(ii) の評価資源の不足は，古典日本語の現代語機械翻訳およびその評価の研究の不足に直結していたと言える．過去に，日本語の現代語機械翻訳に取り組んだ先行研究はいくつかあるものの [5, 6, 7]，非公開の対訳テキストをモデル学習・評価に用いたものであり，後続の研究者がそれらの再現や発展のための研究を行うことは困難であった．

こうした状況を打開すべく取り組まれた研究として，橋本ら [8] および東山ら [9] の研究を挙げる．橋本ら [8] は，中世・近世古文書資料を対象に，人手による現代語訳作成作業と，商用 LLM 現代語訳の品質評価作業を行い，両作業で作成したデータを公開した．東山ら [9] は，橋本らの人手評価結果を用いて，既存の機械翻訳評価指標のメタ評価実験を行い，BLEU [10] や BLEURT [11] といった評価指標の有用性を確認した．また，同研究では，橋本らが用いた資料とオープンな近世古典籍資料等を機械翻訳評価データセットとして整備し，商用・オープン LLM の現代語訳性能の自動評価を行った．これらの研究で報告された評価結果の知見および構築されたデータセットは，課題 (ii) を緩和し，後続の研究を推進する役割を果たすと期待される．

本研究では，上述の先行研究の流れを汲み，日本語の現代語機械翻訳研究をさらに推進することをねらいとする．具体的には，(1) 東山ら [9] のデータセットを拡張し，収録資料のドメインを拡大するとともに，(2) ドメインごとの精度差を明らかにすべく既存 LLM の詳細な性能評価に取り組む．実験結果から，(a) 漢文テキストが多い古文書・古記録ドメインではどの LLM も翻訳精度が低い傾向にあるこ

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

と，(b) Claude 4.5，次いで Gemini 2.5 が，ドメイン横断的に高い翻訳精度を示し，オープン LLM の精度を大きく上回っていること，(c) このような精度差が生じていることは，翻訳結果実例の品質を考慮しても納得しやすい結果であることを確認した．構築したデータセット JHPT V2 は公開予定である[1)]．

2 対訳データセット

著者らの研究 [9] では，歴博中世文書，千葉県史，信州地域史料アーカイブ，福井県文書館の 4 種類のデータソースから，中世・近世日本語資料対訳データセット JHPT V1 を作成した．本研究では，「江戸料理レシピデータセット」[12][2)] を利用して対訳テキストを作成し，JHPT V1 に加えることで，JHPT V2 を作成した．データの作成過程を以降で説明する．

JHPT V1 の構築　V1 データのサイズは，データソース別では表 1 の上 4 行に示す内訳，ドメイン（時代・分野）別では表 2 の上 4 行に示す内訳となっている．つまり，時代別では，文字数において中世資料の約 8%に対して近世資料が約 92%を占めており，形態別では，1 枚物の古文書・古記録が約 56%，冊子体の古典籍が約 64%となっている．また，形態に加えて内容でも分類すると，近世古典籍の文学・地誌が約 49%と最大の割合を占めている．V1 のデータ構築は，対象資料の選定，HTML や PDF ファイルからのテキスト抽出，セグメント単位（翻刻・校訂時に付加された句点や読点の位置で区切った単位）での分割と原文・現代語訳セグメントの対応付け，ルビの分離などのテキスト加工等という工程により構築された．

JHPT V2 の構築　V2 では，表 1（表 2）の第 5 行に相当するデータを追加した．同データは，「江戸料理レシピデータセット」として公開されている近世古典籍『万宝料理秘密箱』の原文（翻刻文）・現代語訳テキストのうち，『巻之二』に相当する部分を利用して作成したものである．データの実例を表 4（第 7・8 行）に示す．この原データを採用した理由は，再頒布可能なライセンスで原文・現代語訳が公開されている貴重な資源であったこと，V1 にない料理という新たなジャンルのテキストであったことの 2 点である．構築の工程は V1 と概ね同様である．テキスト加工については，原文の概ね文末に相当する位置までで原文・現代語訳をそれぞれ 1 セグメントにまとめあげた．また，現代語訳が，レシピ特有の手順番号表現（例：「①を一度こす」）を含む場合は，機械翻訳評価に利用しやすい原文に忠実な訳となるよう，個々の原文に応じて，丸付き数字＋格助詞の個所を削除する，「その」「右の」など原文通りの表現に置き換えるといった編集を加えた．

表 1　JHPT のデータサイズ（#Doc：文書数，#Seg：セグメント数，#Char：原文文字数）

	データソース	#Doc	#Seg	#Char
V1	歴博中世文書	12	84	1,828
	千葉県史	22	183	6,830
	福井県文書	7	29	776
	信州地域史料	10	374	17,199
V2	江戸料理レシピ	1	56	2,941
	合計	53	726	29,574

表 2　JHPT のドメイン別データサイズ

	ドメイン	#Doc	#Seg	#Char
V1	中世古文書・古記録	16	96	2,038
	近世古文書・古記録	25	205	7,521
	近世古典籍-記録等	3	96	4,004
	近世古典籍-文学・地誌	7	273	13,070
V2	近世古典籍-料理	1	56	2,941

3 評価実験

JHPT V2 を用いて，商用・オープン LLM の翻訳精度自動評価を行う．特に，各 LLM の全体的な翻訳精度を評価した文献 [9] の実験との相違点として，本研究では，(1) ドメインごとの翻訳精度の違いの評価，(2) より高性能と考えられる追加の商用 LLM の評価を行う．

自動評価指標　文献 [9] で有効性が確認された評価指標のうち，BLEU および BLEURT を使用する．BLEU は，機械翻訳文 H 中の単語 n-gram が参照訳 R に含まれる割合を基にしたスコアを計算する表層ベースの指標である．BLEURT は，多言語の機械翻訳人手評価データで微調整されたマスク言語モデルを用いて，原文 S および H，R の 3 つ組に対して 0〜1 のスコアを予測する回帰型の指標である．BLEU の計算には SacreBLEU [13] を使用し[3)]，BLEURT 学習済みモデルには BLEURT-20 [14] を用いた．

1） https://github.com/shigashiyama/jhpt
2） https://codh.rois.ac.jp/edo-cooking/
3） Signature：“nrefs:1|case:mixed|eff:yes|tok:ja-mecab-0.996-IPA|smooth:exp|version:2.5.1”.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

表 3　ドメイン別の翻訳精度（オープン LLM，商用 LLM における最高スコアをそれぞれ**太字**で表示）

システム	中世古文書・古記録		近世古文書・古記録		近世古典籍（文学・地誌）		近世古典籍（記録等）		近世古典籍（料理）	
	BLEU	BLEURT	BLEU	BLEURT	BLEU	BLEURT	BLEU	BLEURT	BLEU	BLEURT
原文	8.6	43.8	4.0	40.4	13.2	47.2	7.6	45.5	7.7	35.1
Sarashina2.2-3b-instruct-v0.1	12.7	58.6	12.8	59.2	20.1	63.4	14.4	63.6	15.7	61.6
Llama-3-ELYZA-JP-8B	15.2	53.1	12.1	53.8	22.3	53.2	20.4	61.9	17.5	52.7
Gemma-2-Llama-Swallow-9b-it-v0.1	16.0	57.3	15.6	58.4	24.3	63.4	19.2	65.8	20.1	61.3
llm-jp-3.1-13b-instruct4	17.6	52.3	15.5	56.0	**26.9**	62.0	17.4	62.3	19.4	56.9
Gemma-2-Llama-Swallow-27b-it-v0.1	**19.5**	**60.7**	**16.9**	59.2	26.6	65.5	**20.6**	**67.1**	**22.3**	58.6
DeepSeek-R1 (thinking: off)	15.6	60.6	16.6	**60.8**	26.4	**66.6**	18.3	66.6	21.4	**64.3**
Claude Opus 4.5	**27.2**	**65.8**	**23.0**	**62.8**	**38.4**	**70.4**	**27.5**	**68.4**	**26.5**	**65.4**
Claude Sonnet 4.5	23.6	65.2	22.0	61.7	36.9	69.0	25.4	68.6	24.3	63.2
Gemini 2.5 Flash (thinking_budget: 0)	21.2	62.8	20.2	61.1	33.0	67.7	21.6	67.6	23.5	64.7
Gemini 3 Flash (thinking_level: minimal)	18.8	63.6	19.2	64.3	29.5	69.0	17.4	67.0	19.4	65.1
Gemini 3 Flash (thinking_level: medium)	19.4	64.3	19.3	64.6	30.1	70.0	17.9	68.5	19.8	65.2
GPT-5 (reasoning effort: minimal)	15.8	59.4	18.0	60.8	24.5	65.5	18.6	66.7	21.5	61.9
GPT-5.2 (reasoning effort: none)	17.6	59.3	18.1	60.1	29.1	66.7	21.8	66.1	21.3	63.6
GPT-5.2 (reasoning effort: medium)	18.6	61.3	19.5	61.2	29.6	68.0	22.3	66.8	23.1	64.0

評価対象 LLM　表 3 に掲載する 6 種類の商用 LLM と，6 つのオープン LLM を評価する．商用 LLM については，文献 [9] で評価されたモデルの他，より高い翻訳精度が期待される Claude Opus 4.5, Gemini 3 Flash，GPT-5.2 の 3 モデルを加えた．オープン LLM については，文献 [9] で評価されたモデルのうち翻訳精度が比較的高かったものを対象とした．各 LLM では，文献 [9] と同じく，原文に忠実な現代語訳を要求する単純なプロンプトを使用し，temperature=0.2（GPT-5/5.2 以外）とした．また，「思考」機能を利用できるモデルでは，thinking なしまたは最小限（DeepSeek-R1，Gemini），reasoning effort 最小（GPT-5/5.2）を基本の設定としつつ，思考レベルを上げた設定も一部評価した．各 LLM について，2 回の推論結果の精度の平均を報告する[4)]．

4　実験結果

4.1　ドメイン別翻訳精度

JHPT V2 の各ドメインにおける各 LLM の翻訳精度を表 3 に示す．

まず，LLM の全体的な精度をドメイン別に見ると，中世・近世とも「古文書・古記録」での精度が低い傾向が見られる．これは，存在するデジタルデータのテキスト量自体が少なく，当該ドメインのごく少量のテキストでしか各 LLM が事前学習を行っていないと想定されることや，同時に，当該ドメインに多い漢文形式のテキストを「読む」能力が各 LLM に乏しいことなどが理由と考えられる．参考に，機械翻訳文の代わりに原文を入力した際のスコア（「原文」行）を算出したところ，「近世古文書・古記録」では BLEU スコアが最も低く，原文と参照訳の表層的な乖離が大きいことがわかる．一方，「近世古典籍-文学・地誌」では BLEU スコアが最も高い．

LLM ごとの精度については，各商用 LLM が 5 ドメイン・2 指標を通じて，大半のオープン LLM よりも高いスコアを示した．特に，Claude の最高性能モデル Opus 4.5 は全モデルの中で最高精度を示し，Sonnet 4.5 も上回った．GPT-5.2（思考最小）では，GPT-5 と比べて BLEU を中心に各ドメインでのスコアが向上した．Gemini 3 Flash（思考最小）では，2.5 Flash と比べて 5 ドメインで BLUE スコアが低下したが 4 ドメインで BLEURT スコアが向上し，どちらのモデルが良いか一概に判断しがたい．GPT-5.2, Gemini 3 では，思考レベルを上げ推論コストを費やすことで精度向上したが，元々の翻訳精度が高い Claude モデルとの差を埋めるには至っていない．

オープン LLM については，DeepSeek-R1 や Gemma-2-Llama-Swallow-27b が各ドメインで比較的高い精度を示し，これらは GPT-5/5.2 と近い水準のスコアを達成しているといえる．しかし，Gemini 2.5 や Claude の 2 モデルとはスコアに開きがある．次節の事例分析でも示すように，これら高性能な商用 LLM には実際の品質としても劣っていると考えられる．

4）ただし，GPT-5.2（reasoning effort: medium）については，API レート制限に達したことから 1 回のみの実行となった．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

表 4　原文，参照訳および LLM 出力事例．赤字は誤訳，青字の下線部は他の誤りで，[Ommision] は原文の語句の訳し漏れ，[Addition] は過剰生成，[Dependency] は当該語句に関する係り受け構造の誤りを意味する．

原文	此段前文申上候通り楫痛殊ニ汐行悪敷候故、是迄乗参御番所乗越候段不調法可申上様無御座候、
参照訳	これは前に申し上げた通り楫が故障して、そのうえ潮流が悪かったので今までは寄港していた浦賀番所を素通りしてしまったもので、不始末で申し訳ありません。
Claude Opus 4.5	この件は前文で申し上げました通り、舵が痛んでおり、特に潮の流れが悪かったため、これまで乗ってきて御番所を乗り越えてしまいましたことは不調法であり、申し上げようもございません。
Gemini 3 Flash (thinking_level: medium)	この件については、以前申し上げました通り、舵の傷みに加えて、とりわけ潮の流れが悪かったため、ここまで航行してくる中で [Addition] [Omission: 乗参]番所を通り過ぎてしまったことは、不手際であり、弁解の余地もございません。
DeepSeek-R1 (thinking: off)	このたびは、先に申し上げましたように、櫓の痛みが特にひどく、潮の流れも悪いため、これまで通りに船を出して [Addition] [Omission: 乗参]番所まで参上することができず、まことに不都合でございますが、申し上げるべき方法もございません。
Gemma-2-Llama-Swallow-27b	この段の前文に申し上げました通り、楫の痛みと潮の流れが悪いため、これまで御番所に乗参して [Dependency]乗り越えてきたことは、不都合であると申し上げても差し支えございません。
原文	是は 大麥を よくよばして ずいぶん 水にて能あらひ 小茶碗に すこし入レ たまごに昆布のだしに すこし 醤油と やき塩とを入レ よきかげんして 右茶碗にかけて むすべし
参照訳	これは、大麦をよくふやかし、水で丁寧に洗い、小茶碗に少し入れる。卵に昆布出汁と少量の 醤油、焼塩を入れ、丁度よい味加減にし、右の茶碗にかけて、蒸す。
Claude Opus 4.5	これは大麦をよく水に浸して [Addition]十分にふやかし、水でよく洗い、小さな茶碗に少し入れ、卵に昆布の出汁と少しの醤油と焼き塩を入れ、よい加減に調味して、その茶碗にかけて蒸すこと。
Gemini 3 Flash (thinking_level: medium)	これは、大麦をよくふやかして、水で十分に洗い、小さな茶碗に少し入れます。卵に昆布のだし、少々の醤油と焼き塩を加えて味を調え、それを先ほどの茶碗に注いで蒸してください。
DeepSeek-R1 (thinking: off)	これは、大麦をよく洗って、十分に水でよくすすぎ、小さい茶碗に少し入れ、卵と昆布のだし汁に少し醤油と焼き塩を加えて適当な味加減にし、先ほどの茶碗にかけて蒸らすのである。
Gemma-2-Llama-Swallow-27b	これは、大麦を良く洗い、たっぷりの [Addition]水でよく洗い、小さな茶碗に少し入れ、卵に昆布のだし汁に少し醤油と焼き塩を加えて、良い加減に煮て、その茶碗にかけ、混ぜ合わせます。

4.2　翻訳結果事例

JHPT の原文・参照訳の実例と，LLM 4 モデルの出力事例を表 4 に示す．

一つ目の事例（千葉県史）では，原文の「是迄乗参御番所乗越候段」の個所を中心に，各モデルで正確さを欠く訳となっている．ただし，Claude Opus と Gemini 3 は，「御番所を通り過ぎたことの非を認め謝罪する」中心的な文意を伝えている．他の 2 モデルは，「不調法」[5] を「不都合」と訳しているとみられ，謝罪の正確なニュアンスを示せていない．

二つ目の事例（江戸料理レシピ）では，原文の「よばして」の個所および調理の工程全体について，Claude Opus と Gemini 3 Flash は概ね適切に訳せている．他の 2 モデルでは，「よばして」を「洗って」や「洗い」に誤訳するなどの誤りを含んでいる．

以上のように，Claude と Gemini のモデルは比較的正確な訳を示した一方，DeepSeek-R1 と Gemma-2-Llama-Swallow は，オープン LLM の中では高い自動評価スコアながら，誤りが目立つ結果となった．

5　おわりに

古典日本語・現代日本語の対訳データセット JHPT の拡張と，既存 LLM のドメイン別性能評価を行った．今後，古典文学等にも焦点を当ててデータセットを更に拡大しつつ，オープン LLM の事後学習等による性能改善にも取り組む予定である．

5）『日本国語大辞典 第二版』では，「不調法」について「手ぎわが悪く、下手なこと」や「不始末」という語釈がある．後述の「よばす」については，麦を「煮る」「蒸す」の他，方言として「ふやかす」という語釈がある．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

謝辞

国立歴史民俗博物館より，同館所蔵資料および関連データを提供いただいた．

参考文献

[1] 国文学研究資料館. 国書データベース. https://kokusho.nijl.ac.jp/.

[2] 橋本雄太, 加納靖之, 一方井祐子, 小野英理. 『みんなで翻刻』の運用成果と参加動向の報告. じんもんこん 2020 論文集, pp. 39–46. 情報処理学会, 2020.

[3] カラーヌワット・タリン, 北本朝展. くずし字認識の進化とサービス化の展開. じんもんこん 2020 論文集, pp. 3–10. 情報処理学会, 2020.

[4] 青池亨. CPU 環境で高速に動作する軽量 OCR「NDL 古典籍 OCR-Lite」の開発. じんもんこん 2020 論文集, pp. 181–186. 情報処理学会, 2020.

[5] 星野翔, 宮尾祐介, 大橋駿介, 相澤彰子, 横野光. 対照コーパスを用いた古文の現代語機械翻訳. 言語処理学会第 20 回年次大会発表論文集, pp. 816–819. Association for Natural Language Processing, March 2014.

[6] Masashi Takaku, Tosho Hirasawa, Mamoru Komachi, and Kanako Komiya. Neural machine translation from historical Japanese to contemporary Japanese using diachronically domain-adapted word embeddings. In **Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation**, pp. 534–541, Hanoi, Vietnam, October 2020. Association for Computational Linguistics.

[7] Hisao Usui and Kanako Komiya. Translation from historical to contemporary Japanese using Japanese T5. In **Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages**, pp. 27–35, Tokyo, Japan, December 2023. Association for Computational Linguistics.

[8] 橋本雄太, 太田那優. 大規模言語モデルを利用した古文書資料の現代語訳の品質評価. 研究報告人文科学とコンピュータ, Vol. 2025-CH-138, No. 3, pp. 1–7, 2015.

[9] 東山翔平, 大内啓樹, 橋本雄太. 中世・近世日本語資料の現代語機械翻訳における自動評価指標の検証. じんもんこん 2025 論文集, Vol. 2025, pp. 315–322, 2025.

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[11] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics.

[12] 北本朝展, 山本和明. 人文学データのオープン化を開拓する超学際的データプラットフォームの構築. じんもんこん 2016 論文集, Vol. 2016, pp. 117–124, 2016.

[13] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[14] Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. Learning compact metrics for MT. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 751–762, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[15] SB Intuitions. Sarashina2.2-instruct：コンパクトかつ性能の高い日本語 instruct モデル, 2025. https://www.sbintuitions.co.jp/blog/entry/2025/03/07/093143.

[16] Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, and Akira Sasaki. elyza/Llama-3-ELYZA-JP-8B, 2024. https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B.

[17] Swallow LLM Team. Gemma-2-Llama Swallow, 2025. https://swallow-llm.github.io/gemma2-llama-swallow.ja.html.

[18] LLM-jp. LLM-jp-3.1 シリーズ instruct4 の公開, 2025. https://llm-jp.nii.ac.jp/ja/blog/blog-887/.

[19] DeepSeek-AI et al. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. 2025. arXiv:2501.12948.

[20] Anthropic. Introducing Claude Opus 4.5, 2025. https://www.anthropic.com/news/claude-opus-4-5.

[21] Anthropic. Introducing Claude Sonnet 4.5, 2025. https://www.anthropic.com/news/claude-sonnet-4-5.

[22] OpenAI. GPT-5 system card, 2025. https://cdn.openai.com/gpt-5-system-card.pdf.

[23] OpenAI. Update to GPT-5 system card: GPT-5.2, 2025. https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf.

[24] Google. A new era of intelligence with Gemini 3, 2025. https://blog.google/products-and-platforms/products/gemini/gemini-3/.

[25] Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. 2025. arXiv:2507.06261.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

表 5　表 4 の原文に対する参照訳および LLM 出力事例

原文	此段前文申上候通り楫痛殊ニ汐行悪敷候故、是迄乗参御番所乗越候段不調法可申上様無御座候、
参照訳	これは前に申し上げた通り楫が故障して、そのうえ潮流が悪かったので今までは寄港していた浦賀番所を素通りしてしまったもので、不始末で申し訳ありません。
Claude Sonnet 4.5	この区間は[Addition]前文で申し上げました通り、櫓が痛んでおり、特に潮の流れが悪いため、これまで乗り継いで御番所を通り過ぎてしまったことは、不手際としてお詫び申し上げる方法もございません。
GPT-5.2 (reasoning effort: medium)	この件は前に申し上げたとおり、舵が傷み、ことに潮の具合も悪かったため、これまで乗り進んで御番所を通り越してしまったことは不調法であり、申し上げておわびするほかございません。
DeepSeek-R1 (thinking: on)	このたび前文で申し上げましたように、舵の故障が特にひどく、潮の流れも悪い状況であるため、これまでに船を[Omission: 乗参]御番所まで乗り入れることができなかった事情について、不都合をお伝えする方法もございませんでした。
llm-jp-3.1-13b-instruct4	この段、前文で申し上げました通り、楫（かじ）の具合が特に悪く、潮の流れも非常に悪いため、これまで[Omission: 乗参]乗場の御番所を乗り越したことを無礼千万に[Dependency]申し上げるつもりは毛頭ございません。
sarashina2.2-3b-instruct-v0.1	この段、先ほど述べました通り、船酔いが激しく、また潮の流れが悪く、これまで[Omission: 乗参]予定されていた番所を乗り越すこととなりました。大変不届きな次第であり、申し訳なく思っております。
原文	是は 大麥を よくよばして ずいぶん 水にて能あらひ 小茶碗に すこし入レ たまごに昆布のだしに すこし 醤油と やき塩とを入レ よきかげんして 右茶碗にかけて むすべし
参照訳	これは、大麦をよくふやかし、水で丁寧に洗い、小茶碗に少し入れる。卵に昆布出汁と少量の醤油、焼塩を入れ、丁度よい味加減にし、右の茶碗にかけて、蒸す。
Claude Sonnet 4.5	これは大麦をよく選別して十分に水でよく洗い、小さな茶碗に少し入れ、卵に昆布の出汁と少しの醤油と焼き塩を入れて適度に加減し、右の茶碗にかけて蒸すのである。
GPT-5.2 (reasoning effort: medium)	これは、大麦をよくゆでて、たっぷりの[Addition]水でよく洗い、小さな茶碗に少し入れる。卵に昆布だしを加え、少ししょうゆと焼き塩も入れてほどよく味をととのえ、それをその茶碗に注いで、蒸しなさい。
DeepSeek-R1 (thinking: on)	これは大麦をよく水に浸して十分に洗い、小さい茶碗に少し入れます。卵に昆布の出汁を加え、少しの醤油と焼き塩を入れて適量に調節し、先ほどの茶碗にかけて蒸すとよい。
llm-jp-3.1-13b-instruct4	これは、大麦をよく水に浸し、よく洗ってから、小茶碗に少し入れ、卵と昆布のだしに少し醤油と焼き塩を入れて混ぜ、適当な加減にして、右の茶碗に入れて蒸しなさい。
sarashina2.2-3b-instruct-v0.1	これは、大麦をよく洗って、十分に水で洗い、小さな茶碗に少量入れ、卵に昆布だしを少し、醤油と焼き塩を加えてよく混ぜ合わせ、適切な味付けをして、右の茶碗にかけて蒸しなさい。

A.1　実験設定詳細

次の Hugging Face リポジトリ ID のオープン LLM を用いた：sbintuitions/Sarashina2.2-3b-instruct-v0.1 [15]，elyza/Llama-3-ELYZA-JP-8B [16]，tokyotech-llm/Gemma-2-Llama-Swallow-9b-it-v0.1，tokyotech-llm/Gemma-2-Llama-Swallow-27b-it-v0.1 [17]，llm-jp/llm-jp-3.1-13b-instruct4 [18]，deepseek-ai/DeepSeek-R1 [19]．また，次の商用 LLM を API 経由で利用した：claude-opus-4-5-20251101 [20]，claude-sonnet-4-5-20250929 [21]，gpt-5.2-2025-12-11 [22]，gpt-5-2025-08-07 [23]，gemini-3-flash-preview [24]，gemini-2.5-flash [25]．

各 LLM で，プロンプトは以下を用いた．

```
以下は日本の歴史資料の文章です。原文を現代日本語に翻訳した現代語訳を出力し、他には何も出力しないでください。現代日本語として自然な文字・語彙・表現・文法を用いて、原文の内容に忠実な訳としてください。英単語の使用を禁止します。
原文：{source_text}
現代語訳：
```

A.2　翻訳結果事例（追加 LLM 分）

表 4 の 2 事例についての他の LLM の翻訳結果を表 5 に示す．DeepSeek-R1 について，思考ありでの推論結果を示したが，思考なしでの推論結果（表 4）と比べて必ずしも誤りは減っていない．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).