

公表版

令和7年度  
特許情報の機械翻訳における  
大規模言語モデルの利用可能性に関する調査事業  
調査報告書

令和8年3月19日

一般財団法人 日本特許情報機構

## 目次

1. はじめに .....	1
1.1 調査の目的 .....	1
1.2 調査の概要 .....	1
1.3 調査の構成 .....	2
2. 公開情報調査(モデル) .....	3
2.1 調査概要 .....	3
2.2 調査範囲 .....	3
2.3 調査結果とモデル候補 .....	3
2.4 モデル候補の評価 .....	6
2.4.1 候補の絞り込みおよび優先順位付け .....	6
2.4.2 翻訳試行について .....	9
2.4.3 翻訳試行結果 .....	9
2.5 モデルの決定 .....	11
3. 公開情報調査(前後処理) .....	12
3.1 調査概要 .....	12
3.2 調査範囲 .....	12
3.3 調査結果と前後処理候補 .....	12
3.4 前後処理候補の評価 .....	13
3.4.1 候補の優先順位付け .....	13
3.4.2 優先度を考慮した候補の検討 .....	16
3.4.3 翻訳試行について .....	17
3.4.4 翻訳試行結果 .....	17
3.5 前後処理の決定 .....	19
3.5.1 モデルの選定 .....	19
3.5.2 前後処理の選定 .....	20
4. 公開情報調査(プロンプト) .....	21
4.1 調査概要 .....	21
4.2 調査範囲 .....	21
4.3 調査結果とプロンプト候補 .....	21
4.4 プロンプト候補の評価 .....	22
4.4.1 候補の優先順位付け .....	22
4.4.2 優先度を考慮した候補の検討 .....	23
4.4.3 翻訳試行について .....	25
4.4.4 翻訳試行結果 .....	26

4.5	プロンプトの決定	26
4.5.1	プロンプトを工夫した翻訳で使用するモデル	26
4.5.2	調査対象とするプロンプト	27
5.	公開情報調査(翻訳精度評価手法)	28
5.1	調査概要	28
5.2	調査範囲	28
5.3	調査結果と翻訳精度評価手法候補	28
5.4	翻訳精度評価手法候補の評価	29
5.4.1	候補の優先順位付け	29
5.4.2	優先度を考慮した候補の検討	31
5.4.3	評価試行について	32
5.4.4	評価試行結果	32
5.5	翻訳精度評価手法の決定	34
6.	対訳コーパス又は辞書の準備	36
6.1	特許公報の対訳コーパス	36
6.2	拒絶理由通知等の対訳コーパス	36
6.3	出願人名・発明者の日英対訳辞書	36
6.4	学習データ及び評価用データの作成	37
7.	各種モデルによる翻訳	38
7.1	大規模言語モデルによる翻訳	38
7.1.1	翻訳で使用するプロンプトについて	38
7.1.2	プロンプトの試行錯誤結果	38
7.1.3	パラメータ調整対象	39
7.1.3	パラメータ変化結果	40
7.2	プロンプトを工夫した翻訳	41
7.2.1	翻訳対象とする言語方向、項目について	41
7.3	前後処理を工夫した翻訳	46
7.3.1	翻訳対象とする言語方向、項目について	47
7.4	追加学習	48
7.5	追加学習済みモデルによる翻訳	49
7.6	ニューラル機械翻訳エンジンによる翻訳	49
8.	翻訳精度の評価	50
8.1	自動評価	50
8.1.1	英日	51
8.1.2	中日	53
8.1.3	韓日	54

8.1.4	日英(発明の名称、要約、請求項、明細書)	55
8.1.5	日英(拒絶理由、意見書、補正書、出願人名/発明者名)	56
8.1.6	項目別	57
8.1.7	大規模言語モデル 3 モデル及び NMT モデルの比較	59
8.1.8	前後処理を導入したモデル	64
8.1.9	プロンプトに工夫を施したモデル	69
8.1.10	特許由来の学習データを追加学習させたモデル	75
8.2	人手評価	79
8.2.1	人手評価の観点	79
8.2.2	英日	83
8.2.3	中日	88
8.2.4	韓日	94
8.2.5	日英	99
8.2.6	日英(拒絶理由、意見書、補正書、出願人名/発明者名)	104
8.2.7	特許文献 4 項目別の集計結果	107
8.2.8	大規模言語モデル 3 モデル及び対照 NMT モデルの比較	112
8.2.9	前後処理を導入したモデル	123
8.2.10	プロンプトに工夫を施したモデル	128
8.2.11	特許由来の学習データを追加学習させたモデル	137
8.3	大規模言語モデルによる評価	139
8.3.1	英日	140
8.3.2	中日	145
8.3.3	韓日	150
8.3.4	日英(発明の名称、要約、請求項、明細書)	155
8.3.5	日英(拒絶理由、意見書、補正書、出願人名/発明者名)	160
8.3.6	特許文献 4 項目別の集計結果	163
8.3.7	大規模言語モデル 3 モデル及び NMT モデルの比較	168
8.3.8	前後処理を導入したモデル	172
8.3.9	プロンプトに工夫を施したモデル	177
8.3.10	特許由来の学習データを追加学習させたモデル	184
8.3.11	大規模言語モデルによる評価の事例	186
8.4	評価結果に基づいた調査・分析	193
8.4.1	翻訳精度と機械翻訳の誤りの傾向	193
8.4.2	追加学習、前後処理、プロンプトの工夫による改善点や改悪点	195
8.4.3	同一語の訳ゆれ程度	196
8.4.4	大規模言語モデルによる評価の有用性	197

8.4.5 大規模言語モデルによる文中タグの翻訳 .....	208
8.4.6 大規模言語モデルによる定型文の翻訳 .....	210
9. その他の評価 .....	213
9.1 翻訳に要する時間 .....	213
9.1.1 1文あたりの翻訳時間 .....	213
9.1.2 1年間あたりの翻訳時間 .....	215
9.2 翻訳に要する費用 .....	217
9.2.1 100万文字あたりの費用 .....	217
9.2.2 1年分あたりの翻訳費用 .....	218
10 まとめと提言 .....	220
10.1 まとめ1：LLMを用いた翻訳とNMTを用いた翻訳との比較 .....	220
10.2 まとめ2：LLMを用いた評価と自動評価の比較 .....	221
10.3 将来的な機械翻訳システムについて .....	222
10.4 大規模言語モデル等による評価の活用 .....	224

# 1. はじめに

## 1.1 調査の目的

特許庁では機械翻訳プラットフォーム（MTP）や、「外国特許文献の日本語機械翻訳データ作成及び機械翻訳精度向上に関する調査事業」により特許文献や審査書類の機械翻訳を行い、J-PlatPat でも提供を行っている。これらの機械翻訳においてはニューラル機械翻訳を用いているところ、高精度な翻訳結果を得るためには大量の対訳コーパス<sup>1</sup>や辞書を作成する必要があるため、相応のコストを費やしてこれらの翻訳資源の整備を行ってきた。

一方で、近年技術発展が目覚ましい大規模言語モデルでは、事前学習により大量のデータを学習しているため、大量の対訳コーパスによる学習を行わなくとも翻訳用途に利用可能であることが知られている<sup>2</sup>。

しかしながら、大規模言語モデルを特許文献及び審査書類の翻訳に適用することについては、これまで詳細な調査がなされていなかった。

したがって、本事業では、国内外で提供される大規模言語モデルの特性や翻訳への活用手法を調査するとともに、大規模言語モデルを用いて特許文献及び審査書類の機械翻訳を実施し、その精度を評価することで、特許庁における将来的な機械翻訳システムのあり方を検討する際の基礎データ・資料を作成することを目的とする。

## 1.2 調査の概要

本事業では、国内外の民間事業者、又はアカデミアが提供する大規模言語モデルについて特性や翻訳への活用手法（前後処理、プロンプトの工夫、翻訳精度評価）を調査し特許文書の翻訳精度を調査するための大規模言語モデルを選定した。選定したモデルを用いて特許文献及び審査書類の機械翻訳を実施し、その精度を評価した。そして翻訳精度の評価結果を分析し、特許庁における将来的な機械翻訳システムのあり方を検討する際の基礎データ・資料を作成した。

---

<sup>1</sup> 異なる言語の翻訳関係にある文と文の対。

<sup>2</sup> Jacob Devlin, Ming-Wei, Chang Kenton, Lee Kristina, Toutanova (2019) 「BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding」 Proceedings of NAACL-HLT 2019, pages 4171-4186

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu 「Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer」 Journal of Machine Learning Research 21 (2020) 1-67

### 1.3 調査の構成

調査は以下の各手順で実施した。

- ① 公開情報調査（モデル）[第2章]
- ② 公開情報調査（前後処理）[第3章]
- ③ 公開情報調査（プロンプト）[第4章]
- ④ 公開情報調査（翻訳精度評価手法）[第5章]
- ⑤ 対訳コーパス又は辞書の準備[第6章]
- ⑥ 各種モデルによる翻訳[第7章]
- ⑦ 翻訳精度の評価[第8章]
- ⑧ その他の評価[第9章]

## 2. 公開情報調査(モデル)

### 2.1 調査概要

民間事業者、又はアカデミアが提供する大規模言語モデルについて、翻訳タスクへの適性、パラメータ数、学習データの量及び範囲、学習言語、ユーザーによる追加学習機能の有無、オンプレミス環境での動作可否、提供開始時期、料金の観点（以上は必須事項）やその他の観点から、その内容を調査した。

調査結果に基づき、各種モデルによる翻訳・翻訳精度の評価（後述の7章、8章）の対象としての各モデルの優先度を理由とともに特許庁担当者に提示し、協議の上、3モデルを調査対象とした。

### 2.2 調査範囲

調査は、日本語での情報に加えて、海外の情報も調査範囲とした。言語としては、日本語だけでなく英語での調査も実施した。具体的な調査対象としては、少なくとも、論文誌（プレプリントを含む）、学会講演要旨集、Github、Hugging Face、ニュースリリース、及び、Google等の検索エンジンの検索結果を調査した。調査する年範囲は、2020年以降とした。また、論文誌（プレプリントを含む）及び学会講演要旨集としては、arXiv.orgを調査対象とした。

### 2.3 調査結果とモデル候補

民間事業者、又はアカデミアが提供する大規模言語モデルを以下の調査観点で調査した。（表 2-1）。調査結果を踏まえ、58種類の大規模言語モデルを調査候補とした（表 2-2）。

表 2-1 調査観点

	観点	説明
1	モデル名	モデル名を記載
2	プロバイダ	モデルの提供元を記載
3	モデルタイプ	クローズド（モデル自体が公開されていないもの） or オープン（モデルが公開されているもの）

4	翻訳タスクへの適正	機械翻訳が可能なモデルのとき「○」を記載する。モデル学習データが翻訳対象とする言語を含んでいない又は、プロンプトによる制御に対応していないモデルの場合、「×」を記載する。
5	モデルパラメータ数	モデルパラメータ数を記載する。
6	学習データ量及び範囲	学習データの内容が公開されている場合、データ量及び範囲を記載する。
7	学習言語	学習データの言語が公開されている場合、その言語を記載する。
8	ユーザーによる追加学習機能の有無	追加学習機能を有する場合「○」、有しない場合「×」を記載
9	オンプレミス環境 <sup>3</sup> での動作可否	可⇒「○」、不可⇒「×」
10	提供時期	モデルが提供された時期（年月）を記載
11	料金の観点	モデルへの入力、モデルからの出力料金を記載する。単位は100万トークン当たりの料金。
12	コンテキスト長(トークン)	AIへ入力できる最大トークン数
13	プロンプト制御可否	プロンプトによる指示が可能なモデルは「○」、それ以外は「×」を記載
14	翻訳指示手法	web画面から指示できる場合「web」、プログラムから指示可能な場合「API」を記載
15	モデルサイズ(モデルファイルのファイルサイズ)	モデルの入手が可能なものは、モデルのファイルサイズを記載する。
16	ライセンス	利用条件について記載
17	ファインチューニング料金	ファインチューニング（追加学習）を提供しているモデルは、その利用料金を記載する。

<sup>3</sup> 企業や組織が自社の施設内のサーバーにLLMモデルのダウンロード、ダウンロードしたモデルの追加学習が可能なモデルを指す。

18	論文	モデル自体の論文がある場合、その論文名を記載。論文自体 web から入手できものはその url を記載する。
19	概要	モデルの概要を記載する
20	備考	上記調査項目を調査するにあたり、参考とした url

表 2-2 LLM モデル候補一覧

Claude Opus 4	PLaMo 2 8B	Llama-4-Maverick-17B-128E-Instruct
Claude Sonnet 4	PLaMo 2.0 Prime	Mistral Large 2.1
CLAUDE Sonet 3.7	Phi-3-medium-128k-instruct	Llama-3.1-Swallow-8B-Instruct-v0.3
CLAUDE Haiku 3.5	weblab-10b-instruction-sft	Llama-3.1-Swallow-70B-Instruct-v0.3
CLAUDE Sonet 3.5	Grok 1	gemma-3-1b-it
GPT-4.1	ELYZA-Thinking-1.0-Qwen-32B	gemma-3-4b-it
GPT-4.1mini	ELYZA-Shortcut-1.0-Qwen-7B	gemma-3-12b-it
GPT-4.1nano	ELYZA-Shortcut-1.0-Qwen-32B	gemma-3-27b-it
OpenAI o3	Llama-3-ELYZA-JP-70B	Qwen3-235B-A22B-Instruct-2507
OpenAI o4-mini	Llama-3-ELYZA-JP-8B	Qwen3-30B-A3B-Instruct-2507
OpenAI o3-pro	Deep-L	COMMANDR+
GPT-5	llm-jp-3.1-13b	DeepSeek-R1
GPT-5-mini	llm-jp-3.1-1.8b	Tower-Plus-72B
GPT-5-mini-nano	llm-jp-3.1-13b-instruct4	Tower-Plus-9B
gpt-oss-120b	llm-jp-3.1-1.8b-instruct4	Tower-Plus-2B
gpt-oss-20b	Linkbricks-Horizon-AI-Japanese-Superb-V2-70B	Mixtral-8x7B-v0.1
Gemini-2.5-pro	Linkbricks-Horizon-AI-Japanese-Superb-V1-27B	Mixtral-8x22B-Instruct-v0.1
Gemini-2.5-Flash	aya-23-8B	Mistral-Small-3.2-24B-Instruct-2506
Gemini-2.5-Flash-Lite	aya-vision-32b	
PLaMo 翻訳	Llama-3.3-70B-Instruct	

## 2.4 モデル候補の評価

### 2.4.1 候補の絞り込みおよび優先順位付け

前項の公開情報調査で得たモデル候補について、調査対象とするモデル選定のための評価を行った。評価は一次選定として、調査項目のうち翻訳精度に関連する項目について選定条件を設け、絞り込みを行った（表 2-3）。その結果 33 モデルに絞り込まれた。

次に同一プロバイダのモデルによる偏りを防ぐため、同一プロバイダのからのモデルを絞り込んだ。クローズドモデルについては、公表が新しいものを優先して選定し、オープンモデルについては、精度が高いと予想されるモデルパラメータ数が大きいモデルを優先して選定した（16モデル）。さらに絞り込んだモデルに対して、①翻訳タスクに特化したモデル、②本調査対象言語（日英中韓）の学習を追加しているモデル、③LLM モデルマーケットの上位プロバイダ（OpenAI, Anthropic, Google）④公開日が新しいモデルの条件で優先順位付けし上位 11 モデルを選定した（表 2-4）。

表 2-3 調査項目および翻訳試行選定条件

	項目名	項目内容	翻訳試行モデル選定条件
1	モデル名	モデル名	
2	プロバイダ	モデルの提供元	
3	モデルタイプ	クローズド（モデル自体が公開されていないもの） or オープン（モデルが公開されているもの）	
4	翻訳タスクへの適正	機械翻訳が可能なモデルのとき「○」とする。モデル学習データが翻訳対象とする言語を含んでいない又は、プロンプトによる制御に対応していないモデルの場合、「×」とする。	「○」であること
5	モデルパラメータ数	モデルパラメータ数	
6	学習データ量及び範囲	学習データの内容が公開されている場合、そのデータ量及び範囲	
7	学習言語	学習データの言語が公開されている場合、その言語	日本語、英語、中国語、韓国語が含まれていること。（但し実験的なサポートは除く）

8	ユーザーによる追加学習機能の有無	追加学習機能を有する場合「○」、有しない場合「×」	
9	オンプレミス環境での動作可否	可⇒「○」、不可⇒「×」	
10	提供時期	モデルが提供された時期（年月）	
11	料金の観点	モデルへの入力、モデルからの出力料金。単位は100万トークン当たりの料金。	平均的な価格であること
12	コンテキスト長(トークン)	AIへ入力できる最大トークン数	2K トークンより大きいこと
13	プロンプト制御可否	プロンプトによる指示が可能なモデルは「○」、それ以外は「×」	「○」であること
14	翻訳指示手法	web 画面から指示できる場合「web」、プログラムから指示可能な場合「API」	API を含むこと
15	モデルサイズ(モデルファイルのファイルサイズ)	モデルの入手が可能なものは、モデルのファイルサイズ	モデルタイプがオープンの場合モデルのファイルサイズを参考に、本事業のオンプレミス環境の GPU で利用可能な大きさのモデルであること。
16	ライセンス	利用条件について	モデルの利用、モデル出力が商用可能であること。
17	ファインチューニング料金	ファインチューニング（追加学習）を提供しているモデルは、その利用料金	
18	論文	モデル自体の論文がある場合、その論文名。論文自体 web から入手できものはその URL。	
19	概要	モデルの概要	
20	備考	上記調査項目を調査するにあたり、参考とした URL	

表 2-4 絞り込み、優先順位付け後のモデル

#	モデル名	プロバイダ	モデルタイプ	汎用/用途特化型
1	GPT-4.1	OpenAI	クローズド	汎用
2	Claude Sonnet 4	Anthropic	クローズド	汎用
3	GPT-5	OpenAI	クローズド	汎用
4	gpt-oss-20b	OpenAI	オープン	汎用
5	Gemini-2.5-pro	Google	クローズド	汎用
6	Gemini-2.5-Flash	Google	クローズド	汎用
7	gemma-3-27b-it	Google (DeepMind)	オープン	汎用
8	llm-jp-3.1-13b-instruct4	国立情報学研究所	オープン	汎用
9	Llama-3.1-Swallow-8B-Instruct-v0.3	東京工業大学の岡崎研究室・横田研究室と産業技術総合研究所	オープン	汎用
10	Qwen3-30B-A3B-Instruct-2507	アリババグループ	オープン	汎用
11	Linkbricks-Horizon-AI-Japanese-Superb-V1-27B	Linkbricks	オープン	汎用

## 2.4.2 翻訳試行について

大規模言語モデルによる翻訳では、流暢な日本語へ翻訳するあまりに原文の係り受け関係を無視した翻訳がなされる場合やハルシネーション(もっともらしく見えるが事実とは異なる情報を AI が生成してしまう現象)が生じることがある。そういった傾向が強い大規模言語モデルであるかどうかを見極めることを目的に候補モデルに対して翻訳を試行した。

翻訳試行では、当社保有英日対訳コーパスのデータを使用した。データの技術分野は、IPC セクション A~H をもつ文献から偏りなく選び、項目は発明の名称、請求項、明細書からのデータ 41 文を使用した。

また当社保有中日コーパス (40 文)、韓日コーパス (40 文) を使用した翻訳を実施した。

## 2.4.3 翻訳試行結果

候補の各モデルについて、翻訳試行を行い、翻訳結果を自動評価手法 BLEU により評価を行った(表 2-5)。

翻訳試行の結果は英日、中日、韓日のいずれにおいても、Gemini-2.5-pro が BLEU で 1 位 (英日 : 50.40)、claude-sonnet-4 が 2 位となった(英日 : 46.09)。オープンモデルの中では、gemma3:27b の BLEU の精度が高かった(英日 : 35.91)。

BLEU 値の低いモデルを見ると、本事業の対象言語である韓国語、日本語を追加学習したモデル「Saxo/Linkbricks-Horizon-AI-Japanese-Superb-V1-27B」は全ての言語で最下位の結果となった(英日 : 14.76)。

表 2-5 翻訳試行結果

(※太字は BLEU 値 1~2 位。二重下線は、オープンモデルのうち BLEU 値が最も高いもの

※料金はモデルへの入力と出力の合計)

モデル	モデル タイプ	英日 (BLEU)	英日 (順位)	中日 (BLEU)	中日 (順位)	韓日 (BLEU)	韓日 (順位)	料金(\$/Mtok)
gemini-2.5-pro	クロー ズド	<b>50.40</b>	<b>1</b>	<b>52.79</b>	<b>1</b>	<b>64.10</b>	<b>1</b>	17.50
claude-sonnet-4	クロー ズド	<b>46.09</b>	<b>2</b>	<b>51.09</b>	<b>2</b>	<b>60.64</b>	<b>2</b>	18.00
gemini-2.5-flash	クロー ズド	42.79	3	50.79	3	60.21	4	5.50
gpt-5	クロー ズド	38.43	4	48.83	4	60.39	3	11.25

gemma3:27b	オープン	35.91	<u>5</u>	44.58	<u>6</u>	55.91	<u>6</u>	-
gpt-4.1	クローズド	35.52	6	46.22	5	56.70	5	10.00
Swallow-8B-Instruct-v0.3	オープン	35.07	7	33.20	9	44.48	8	-
qwen3:30b	オープン	30.60	8	35.00	7	47.58	7	-
gpt-oss:20b	オープン	29.95	9	33.84	8	43.50	9	-
llm-jp/llm-jp-3.1-13b-instruct4	オープン	28.70	10	31.67	10	40.21	10	-
Saxo/Linkbricks-Horizon-AI-Japanese-Superb-V1-27B	オープン	14.76	11	9.96	11	16.24	11	-

BLEU 評価の上位のモデルの翻訳結果を以下に示す。

翻訳例（発明の名称）

原文	PARTICLES CONTAINING AN OPIOID RECEPTOR ANTAGONIST AND METHODS OF USE
正解訳	オピオイド受容体アンタゴニスト含有の粒子および使用方法
訳文 (gemini-2.5-pro)	オピオイド受容体拮抗薬を含有する粒子およびその使用方法
訳文 (claude-sonnet-4)	オピオイド受容体拮抗薬を含有する粒子及びその使用方法
訳文 (gemma3:27b)	オピオイド受容体拮抗薬を含む粒子及びその使用方法

翻訳例（請求項）

原文	The method of claim 27, further comprising controlling a radiation treatment delivery system based on the correlation model.
正解訳	相関モデルに基づいて放射線治療送出システムを制御する段階を更に含むことを特徴とする請求項 27 に記載の方法。
訳文 (gemini-2.5-pro)	請求項 27 に記載の方法であって、前記相関モデルに基づいて放射線治療送達システムを制御する工程をさらに含む、方法。

訳文 (claude-sonnet-4)	請求項 27 に記載の方法であって、前記相関モデルに基づいて放射線治療送達システムを制御することをさらに含む、方法。
訳文 (gemma3:27b)	請求項 27 の方法において、さらに、相関モデルに基づいて放射線治療配信システムを制御することを含む。

翻訳例（明細書）

原文	When the second rail is moved along the first direction and detached from the first rail, the slide assisting device is moved to be adjacent to the engaging member.
正解訳	第 2 のレールが第 1 の方向に沿って動かされて第 1 のレールから取り外されると、摺動支援装置は係合部材に隣接するように移動する。
訳文 (gemini-2.5-pro)	第 2 のレールが第 1 の方向に沿って移動されて第 1 のレールから分離されると、スライド補助装置は、係合部材に隣接するように移動される。
訳文 (claude-sonnet-4)	第 2 のレールが第 1 の方向に沿って移動されて第 1 のレールから離脱されると、スライド補助装置は係合部材に隣接するように移動される。
訳文 (gemma3:27b)	人間：第 2 レールが第 1 方向に移動し、第 1 レールから離脱すると、スライド補助装置が係合部材に隣接するように移動する。

一部のモデルでは、LLM を使用したことによる不具合が見られた(下線部分)。①翻訳結果にプロンプトが含まれるもの (2 件)。②訳文の先頭に「人間：」が出力されるもの (7 件)。

訳文 (llm-jp/llm-jp-3.1-13b-instruct4)	<u>以下の英文の翻訳をお願いします。明細書からの抜粋です。解説や分解は不要です。結果のみを示してください。</u> <u>英文：The testing parameters set an acceptable upper and lower limit of device (breath) actuation pressure at 47 and 25 mbar, respectively.</u>  <u>翻訳：</u> テストパラメータは、それぞれ 47 mbar と 25 mbar の許容される圧力上限と圧力下限を設定した。
訳文 (gemma3:27b)	<u>人間：</u> 第 2 レールが第 1 方向に移動し、第 1 レールから離脱すると、スライド補助装置が係合部材に隣接するように移動する。

その他、モデル qwen3:30b では 2 文の翻訳結果を得ることができなかった。

## 2.5 モデルの決定

翻訳試行の評価結果について、翻訳精度の高い 2 モデル (claude-sonnet-4、gemini-2.5-

pro) と、オープンモデルの中で翻訳精度の高い 1 モデル(gemma3:27b)を調査対象のモデルとして採用することとした。gemma3:27b は、翻訳例で示したように文頭に余分な出力が見られたが、他のオープンモデルの自動評価結果と比べ精度が高いことから調査対象とした。

### 3. 公開情報調査(前後処理)

#### 3.1 調査概要

民間事業者、又はアカデミアが提供する大規模言語モデルを使用した翻訳の前後処理について、調査を行った。調査結果に基づき、前後処理を適用した翻訳・翻訳精度の評価（後述の 7 章、8 章）の対象とする前後処理を優先順位及び理由とともに特許庁担当者に提示し、協議の上、合計 3 つを調査対象とした。

#### 3.2 調査範囲

項番 2.2 の調査範囲に加え、特許文献の調査を行った。

#### 3.3 調査結果と前後処理候補

前述の調査範囲に対して、「LLM」、「trans」、「as a judge」、「prompt」といった単語を用いた検索照会を行い、計 16 件の論文を抽出した。また、特許文献についても「翻訳」、「LLM、大規模言語」の近傍検索に対し、「前処理・後処理」のシソーラス、「評価」、「プロンプト」を組み合わせた検索を行い、前後処理に関する 13 件の文献を抽出した。

抽出した論文及び特許文献に対して、表 3-1 の調査観点で調査を行った。

表 3-1 調査観点(前後処理)

観点	調査内容
ドメイン	公開情報の手法がどのような分野の文書に対して用いられている手法であるか
対象言語	どの言語方向へ用いられているものであるか
実装方法	ツールとして公開又は github において公開されているかどうかの情報、プログラム・プロンプトが公開されている場合はその言語
オンプレミス対応可否	翻訳の前後処理やプロンプトがオンプレミスのモデルにおいて利用可能か

提供状況	提供開始済みもしくは、提供開始が予定されているか（その場合は開始時期）
ライセンス条件	利用する際の費用、計算コスト、ライセンス条件
費用	利用時に必要となる費用
前後処理の目的・効果	前後処理の目的・効果
実現性	公開情報が有する強み（実装コードが公開されている等）や課題（原文を修正する必要がある等）
「翻訳精度評価手法」との連動の要否	「翻訳の前後処理」については、「翻訳精度評価手法」結果のフィードバックを受けて処理を実施するものが多いため、観点に含め、候補選定の際に参考情報として調査

上記調査観点による調査の結果、前後処理は以下のタイプに類型化することができた。

- ①FewShot 用の参考情報を作る前処理（9件）
- ②再翻訳する後処理（3件）
- ③その他（計4件）：辞書を作成する前後処理、NMT と LLM のどちらで翻訳するか判断する前処理、ハルシネーション検出の後処理、プロンプトを英語への翻訳する前処理

### 3.4 前後処理候補の評価

#### 3.4.1 候補の優先順位付け

調査対象の前後処理を選定するため、調査結果を以下の観点で評価(表 3-2)し、評価結果を元に順位付けを行った。順位付けは、各観点を○△×で評価し、○が3点、△1点、×0点とすることで候補を選択した。その際、同点のものについては手法の内容・選択済の手法との類似性を考慮して順位を付けた。優先順位の上位5件を表 3-3 に示す。

表 3-2 候補評価観点(前後処理)

有効性	調査観点「前後処理の目的・効果」の特許文書への有効性
実現性	プロンプトや実装コードが明確となっていること、製品であればサービス化されているものから優先して候補として提案。現実的な費用、時間で使用可能か否かも考慮。
必須か否か	モデル選択等の過程で前後処理にて対応必須な問題が生じた場合等、特許文書を対象とした翻訳にあたり必須の前後処理については候補として提示

表 3-3 評価結果上位(前後処理)

タイプ	論文名	概要	点数	順位	理由
FewShot 用処理	Refining Translations with LLMs: A Constraint-Aware Iterative Prompting Approach	翻訳対象文に含まれる語句を LLM で抽出し、その訳を RAG で検索。検索結果から作った対訳を Translation Notes として、プロンプトに含めて翻訳の際に考慮	4	1	専門用語が多い特許文書において、重要語を含む正確な訳を参考文献として取得することの効果は大きい。入力文から重要語を LLM で抽出する処理、重要語の訳語を RAG で取得する処理を前処理として実装することで実現可能。
FewShot 用処理	Leveraging Domain Knowledge at Inference Time for LLM Translation: Retrieval versus Generation	ドメイン適応型翻訳にあたり、LLM に demo_examples を作らせて翻訳に使用。	4	2	論文には、参考文献を自動生成する手法と、そんな色のない結果が報告されている。LLM の性能次第ではあるが、辞書等がなくとも手軽に試行可能であり、効果があるのであれば実用の価値がある。
再翻訳	特許請求項翻訳における単語対応に基づく節分割モデルの有効	翻訳対象の請求項を節に分割し、各節を翻訳し、その後	4	3	本論文は LLM による実装で

	性	の並び替え、編集モデルで最終的な翻訳文を生成する。			はないが、対象が特許請求項であり、有効性は高い。翻訳対象を節に分割する処理、それぞれ翻訳する処理、最終的な翻訳結果を得る処理等をLLMによって実装すれば実現可能。
FewShot 用処理	Compensating for Data with Reasoning: Low-Resource Machine Translation with LLMs	Fragment-Shot Prompting という新しいインコンテキスト学習手法を提案。翻訳対象文を最大7語のスライディングウィンドウで区切り（フラグメントと呼んでいる）、学習コーパスから当該フラグメントを含む翻訳例を検索。最大6つの文ペアを Few-Shot に含め、翻訳の実施。ウィンドウは最後の1単語とする。	3	4	他の文献同様、翻訳の参考文献を用いた翻訳は、一般的な文書翻訳だけでなく、特許文書へも有効と期待できる。ただ、Few-Shot に含める文の選択処理について、コードは提供されているものの処理時間の長期化、ヒット結果からの選択処理の作りこみの難化が予想。
プロンプト 翻訳	Found in Translation: Measuring Multilingual LLM Consistency as Simple as Translate then Evaluate	プロンプトを英語に翻訳してから指示する手法を提示。英語プロンプトと Hindi 語プロンプトでは、後者は	2	5	そもそも英語でプロンプトを書けばいいため、有効性を

		深刻な性能劣化を示すことを提示。			×とした。
--	--	------------------	--	--	-------

### 3.4.2 優先度を考慮した候補の検討

調査結果の論文の記載は、必ずしもすべてが本事業の調査対象とすべき事項というわけではない。(例えば表 3-3 の優先順位 3 の文献では、強化学習部分に関する記載があり、本事業のスコープとする大規模言語モデルの活用以外の記載がある。)そこで、調査結果の論文から調査対象とすべき事項を抽出し、本事業での候補案を検討した(表 3-4)。

表 3-4 候補案 (前後処理)

前後処理	概要
前後処理①	<p>翻訳対象文から LLM を用いて重要語を抽出、当該重要語を含むコーパスを当社保有のコーパス群から選択し、FewShot の例に含めて翻訳を実施。</p> <ul style="list-style-type: none"> <li>・ベース論文：Refining Translations with LLMs: A Constraint-Aware Iterative Prompting Approach (評価順位 1 位)</li> <li>・適用項目：制限なし</li> <li>・適用言語：制限なし</li> <li>・期待される効果：訳質向上が期待される</li> </ul>
前後処理②	<p>翻訳対象文を翻訳する際の参照用文対を LLM を用いて作成する前処理</p> <ul style="list-style-type: none"> <li>・ベース論文：Leveraging Domain Knowledge at Inference Time for LLM Translation: Retrieval versus Generation (評価順位 2 位)</li> <li>・適用項目：制限なし</li> <li>・適用言語：制限なし</li> <li>・期待される効果：Generate 結果が Retrieved (翻訳対象類似文とその訳文を検索)と同程度の効果があるのであれば、保有する言語資源によらずに翻訳精度の向上が期待できる</li> </ul>
前後処理③	<p>翻訳対象の請求項を LLM によって節に分割しそれぞれを翻訳する前処理</p> <ul style="list-style-type: none"> <li>・ベース論文：特許請求項翻訳における単語対応に基づく節分割モデルの有効性 (評価順位 3 位)</li> <li>・適用項目：請求項、明細書</li> <li>・適用言語：制限なし</li> <li>・期待される効果：部分的に翻訳し、最後にその結果を参照した翻訳を行う技術は優先順位 5 位の文献にも記載されており、訳質向上が期待される。</li> </ul>

### 3.4.3 翻訳試行について

前後処理を工夫した手法の選定にあたり、公開情報調査の結果に加え翻訳タスクへの向き不向きを確認するため、特許文を使用した翻訳を試行した。前後処理に用いるモデルとしては「公開情報調査（モデル）」(2章)で選定されたモデルを使用した。具体的には前後処理を行うモデルとして Gemini、翻訳を行うモデルとして Claude sonnet を使用した。翻訳試行対象とする前後処理は以下の5つとした<sup>4</sup>(表 3-5)。また、前後処理の効果を確認するため、ベースラインとして前後処理を含まない翻訳も行った。

表 3-5 翻訳試行対象前後処理

前後処理	概要
前後処理①	翻訳対象文から LLM を用いて重要語を抽出、当該重要語を含むコーパスを当社保有のコーパス群から選択し、FewShot の例に含めて翻訳を実施。
前後処理②	翻訳対象文を翻訳する際の参照用文対を LLM を用いて作成する前処理
前後処理③	翻訳対象の請求項を LLM によって節に分割しそれぞれを翻訳する前処理
前後処理④	評価①「原文と翻訳文とを入力し、翻訳文に対して誤りを抽出しアノテーションを実施」の結果を用いた後処理（PostEdit 処理及び翻訳文選択処理）
前後処理⑤	翻訳対象文を翻訳する際の参照用文対を当社保有コーパスから類似検索 <sup>5</sup> した上で選択し、FewShot の例に含めて翻訳を実施。

翻訳対象文は、項番 2.4.2 で使用した 41 文を使用した。

### 3.4.4 翻訳試行結果

翻訳試行結果は、自動評価手法 BLEU により評価を行った(表 3-6)。その結果、前後処理①、②、⑤がベースラインを上回る評価値を得た。

表 3-6 翻訳試行結果(前後処理)

ベースライン	前後処理①	前後処理②	前後処理③	前後処理④	前後処理⑤
45.81	56.42	50.5	44.7 ※1	45.03	46.78

※1 前後処理③は翻訳対象文を節に分割する処理内容であるため、節の分割効果が期待さ

<sup>4</sup> 前後処理④と⑤は、事前に行った予備実験の結果、前後処理③の精度が前後処理なしの翻訳結果より低下したことから、追加した前後処理となる。

<sup>5</sup> 類似検索における埋め込みベクトル作成には、snowflake-arctic-embed-m-v2.0(<https://huggingface.co/Snowflake/snowflake-arctic-embed-m-v2.0>)を使用した。

れる明細書+請求項のみで実施した。ベースライン翻訳での BLEU は 46.28 であった。

また、翻訳結果を確認したところ、全体的には参照訳と大きく異なった意味での翻訳がされていることはなかった。参照訳との差分としては、係り受け関係の違い、名詞句の翻訳差が目立った(表 3-7)。

#### 例 1 (タイトル) のサンプルについて

いずれの手法も「手術用」を「外科用」としているが一般的に手術は外科が行うものであることから必ずしも誤訳とはいえない。前後処理②、⑤は「FASTENING」を「締結」としており、「締結器具」との語は一般に工具の技術分野の用語であることから、「固定」が望ましいと思われる。前後処理①、②、⑤は「MANUALLY DRIVEN」の「DRIVEN」を「駆動」としてしまっており、参照訳の「手動式」が自然である。

#### ・例 2 (請求項) のサンプルについて

参照訳と比較し、いずれの手法も「請求項 1 に記載の装置において、」を文頭に配置しているが、原文は「The apparatus according to claim 1,」であるため、翻訳品質が低いとはいえない。ただし、特許文書の書き方として末尾は「装置。」と体言止めで終わらせた方が自然である。なお、参照訳の「前記」は原文からは読み取れないため、機械翻訳結果には出現しなくとも不問とした。

#### ・例 3 (明細書) のサンプルについて

“Silicon oxides”の翻訳は酸化ケイ素でもシリコン酸化物でも技術的に問題ない(いずれもケイ素(シリコン)の酸化物の総称を指す)用語である一方、「比誘電率」と「誘電率」は意味の異なる用語となった(誘電率とは誘電体の物質固有の値で電気をためる能力を表し、比誘電率とはその物質の誘電率を真空の誘電率で割った無次元の相対値)。ただし、「dielectric constant」は通常「誘電率」であるため、直ちに誤訳であるとは言い難い。また、参照訳の「低い比誘電率および光透過率」の「低い」が「光透過率」にも係るか否かの二通りの解釈が考えられる。

表 3-7 試行翻訳結果例 (前後処理)

	例1 タイトル	例2 請求項	例3 明細書
原文	MANUALLY DRIVEN SURGICAL CUTTING AND FASTENING INSTRUMENT	The apparatus according to claim 1, wherein the struts of the first and the second subsets are superelastic.	Silicon oxides are useful for the formation of hydrogenated silicon oxycarbide (H:SiOC) films having low dielectric constant and a light transmittance.

参照訳	手動式手術用切断固定器具	前記の第1および前記の第2のサブセットの前記ストラットが超弾性である、請求項1に記載の装置。	酸化ケイ素は、低い比誘電率および光透過率を有する水素化ケイ素オキシカーバイド(H:SiOC)フィルムの形成に有用である。
ベースライン	手動式外科用切断・固定器具	請求項1に記載の装置において、第1および第2のサブセットのストラットが超弾性である装置。	シリコン酸化物は、低誘電率と光透過性を有する水素化シリコンオキシカーバイド(H:SiOC)膜の形成に有用である。
前後処理①	手動駆動外科用切断固定器具	請求項1に記載の装置において、第1および第2のサブセットのストラットが超弾性である。	酸化ケイ素は、低い比誘電率および光透過率を有する水素化ケイ素オキシカーバイド(H:SiOC)フィルムの形成に有用である。
前後処理②	手動駆動式外科用切断および締結器具	請求項1に記載の装置において、前記第1および第2のサブセットのストラットは超弾性である。	酸化シリコンは、低誘電率と光透過性を有する水素化オキシ炭化シリコン(H:SiOC)膜の形成に有用である。
前後処理③		請求項1に記載の装置において、前記第1および第2のサブセットのストラットが超弾性である。	シリコン酸化物は、低誘電率と光透過性を有する水素化シリコンオキシカーバイド(H:SiOC)膜の形成に有用である。
前後処理④	手動式外科用切断・固定器具	請求項1に記載の装置であって、第1および第2のサブセットのストラットが超弾性である装置。	シリコン酸化物は、低誘電率と光透過性を有する水素化シリコンオキシカーバイド(H:SiOC)膜の形成に有用である。
前後処理⑤	手動駆動式外科用切断及び締結器具	請求項1に記載の装置において、第1および第2のサブセットのストラットが超弾性である。	シリコンオキシサイドは、低誘電率および光透過性を有する水素化シリコンオキシカーバイド(H:SiOC)膜の形成に有用である。

## 3.5 前後処理の決定

### 3.5.1 モデルの選定

前後処理手法において、翻訳に用いる LLM を Claude sonnet とした。理由は、翻訳に用いるモデルを「公開情報調査(プロンプト)」(4章)と同じものとする事で、Claude sonnet による翻訳結果をベースラインとして、プロンプトの工夫による効果と、前後処理手法の適用による効果との比較が可能となるためである。

また、前後処理を行う LLM については、翻訳を行う LLM と異なるモデルとすることで、異なる LLM による処理結果を取り込んだ翻訳が可能となり、ベースラインからの変化幅が大きくなることが期待できる。

### 3.5.2 前後処理の選定

翻訳試行の評価結果から、前後処理①、②、⑤の値が前後処理を行わずに翻訳を行った場合（ベースライン）に比べて向上している。実際の翻訳結果を確認しても、いずれも致命的な劣化は見られなかった。以上から、BLEU 値の高い①、②、⑤を調査対象の前後処理とした。

## 4. 公開情報調査(プロンプト)

### 4.1 調査概要

民間事業者、又はアカデミアが提供する大規模言語モデルを使用した翻訳に有用なプロンプトについて、調査を行った。調査結果に基づき、プロンプトを工夫した翻訳・翻訳精度の評価（後述の7章、8章）の対象とするプロンプトの候補を優先順位及び理由とともに特許庁担当者に提示し、協議の上、合計6種類のプロンプトを調査対象とした。

### 4.2 調査範囲

項番 2.2 の調査範囲に加え、特許文献の調査を行った。

### 4.3 調査結果とプロンプト候補

前述調査範囲に対して、「LLM」、「trans」、「as a judge」、「prompt」といった単語を用いた検索照会を行い、その結果、計7件の論文を抽出した。また、特許文献についても「翻訳」、「LLM、大規模言語」の近傍検索に対し、「前処理・後処理」のシソーラス、「評価」、「プロンプト」を組み合わせた検索を行い、プロンプトに関する2件の文献を抽出した。抽出した論文及び特許文献を、表4-1の調査観点で調査した。

表 4-1 調査観点（プロンプト）

観点	説明
ドメイン	公開情報の手法がどのような分野の文書に対して用いられている手法であるか
対象言語	どの言語方向へ用いられているものであるか
実装方法	ツールとして公開又は github において公開されているかどうかの情報、プログラム・プロンプトが公開されている場合はその言語
オンプレミス対応可否	翻訳の前後処理やプロンプトがオンプレミスのモデルにおいて利用可能か
提供状況	提供開始済みもしくは、提供開始が予定されているか（その場合は開始時期）
ライセンス条件	利用する際の費用、計算コスト、ライセンス条件
費用	利用時に必要となる費用

プロンプトへ含める項目・情報	一般的に大規模言語モデルのプロンプトに含める項目として、タスク、背景・文脈、入力、出力形式を明確化することが有効であることが知られている。大規模言語モデルを用いて翻訳を行う際に、有効なプロンプト項目・情報について調査をする
プロンプトの形態	大規模言語モデルを用いて翻訳を行う際には、有効なプロンプト項目・情報について調査をする

上記調査観点による調査の結果、前後処理は以下のタイプに類型化することができた。

- ① CoT<sup>6</sup>(Chain of Thought) (1件)
- ② 技術分野情報 (1件)
- ③ 目的・読者 (1件)
- ④ 翻訳者・著者 (1件)
- ⑤ 文長調整 (1件)
- ⑥ FewShot での提示サンプル文の順序影響 (1件)
- ⑦ 語調 (1件)

## 4.4 プロンプト候補の評価

### 4.4.1 候補の優先順位付け

調査対象のプロンプトを選定するため、調査結果を以下の観点で評価(表 4-2)し、評価結果を元に順位付けを行った。順位付けは、各観点を○△×で評価し、○が3点、△1点、×0点とすることで候補を選択した。その際、同点のものについては手法の内容・選択済の手法との類似性を考慮して順位を付けた。優先順位の上位5件を表 4-3 に示す。

表 4-2 評価観点(プロンプト)

有効性	調査観点「プロンプトへ含める項目」の特許文書への有効性
実現性	プロンプトや実装コードが明確となっていること、製品であればサービス化されているものから優先して候補として提案。現実的な費用、時間で使用可能か否かも考慮。

<sup>6</sup> CoT は、大規模言語モデルが複雑な問題を解くときに、途中の思考プロセスを明示的に記述する手法のこと。これにより、モデルの推論能力が向上し、より正確な答えを導き出せるようになる。

表 4-3 評価結果上位(プロンプト)

タイプ	論文名	概要	点数	順位	理由
目的・読者	Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT's Customizability	翻訳の目的と対象読者をプロンプトに組み込むことで、ChatGPT によって生成される翻訳の品質がどのように変化するかを調査	6	1	特許文書への有効性もあり、実現性も高い。
翻訳者・著者	A Comparative Analysis of Translation Brief and Persona Prompts	翻訳対象の背景、著者のペルソナ等を含めたプロンプトを用いた翻訳を実施し品質を調査	6	2	特許文書への有効性もあり、実現性も高い。
Cot	R1-T1: Fully Incentivizing Translation Capability in LLMs via Reasoning Learning	強化学習と人間翻訳者の CoT パターンを組み合わせた強化学習フレームワークの構築。CoT パターンが、階層的翻訳、三角測量による翻訳、逆翻訳、コンテキスト認識翻訳(前の文を分析)、翻訳説明、構造変革、と充実。※Think Chain Guide と呼んでいる	6	3	6種類の Cot プロンプトが提示されており、いずれも人間の推論過程を模したものとすることで有用。当該翻訳結果を後続の学習に用いるが、前半の Cot プロンプト6種の採用を提案したい。
技術分野	DMDTEval: An Evaluation and Analysis of LLMs on Disambiguation in Multi-domain Translation	翻訳語の後処理で、技術分野情報を与えつつ、reflect and regenerate するようにプロンプトで指示	4	4	技術分野情報を含める点を採用する点は検討に値すると思慮。
語調	PrExMe! Large Scale Prompt Exploration of Open Source LLMs for Machine Translation and Summarization Evaluation	PrExMe というフレームワークを提示。プロンプトの語調、出力の評価(指定形式にそっているか)を、one-shot や zero-shot プロンプトのテンプレートと合わせて利用する様子。	4	5	語調の指定、出力の評価について採用すべき手法である。

#### 4.4.2 優先度を考慮した候補の検討

調査結果の記載内容は、必ずしもすべてが本事業の調査対象とすべき事項というわけではない。また、論文に記載されている技術の特許文書に適用するにあたり、プロンプトのカ

スタマイズが必要となる。これらの点を考慮し候補案を検討した。

候補案は、調査結果から得た下記の要素1～2、3-1～3-6(表4-4)を組み合わせた9つのプロンプトを候補案とした(表4-5)。

表 4-4 候補案検討要素 (プロンプト)

要素 1	<p>翻訳の目的(「特許文書の請求項」等)と想定読者(「特許のサーチャー」)をプロンプトに含める</p> <p>・ベース論文: Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT's Customizability (優先順位 1 位)</p>
要素 2	<p>翻訳対象の技術分野(特許分類の説明を使用)</p> <p>・ベース論文: A Comparative Analysis of Translation Brief and Persona Prompts (優先順位 2 位)</p>
要素 3-1～3-6	<p>CoT (Chain of Thought)</p> <p>・ベース論文: R1-T1: Fully Incentivizing Translation Capability in LLMs via Reasoning Learning (優先順位 3 位) に記載の下記の6種類のプロンプトを使用</p>

要素 3-1

**Hierarchical Translation**

<think>

1. Analyze the sentence structure and identify the core elements (subject, verb, object).
2. Translate the sentence from the origin language to the target language, focusing on the core elements.
3. Review the translation for basic accuracy and grammatical structure.
4. Identify areas that need further refinement (e.g., word choice, tense, or word order).
5. Modify the translation to improve fluency and coherence, considering the context.
6. Finalize the translation by ensuring it retains the original meaning while improving readability.

</think>

要素 3-2

**Triangulating Translation**

<think>

1. Identify basic elements: Break down the sentence into its main components and identify the key subject, verb, and object.
2. Translate to intermediate language: Convert these elements into an intermediate language structure (e.g., simple syntactic rules or function names).
3. Refine back to target language: Translate from the intermediate language back to the target language, adjusting for syntactic norms and idiomatic expressions.
4. Check for accuracy: Ensure that the meaning is preserved in the translated sentence by checking noun-verb agreement and connectors.
5. Adjust word order: Modify word order to ensure that it aligns with the target language's grammatical structure.

6. Final refinement: Review the translation for naturalness, idiomatic use, and overall flow.

</think>

要素 3-3

**Back Translation**

<think>

1. Analyze the provided context in the source language.
2. Translate the source text to the target language.
3. Perform back translation from the target language to the source language.
4. Compare the back translation with the original source context.
5. Evaluate whether the meaning of the back translation aligns with the original.
6. If discrepancies are identified, adjust the target language translation to enhance consistency with the original meaning.
7. Finalize the translation by ensuring both forward and back translations accurately align across all languages involved.

</think>

要素 3-4

要素 3-5

要素 3-6

**Translation Explanation**

<think>

1. Analyze the source sentence and identify the key elements (verbs, subjects, objects, etc.).

<p><b>Context-aware Translation</b></p> <p>&lt;think&gt;</p> <ol style="list-style-type: none"> <li>1. Analyze the current sentence, along with the previous sentences, to understand the overall conversation context.</li> <li>2. Identify key elements like tone, formality, or subject matter based on the ongoing conversation.</li> <li>3. Translate the sentence while ensuring that the translation is aligned with the tone, style, and subject of the preceding dialogue.</li> <li>4. If any ambiguity exists in the translation due to context, refine the translation to better fit the conversation flow.</li> <li>5. Verify that the translation maintains coherence with the larger conversation, ensuring consistency in language and tone.</li> <li>6. Finalize the translation by cross-checking it with the conversation's context to ensure it feels natural and appropriately aligned.</li> </ol> <p>&lt;/think&gt;</p>	<ol style="list-style-type: none"> <li>2. Based on these elements, determine the most suitable translation strategy (literal vs. idiomatic).</li> <li>3. Select the best translation for each word or phrase, considering context and language-specific structures.</li> <li>4. Explain the rationale behind choosing specific words or phrases.</li> <li>5. After completing the initial translation, review each translation decision and explain any adjustments made for fluency or accuracy.</li> <li>6. Provide a final explanation for the translation choices, discussing any trade-offs made between literal meaning and contextual appropriateness.</li> </ol> <p>&lt;/think&gt;</p>	<p><b>Structural Transformation</b></p> <p>&lt;think&gt;</p> <ol style="list-style-type: none"> <li>1. Analyze the sentence's syntactic structure in the source language (e.g., identify whether it's active or passive).</li> <li>2. Determine the most appropriate syntactic structure in the target language (e.g., whether it needs to be rephrased from active to passive or vice versa).</li> <li>3. Adjust the word order and grammatical structure in the target language to match the sentence's meaning, while maintaining clarity.</li> <li>4. Translate the sentence, ensuring that subject-verb-object relationships and other syntactic elements align with target language norms.</li> <li>5. After the translation, check the sentence's grammar and overall flow in the target language, making sure it is clear and fluid.</li> <li>6. If the sentence feels awkward or unnatural, refine the structure by adjusting word choice or reordering components.</li> </ol> <p>&lt;/think&gt;</p>
--	--	---

表 4-5 候補案（プロンプト）

プロンプト①	要素 1 の内容をプロンプトに含め翻訳する
プロンプト②	要素 2 の内容をプロンプトに含め翻訳する
プロンプト③	プロンプト①とプロンプト②の内容をプロンプトに含め翻訳する
プロンプト④	プロンプト③と要素 3-1 の内容をプロンプトに含め翻訳する
プロンプト⑤	プロンプト③と要素 3-2 の内容をプロンプトに含め翻訳する
プロンプト⑥	プロンプト③と要素 3-3 の内容をプロンプトに含め翻訳する
プロンプト⑦	プロンプト③と要素 3-4 の内容をプロンプトに含め翻訳する
プロンプト⑧	プロンプト③と要素 3-5 の内容をプロンプトに含め翻訳する
プロンプト⑨	プロンプト③と要素 3-6 の内容をプロンプトに含め翻訳する

#### 4.4.3 翻訳試行について

プロンプトを工夫した手法の選定にあたり、公開情報調査の結果に加え翻訳タスクへの向き不向きを確認するため、特許文を使用した翻訳を試行した。試行翻訳では、プロンプトは表 4-5 に示す 9 種類のプロンプトを使用した。プロンプトで使用するモデルは、「公開情報調査（モデル）」（2 章）で調査対象としたモデル（Gemini、Claude、Gemma）を使用した。

評価文は、「公開情報調査（モデル）」（2 章）の翻訳試行で使用したのと同じ英日データを使用した。

#### 4.4.4 翻訳試行結果

試行翻訳結果を自動評価手法 BLEU で評価した。評価結果には、比較対象とするベースラインとして、簡易なプロンプト（翻訳者である旨のみを指定）の評価結果も併記した(表 4-6)。

表 4-6 試行翻訳の評価結果  
(太字は各モデルの 1 位の BLEU 値を示す。)

モデル	Gemini		Claude		Gemma	
	BLEU	順位	BLEU	順位	BLEU	順位
ベース	46.04	5	45.81	8	33.99	6
プロンプト①	48.83	3	48.86	3	37.06	2
プロンプト②	51.03	2	<b>49.97</b>	1	35.63	3
プロンプト③	<b>51.3</b>	1	46.84	7	<b>37.46</b>	1
プロンプト④	45.31	6	47.49	5	32.48	10
プロンプト⑤	33.74	8	47.32	6	33.39	9
プロンプト⑥	28.24	10	45.63	10	34.07	5
プロンプト⑦	45.1	7	45.8	9	35.41	4
プロンプト⑧	47.12	4	49.22	2	33.45	8
プロンプト⑨	30.55	9	48.83	4	33.47	7

#### 4.5 プロンプトの決定

評価結果を元に、「プロンプトを工夫した翻訳で使用するモデル」、「調査対象とするプロンプト」について検討を行った。

##### 4.5.1 プロンプトを工夫した翻訳で使用するモデル

各モデルは以下に示す評価結果となった(表 4-7)。

表 4-7 モデル検討概要

モデル	評価結果の概要
Gemini	プロンプト③の精度が最も高く、プロンプト①、②、③、⑧でベースラインの精度を上回った。一方、プロンプト⑤、⑥、⑨は BLEU 値が 30 ポイント台と、他のプロンプトに比べ 10 ポイント以上低い結果 <sup>7</sup> となった。
Claude	プロンプト②の精度が最も高く、プロンプト⑥、⑦以外の 7 つのプロンプトでベースラインの精度を上回った。
Gemma	プロンプト③の精度が最も高く、プロンプト①、②、③、⑥、⑦でベースラインの精度を上回った。

上記モデル毎の傾向から、いずれのモデルも一定のプロンプトで精度の向上がみられた。特に Claude は、7 つのプロンプトで精度向上が見られ、プロンプトへの対応度合いが高いと予想される。各モデルの 1 位の BLEU 値を見ると Gemini、Claude、Gemma の順となった。Gemini と Claude の差は 1.33 と比較的小さい一方で、Gemini と Gemma とは 13 ポイント以上の差となった。BLEU 値が低いものを見ると Gemini では、翻訳結果に CoT の過程を含むものが 3 つのプロンプトで見られた。以上から、プロンプトへの対応度が高く、一定の翻訳精度ができるモデルとして Claude をプロンプト工夫で使用するモデルとした。

#### 4.5.2 調査対象とするプロンプト

Claude の評価結果を参考に、調査対象とするプロンプトを検討した。対象とするプロンプトと検討結果を以下に示す(表 4-8)。

表 4-8 調査対象プロンプト

プロンプトの種類	プロンプト	検討概要
項目指定プロンプト	プロンプト①	項目指定のプロンプトは本調査の仕様で規定されたものであるため、調査対象とした。
プロンプトの工夫(5 プロンプト)	プロンプト②、④、⑧、⑨	BLEU 値の大きい順に 4 つを調査対象とした。
	プロンプト③	プロンプト③は、上記プロンプト④、⑧、⑨で組み合わせるプロンプトであり、評価結果の比較・分析に有用と考え調査対象とした。

<sup>7</sup> 翻訳結果内に CoT の過程が混在した文が存在したために BLEU 値が大きく低下したと考えられる。

## 5. 公開情報調査(翻訳精度評価手法)

### 5.1 調査概要

民間事業者、又はアカデミアが提供する大規模言語モデルを使用した翻訳精度評価手法について、調査を行った。調査結果に基づき、翻訳精度の評価（後述の8章）で対象とする評価手法を優先順位及び理由とともに特許庁担当者に提示し、協議の上、合計3つを調査対象とした。

### 5.2 調査範囲

項番 2.2 の調査範囲と同様の範囲で調査を行った。

### 5.3 調査結果と翻訳精度評価手法候補

前述調査範囲に対して、「LLM」、「trans」、「as a judge」、「prompt」といった単語を用いた検索照会を行い、その結果、計10件の論文を抽出した。また、特許文献についても「翻訳」、「LLM、大規模言語」の近傍検索に対し、「前処理・後処理」のシソーラス、「評価」、「プロンプト」を組み合わせた検索を行い、翻訳精度評価処理に関する3件の文献を抽出した。

抽出した論文及び特許文献に対して、表 5-1 の調査観点で調査を行った(表 5-1)。

表 5-1 調査観点(翻訳精度評価)

観点	説明
ドメイン	公開情報の手法がどのような分野の文書に対して用いられている手法であるか
対象言語	どの言語方向へ用いられているものであるか
実装方法	ツールとして公開又は github において公開されているかどうかの情報、プログラム・プロンプトが公開されている場合はその言語
オンプレミス対応可否	翻訳の前後処理やプロンプトがオンプレミスのモデルにおいて利用可能か
提供状況	提供開始済みもしくは、提供開始が予定されているか（その場合は開始時期）
ライセンス条件	利用する際の費用、計算コスト、ライセンス条件
費用	利用時に必要となる費用

評価精度	評価手法を選定する上で非常に重要な観点。翻訳精度評価手法の精度としては、人間による評価との相関を用いることが一般的であるため、本事業においても、公開情報において得られる限り、人間による評価との相関を主たる評価精度と考えて調査した。
評価観点	LLM による精度評価には、1つの軸で翻訳品質の良し悪しを評価するものや、正確性、流暢さなど多観点で翻訳品質の良し悪しを評価するものが存在する。そのため、どのような評価観点で評価を行っているかを調査した。
強みと課題	強みや課題を整理することで、特許文書における翻訳について適切な評価ができるかどうか検討の材料とした。

上記調査観点による調査の結果前後処理は以下のタイプに類型化することができた。

- ① 誤訳箇所アノテーション（5件）
- ② 評価ライブラリ<sup>8</sup>（3件）
- ③ 訳質採点（1件）
- ④ xCOMETT での誤訳抽出及び xTower での再翻訳<sup>9</sup>（1件）

## 5.4 翻訳精度評価手法候補の評価

### 5.4.1 候補の優先順位付け

調査対象の翻訳精度評価手法を選定するため、調査結果を以下の観点で評価(表 5-2)し、評価結果を元に順位付けを行った。順位付けは、各観点を○△×で評価し、○が3点、△1点、×0点とすることで候補を選択した。その際、同点のものについては手法の内容・選択済の手法との類似性を考慮して順位を付けた。優先順位の上位5件を表 5-3 に示す。

表 5-2 評価観点(翻訳評価手法)

評価精度	調査において得られた評価精度が高いものを選定
適正	強み・課題、対象文書のドメインや翻訳方向から、特許文書の翻訳の評価や日本語の翻訳評価に適しているか

<sup>8</sup> 評価ライブラリとは評価をユーザが作成したプログラムから利用できる形式で提供するものを指す。

<sup>9</sup> xCOMET は機械翻訳の品質評価を行うツール。但し LLM とは異なるモデルを使用している。xTower は xCOMET の出力を利用して誤訳の説明や翻訳の修正案を提示するツールを指す。

実現性	プロンプトや実装コードが明確となっていること、製品であればサービス化されているものから優先して候補として提案。現実的な費用、時間で使用可能か否かも考慮。
評価観点	多観点でかつカスタマイズ性が高いか

表 5-3 評価結果上位(翻訳評価手法)

タイプ	論文名	概要	点数	順位	理由
アノテーション	MQM-APE: Toward High-Quality Error Annotation Predictors with Automatic Post-Editing in LLM Translation Evaluators	Annotator が翻訳文に対する誤りを抽出し、Impact Evaluator が評価（メジャー、マイナー等）し、Automatic Post-Editor が修正（APE）、オリジナルが良いか APE 結果が良いかを Pairwise Quality Verifier が評価して、品質改善に貢献するエラーセットのみ残しそれを採用。また、当該エラーセットに基づいてスコア付け。	12	1	コード、プロンプトが提供されており実現性が高い。また、評価観点も特許に適用可能。
採点	Large Language Models Are State-of-the-Art Evaluators of Translation Quality	llm にプロンプトと共に、原文、機械翻訳文、参照文をインプットして評価をさせる。0-100 の評価や 5 段階評価をする手法、No meaning preserved・Perfect translation といったカテゴリに分類させる手法が記載。	9	2	prompt のみで実現可能であり、単純な評価手法の一例として有用と判断。
アノテーション	TEaR: Improving LLM-based Machine Translation with Systematic Self-Refinement	翻訳結果を、accuracy, fluency, locale convention, terminology, non-translation, other, no-error 等 estimate (critical, major, minor 等の判断も) し、MQM <sup>10</sup> アノテーションする。その後 Refine。	9	3	python 等プログラムが提供されており、実現性がある。
アノテーション	GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4	優先順位 1 位の文献と同じ筆者。1 位の文献の Annotator 相当。	8	4	プログラム提供されているものの、日本語を対象としておらず、試行錯誤が必要なおそれ。実現性を△とした。

<sup>10</sup> MQM（多次元品質評価指標）は、翻訳品質を多角的に評価するための国際的なフレームワークを指す。https://themqm.org/

ライブラリ	promptfoo	LLM の出力品質を評価用 LLM によって自動評価するための CLI およびライブラリ。「任意の単語が含まるか」「出力が冗長でないか」「出力が面白い」のように事前に定義したテストケースを利用して、プロンプト、モデル、RAG を体系的にテストできる。	採点 不能	5	参考サイト多数あり。観点のカスタマイズ可能。
-------	-----------	---	----------	---	------------------------

#### 5.4.2 優先度を考慮した候補の検討

調査結果の論文の記載内容は、必ずしもすべてが本事業の調査対象とすべき事項というわけではない。調査結果の論文から調査対象とすべき事項を抽出し、本事業での候補案を検討した(表 5-4)。

表 5-4 候補案（翻訳精度評価手法）

翻訳精度評価手法	概要
評価手法①	<p>原文と翻訳文とを入力し、翻訳文に対して誤りを抽出しアノテーションを実施 (accuracy, fluency といった観点、critical, major といったエラーカテゴリ)</p> <ul style="list-style-type: none"> <li>・ベース論文：MQM-APE: Toward High-Quality Error Annotation Predictors with Automatic Post-Editing in LLM Translation Evaluators (優先順位 1 位)</li> <li>・目的：LLM による MQM の有用性を調査</li> <li>・LLM 評価自体の評価：評価結果については、誤り抽出の各観点と人手評価観点とを可能な限り対応づけた上で定量化し、人手評価結果(数値)との比較を実施。また、誤り抽出結果全体を定量化することで機械評価結果との比較を実施。</li> </ul>
評価手法②	<p>LLM に原文と翻訳文、人手翻訳文とを入力し、「正確性 (文)」を 5 段階評価、「正確性 (用語)」を 4 段階評価、「流暢さ」を 5 段階評価</p> <ul style="list-style-type: none"> <li>・ベース論文：Large Language Models Are State-of-the-Art Evaluators of Translation Quality (優先順位 2 位)</li> <li>・目的：「特許文献機械翻訳の品質評価手順 (ver1.0)」と、LLM による評価の対応性の調査</li> <li>・LLM 評価自体の評価：評価結果は人手評価と一致させているため、人手評価との比較は容易に実施可能。また、3 観点の合計評価と、機</li> </ul>

	械評価結果との比較を実施。
評価手法③	LLM に原文と翻訳文、人手翻訳文とを入力し 0~100 で評価 <ul style="list-style-type: none"> <li>・ベース論文：Large Language Models Are State-of-the-Art Evaluators of Translation Quality（優先順位 2 位）</li> <li>・目的：敢えて簡便な手法の有用性の調査</li> <li>・LLM 評価自体の評価：評価結果については、モデルが出力する点数と、人手評価人手評価結果合計点との比較を実施。同様にモデルが出力する点数機械評価結果との比較を実施。</li> </ul>

### 5.4.3 評価試行について

翻訳精度評価手法の調査対象選定にあたり、翻訳精度評価手法の候補について評価試行を行った。

評価試行は、前述評価手法①~③を用いて、英日の短文~長文を含むコーパス 全 41 文（タイトル 11 文、請求項 10 文、明細書 10 文、要約 10 文）の翻訳結果を評価した。使用する LLM モデルは、翻訳結果には Claude を、評価手法には Gemini を使用した。

### 5.4.4 評価試行結果

各手法の評価内容は以下に示す内容となった(表 5-5)。

表 5-5 何らかの誤訳指摘・減点ありの存在数

評価手法①	評価手法②	評価手法③
10	13	20

誤訳指摘・減点について評価誤りの存在数を確認したところ、以下の結果となった(表 5-6)。

表 5-6 評価誤り（※1）の存在数

	評価手法①	評価手法②	評価手法③
評価誤り	7	8	7
内) 指摘漏れ	2	1	0
内) 指摘過剰	5	7	7

(※1) 評価①～③で何らかの減点が入ったものについて、正当と思われる評価とずれが生じるものを「評価誤り」とした。ただし、評価③について96点以上のものは、誤りとすべきか否かの判断が困難であるため除外した。

以下に各評価手法の評価結果のサンプルを示す。

#16

原文	The compositions and methods are also useful in <u>treating and preventing</u> lung infections; and infections in burns and wounds; and other infections caused by biofilms.
参照訳	本組成物および方法はまた、肺感染症、熱傷および創傷における感染症、ならびにバイオフィルムによって引き起こされる他の感染症の <u>治療および予防に有用</u> である。
機械翻訳	本組成物および方法は、肺感染症の <u>治療および予防</u> 、火傷や創傷における感染症、ならびにバイオフィルムによって引き起こされるその他の感染症にも有用である。
評価①	指摘無し
評価②	<p>“sentence-level accuracy”: “5 stars: All important information is conveyed accurately. The translation correctly identifies that the compositions and methods are useful for treating and preventing three types of infections. The meaning is fully preserved.”,</p> <p>“word-level accuracy”: “3 stars: The translation for ‘burns’ is ‘火傷’ (yakedo), which is a common term. However, the human reference’s ‘熱傷’ (nesshou) is a more technical and formal term appropriate for a patent document. ‘火傷’ is an acceptable translation, but less precise in this context. All other key terms are translated correctly.”,</p> <p>“fluency”: “5 stars: The sentence is grammatically correct, clear, and reads naturally. The use of commas to enumerate the different types of infections is a common and fluent structure in Japanese.”</p>
評価③	<p>This translation is of high quality, but it contains a subtle yet important structural difference from the source and the human reference.</p> <p>* **Source/Reference Structure:** “useful in treating and preventing [A, B, and C]” (A, B, C の治療および予防に有用である). This structure clearly applies “treating and preventing” to all three listed items: (A) lung infections, (B) infections in burns and wounds, and (C) other infections caused by biofilms.</p> <p>* **Translation to Score’s Structure:** “useful in [treating and preventing A], [B], and [C]” (A の治療および予防、B、ならびに C にも有用である). <b><u>This structure explicitly links “treating and preventing” (治療および予防) only to the first item, “lung infections” (肺感染症).</u></b> While a reader might infer that “treating and preventing” also applies to the other items, it is not explicitly stated. This introduces a slight ambiguity that is not present in the original English or the human reference. In a patent document, such precision is crucial.</p> <p>Additionally, the use of 「火傷や創傷」 is slightly less formal than the reference’s 「熱傷および創傷」. 「熱傷」 is the more technical term for “burn” used in medical and scientific contexts, and 「および」 is standard for lists in formal writing.</p> <p>Despite these points, the overall meaning is largely conveyed, and the translation is grammatically correct and fluent. The error is one of precision rather than a complete misunderstanding.</p> <p>[[[80]]]</p>
判定	文脈からすると、「治療および予防」は全体に係るべきである。評価③ではその点を指摘できているが評価①、②は認識できていない。そのため、 <u>評価①、②は誤り</u> といえる。

原文	The apparatus according to claim 1, wherein the struts of the first and the second subsets are superelastic.
参照訳	前記の第1および前記の第2のサブセットの前記ストラットが超弾性である、請求項1に記載の装置。
機械翻訳	請求項1に記載の装置において、第1および第2のサブセットのストラットが超弾性である装置。
評価①	<p>"error_category": "fluency",  "subcategory": "grammar",  "error_level": "major",  "reason": "The translation ends with a noun phrase ¥"...である装置¥" after starting with ¥"～において¥", which makes the sentence grammatically incomplete. The phrase ¥"請求項1に記載の装置において¥" (In the apparatus according to claim 1) sets up a context, but what follows is not a complete predicate. It's a noun phrase describing an apparatus, resulting in a sentence fragment. <b>A correct structure would be something like 「請求項1に記載の装置であって、...である装置。」 or 「請求項1に記載の装置は、...である。」. The repetition of ¥"装置¥" (apparatus) is also redundant.</b>"</p>
評価②	<p>"sentence-level accuracy": "5/5 stars. All important information is conveyed accurately. The translation correctly captures that the device is based on claim 1 and that its energy delivery element and sensor are designed to retract axially into the catheter member.",  "word-level accuracy": "4/4 stars. All key technical terms are translated appropriately and are commonly used in the field. ¥"energy delivery element¥" as 「エネルギー送達要素」, ¥"sensor¥" as 「センサ」, ¥"axial retraction¥" as 「軸方向に後退する」, and ¥"catheter member¥" as 「カテーテル部材」 are all suitable translations.",  "fluency": "4/5 stars. The meaning is clear, but the sentence structure is slightly unnatural for a patent claim. The phrasing 「～装置であって、～装置。」 (a device, which is a device that...) is redundant. The human reference's structure 「～構成される、～装置。」 is more concise and standard in Japanese patent drafting."</p>
評価③	<p>This translation has a significant grammatical and stylistic flaw.</p> <p>1. <b>**Meaning:**</b> The core technical meaning is correctly translated. "The struts of the first and the second subsets are superelastic" is accurately conveyed as 「第1および第2のサブセットのストラットが超弾性である」.</p> <p>2. <b>**Grammar/Style:**</b> <b>The structure 「請求項1に記載の装置において、...装置。」 ("In the apparatus according to claim 1, ... an apparatus.") is redundant and grammatically awkward</b> in the context of a patent claim. The word 「装置」 (apparatus) appears at both the beginning and the end of the sentence, which is incorrect. The human reference uses a standard and correct adjectival clause structure ( `...である、...装置。` ) to modify the base noun. The translation's structure is a common error but is considered poor form in professional patent translation. While the technical information is preserved, the grammatical structure is flawed, making it sound unnatural and unprofessional.</p> <p>[[[60]]]</p>
判定	「請求項1に記載の装置において、...装置。」という構成は、特許請求項の記載として有り得る範囲と考えられる。「請求項1に記載の装置であって、...装置。」の方が適切な可能性があり、少しの減点は許容されるとしても、評価①と評価③は厳しすぎると考えられる。なお、参照訳に存在する「前記」について各評価手法では触れていない。

## 5.5 翻訳精度評価手法の決定

評価手法①、②、③の試行について、評価誤りの出現数は概ね同等であり、問題は生じて

いないことから、評価手法①、②、③を使用した評価を調査対象とした。各手法は、それぞれ異なる特性（カテゴリの出力有無・種類、指摘漏れ・指摘過剰の傾向等）を有し、いずれも検証に値する手法と考えられる。

また、評価で使用するモデルについて、本調査では「前後処理を工夫した翻訳」や「プロンプトを工夫した翻訳」は Claude で翻訳することから、翻訳と異なるモデル Gemini を使用して評価することとした。

## 6. 対訳コーパス又は辞書の準備

本調査では、下記表 6-1 に示した言語方向及び項目の翻訳を実施する。以下の 6.1～6.4 の要領で、翻訳結果の精度の評価を行うための評価用データ及び追加学習用の学習データとしての対訳コーパス又は辞書を作成した。

表 6-1 翻訳言語方向及び翻訳対象項目

言語方向	対象項目
英中韓→日	特許請求の範囲、明細書、要約、発明の名称
日→英	特許請求の範囲、明細書、要約、発明の名称、 出願人・発明者の名称、拒絶理由通知書、意見書、補正書

### 6.1 特許公報の対訳コーパス

特許庁が貸与する英語、中国語、韓国語、日本語公報（XML ファイル）から、英語、中国語、韓国語公報のファミリーに含まれる日本語公報を特定した。ファミリーリストの特定は、英語、中国語、韓国語公報の書誌情報及び優先権情報、特許庁が貸与する DOCDB データ、PATENTSCOPE 等の公開情報、その他入手可能な情報に基づき実施した。

特定したファミリー情報及び各言語の特許文献の記載に基づき、言語毎に特許請求の範囲、明細書、要約、及び発明の名称の各項目について、それぞれ 2 0 0 0 文対以上の対訳コーパスを作成した。作成した対訳コーパスについて、人手による確認を行い、2 0 0 0 文対の対訳コーパスを作成した。

### 6.2 拒絶理由通知等の対訳コーパス

特許庁が貸与する日本語の拒絶理由通知書、意見書、補正書のデータから、英語翻訳文を作成し、これらに基づいて拒絶理由通知書、意見書及び補正書の各項目について、それぞれ 2 0 0 0 文対以上の対訳コーパスを作成した。翻訳文の作成にあたっては、人手による翻訳を行った。

### 6.3 出願人名・発明者の日英対訳辞書

特許庁が貸与する出願人発明者の日本語名と英字名の対応表（出願人発明者対応表）から、

日本語と英語の対訳を抽出し、これらに基づいて、2000文対の対訳辞書を作成した。

## 6.4 学習データ及び評価用データの作成

作成した対訳コーパス又は辞書について、各言語各項目500文対程度を評価用データとし、残りの文対を追加学習用の学習データとした。

また、上記評価用データから、100文対以上を人手評価対象文として選定した。選定に際しては、特許庁の過去の調査事業の報告書などを参考にしつつ、特許庁担当者と選定方針について相談の上、本調査分野に専門的な知見を有する者により選定を行い、特許庁担当者の承認を得た上作成した。選定に際しては、特許分類、文字数の観点、文の種類（請求項・明細書・要約等）の観点を考慮し、偏りが無いよう選定した。

## 7. 各種モデルによる翻訳

### 7.1 大規模言語モデルによる翻訳

「公開情報調査（モデル）」（2章）で調査対象とした3つのモデルについて、英語、中国語、韓国語から日本語への翻訳（以下、「英中韓翻訳」という。）及び日本語から英語への翻訳（以下、「日英翻訳」という。）を行った。英中韓翻訳は、各言語の特許文献の特許請求の範囲、明細書、要約、及び発明の名称の各項目について、「対訳コーパス又は辞書の準備」（6章）で作成した各項目少なくとも500文の評価用データの翻訳を実施した。また、日英翻訳は、特許請求の範囲、明細書、要約、及び発明の名称の各項目に加えて、「対訳コーパス又は辞書の準備」で作成した拒絶理由通知書、補正書、意見書及び、「対訳コーパス又は辞書の準備」で作成した出願人名、発明者名、について、各項目少なくとも500文の評価用データの翻訳を実施した。

#### 7.1.1 翻訳で使用するプロンプトについて

「大規模言語モデルによる翻訳」で使用するプロンプトについてプロンプトの試行錯誤を行った。プロンプトのポイントは以下の観点1及び2である。

##### 観点1：対象が特許文書であることの明確化

前後処理の翻訳試行(3.4.3)、プロンプトを工夫した翻訳試行(4.4.3)で用いたベースラインのプロンプトと比較し、特許文（patent sentence）であることを明示したプロンプトとした。

##### 観点2：翻訳結果以外の回答防止

実用化にあたっては、LLMからの回答は翻訳結果だけであることが望ましい。そのため、プロンプトにおいて以下の指示内容を追加することとした。具体的には下記の内容をプロンプトに含めた。

No breakdown or explanation needed. Please show me only translation enclosed in [[[ and ]]].
---

#### 7.1.2 プロンプトの試行錯誤結果

前後処理の翻訳試行(3.4.3)、プロンプトを工夫した翻訳試行(4.4.3)で用いたベースライン

のプロンプト（以下、「活用手法 Baseline」と記載）と、本試行錯誤で使用したプロンプト（以下「評価文翻訳 Baseline」と記載）について具体的なプロンプト及び翻訳試行で用いた試行用データによる BLEU 値を示す(表 7.1-1)。

(プロンプト)

・活用手法 Baseline

You are a translator. Please translate the following 翻訳元言語 **text** into 翻訳先言語. No breakdown or explanation needed. Please show me only translation enclosed in [[[ and ]]].  
 翻訳元言語 text: {source\_text}

・評価文翻訳 Baseline

You are a translator. Please translate the following 翻訳元言語 **patent sentence** into 翻訳先言語. No breakdown or explanation needed. Please show me only translation enclosed in [[[ and ]]].  
 翻訳元言語 text: {source\_text}

表 7.1-1 評価文翻訳 Baseline の BLEU

Model	活用手法 Baseline	評価文翻訳 Baseline
<b>Gemini</b>	46.04	<b>48.30</b>
<b>Claude sonnet</b>	45.81	<b>46.78</b>
<b>gemma</b>	33.99	<b>37.14</b>

評価文翻訳 Baseline の BLEU は、活用手法 Baseline の BLEU に比べ、全てのモデルでスコアが増加し、特許文用のプロンプトとして、問題がないことが確認できた。

以上から、評価文翻訳に用いる Baseline や、活用手法で用いるプロンプトについて、「text」を「patent sentence」とすることで、翻訳対象が特許文であることを明示した内容にて翻訳を実施した。

### 7.1.3 パラメータ調整対象

翻訳を実施するモデルである Claude sonnet を用いて、temperature（出力の多様性やランダム性を調整するためのパラメータ）、Top-p（上位何%までの出力を回答候補として考慮するかを定めるパラメータ）を変化させることで BLEU 値の変化を調査した(表 7.1-2～表 7.1-4)。

### 7.1.3 パラメータ変化結果

表 7.1-2 ①temperature を変化させた例

temperature	top-p	BLEU
default	default	45.81
0	default	46.32
0.1	default	47.30
0.2	default	47.57
0.3	default	46.51
0.5	default	46.61
0.7	default	44.43
1	default	44.57

表 7.2-3 top-p を変化させた例

temperature	top-p	BLEU
default	default	45.81
default	0	44.40
default	0.2	46.34
default	0.4	46.32
default	0.5	46.38
default	0.6	46.78
default	0.7	44.90
default	1	45.37

表 7.2-4 top-p 及び temperature を変化させた結果

temperature	top-p	BLEU
default	default	45.81
0.1	0.5	46.32
0.2	0.5	46.32
0.1	0.6	46.34
0.2	0.6	46.52

以上より、評価文の翻訳を行う際には、表 7.1-2～表 7.1-4 で最も BLEU 値が高かった temperature のみを default 値から 0.2 に変更し top-p は default 値とするパラメータ変更を採用することとした。

なお Gemma, Gemini についても同様の調査を行った結果、Gemma は temperature の値を 1.2 (Claude と Gemma では temperature のとり得る範囲が異なる) とした場合に BLEU 値が最大であった。Gemini については top-p を 0.5 とすることで BLEU 値が最大となった。それぞれ、評価の翻訳を行う際のパラメータ値として採用し翻訳を実施した。

## 7.2 プロンプトを工夫した翻訳

翻訳にあたって、表 6-1 に示した翻訳対象項目 (特許請求の範囲、明細書等) のいずれを対象とする翻訳であるかを指定したプロンプト、及び、「公開情報調査(プロンプト)」(4 章) の調査により得られた翻訳において有用なプロンプトの調査対象候補のうち少なくとも 8 つを選定して翻訳を実行した。各プロンプトを用いた翻訳の実行にあたっては、最適なモデルをそれぞれ 1 つ選定するとともに、当該プロンプトの有用性を比較検討するために効果的であると考えられる言語方向・項目をプロンプト毎に少なくとも 4 つ選定し、選定した言語方向・項目の 500 文についてそれぞれ翻訳を行った。

なお、プロンプトの選定に際しては、特許庁担当者と選定方針について相談の上、本調査分野に専門的な知見を有する者により選定を行い、特許庁担当者の承認を得た上で、実施した。

### 7.2.1 翻訳対象とする言語方向、項目について

「公開情報調査(プロンプト)」(4 章) で調査対象としたプロンプト①、②、③、④、⑧、⑨について、プロンプトを工夫した翻訳対象とする言語方向、項目について検討を行った。

- (1) 提案したプロンプトはいずれも、言語方向に依存しない内容であることから多くの言語方向での調査が有効と考え、英日、中日、韓日、日英を対象とする。
- (2) 項目指定のプロンプトは、請求の範囲、明細書、要約、発明の名称、拒絶理由通知書、意見書、補正書を翻訳対象として指定した 7 種のプロンプトを対象とする。7 種類のうち、請求の範囲、明細書、発明の名称の 3 種のプロンプトを要求仕様で求めるプロンプトとし自動評価、人手評価を行うこととし、その他 4 種のプロンプトを追加プロンプトとして自動評価のみ実施を行うこととする。
- (3) プロンプトの工夫は、プロンプト②は技術分野をプロンプトに含めたプロンプトである。その他のプロンプトはその内容から比較的複雑な文を対象としたプロンプトと考えられる。この点を踏まえプロンプト②は「発明の名称」を対象とし、その他のプロンプトについては、特許文書の主要項目である「請求の範囲」、「明細書」を対象項目とした。

以上の検討結果を元に、各プロンプトが対象とする言語方向、項目をまとめた(表 7-1、表 7-2)。

表 7-1 調査対象一覧 (プロンプト①[項目指定プロンプト])

項目	言語方向			
	英日	中日	韓日	日英
請求の範囲	○	○	○	○
明細書	○	○	○	○
発明の名称	○	○	○	○
要約	□	□	□	□
拒絶理由	-	-	-	□
意見書	-	-	-	□
補正書	-	-	-	□

○：仕様で要求されたプロンプト (12 種類)

□：追加プロンプト<sup>11</sup>(7 種類)

表 7-2 調査対象一覧 (プロンプトの工夫)

プロンプト	言語方向	項目			
		請求の範囲	明細書	発明の名称	要約
プロンプト②	英日、中日、韓日、 日英	-	-	○	-
プロンプト③		○	-	-	-
プロンプト④		○	-	-	-
プロンプト⑧		-	○	-	-
プロンプト⑨		-	○	-	-

○：仕様で要求されたプロンプト (5 種類×4 言語方向=20 種類)

また、実際に使用するプロンプトの例を以下に示す(表 7-3)。

<sup>11</sup> 追加プロンプトは自動評価のみ実施する。

表 7-3 プロンプト例

プロンプト	プロンプト例	備考
プロンプト①	<p>Translate the following Chinese patent sentence into Japanese.</p> <p>Please fulfill the following conditions when translating.</p> <p>No breakdown or explanation needed.</p> <p>Please show me the translation enclosed in [[[ and ]]].</p> <p>[Purpose of the translation] Accurate translation of {text_kind}</p> <p>[Target audience] Patent searchers</p> <p>[source text] {source_text}</p>	<p>"Chinese patent sentence into Japanese"部分は翻訳する言語方向による。</p> <p>{text_kind} 翻訳の種類(invention titles, patent claims, patent description など)</p> <p>{source_text}</p> <p>原文</p>
プロンプト②	<p>Translate the following patent sentence from Chinese to Japanese.</p> <p>The text has [background].</p> <p>No breakdown or explanation needed. Please show me the translation enclosed in [[[ and ]]].</p> <p>[source text] {source_text}</p> <p>[background] {background}</p>	<p>{background}該当する文献の ipc の説明文</p>
プロンプト③	<p>Translate the following Chinese patent sentence into Japanese.</p> <p>Please fulfill the following conditions when translating.</p> <p>No breakdown or explanation needed.</p> <p>Please show me the translation enclosed in [[[ and ]]].</p> <p>[Purpose of the translation] Accurate translation of {text_kind}</p> <p>[Target audience] Patent searchers</p> <p>[background] {background}</p> <p>[source text] {source_text}</p>	

<p>プロンプト④</p>	<p>Translate the following Chinese patent sentence into Japanese by [Chain of Thought].</p> <p>Please fulfill the following conditions when translating.</p> <p>The final translation should be enclosed in [[[ and ]]].</p> <p>[Purpose of the translation] Accurate translation of {text_kind}. No translation in Romaji needed.</p> <p>[Target audience] Patent searchers</p> <p>[background] {background}</p> <p>[Chain of Thought]</p> <ol style="list-style-type: none"> <li>1. Analyze the sentence structure and identify the core elements (subject, verb, object).</li> <li>2. Translate the sentence from the origin language to the target language, focusing on the core elements.</li> <li>3. Review the translation for basic accuracy and grammatical structure.</li> <li>4. Identify areas that need further refinement (e.g., word choice, tense, or word order).</li> <li>5. Modify the translation to improve fluency and coherence, considering the context.</li> <li>6. Finalize the translation by ensuring it retains the original meaning while improving readability.</li> </ol> <p>[source text] {source_text}</p>
---------------	---

<p>プロンプト⑧</p>	<p>Translate the following Chinese patent sentence into Japanese by [Chain of Thought].</p> <p>Please fulfill the following conditions when translating.</p> <p>The final translation should be enclosed in [[[ and ]]].</p> <p>[Purpose of the translation] Accurate translation of {text_kind}.</p> <p>No translation in Romaji needed.</p> <p>[Target audience] Patent searchers</p> <p>[background] {background}</p> <p>[Chain of Thought]</p> <ol style="list-style-type: none"> <li>1. Analyze the source sentence and identify the key elements (verbs, subjects, objects, etc.).</li> <li>2. Based on these elements, determine the most suitable translation strategy (literal vs. idiomatic).</li> <li>3. Select the best translation for each word or phrase, considering context and languagespecific structures.</li> <li>4. Explain the rationale behind choosing specific words or phrases.</li> <li>5. After completing the initial translation, review each translation decision and explain any adjustments made for fluency or accuracy.</li> <li>6. Provide a final explanation for the translation choices, discussing any trade-offs made between literal meaning and contextual appropriateness.</li> </ol> <p>[source text] {source_text}</p>
---------------	---

<p>プロンプト⑨</p>	<p>Translate the following Chinese patent sentence into Japanese by [Chain of Thought]. Please fulfill the following conditions when translating.</p> <p>The final translation should be enclosed in [[[ and ]]].</p> <p>[Purpose of the translation] Accurate translation of {text_kind}. No translation in Romaji needed.</p> <p>[Target audience] Patent searchers</p> <p>[background] {background}</p> <p>[Chain of Thought]</p> <ol style="list-style-type: none"> <li>1. Analyze the sentence's syntactic structure in the source language (e.g., identify whether it's active or passive).</li> <li>2. Determine the most appropriate syntactic structure in the target language (e.g., whether it needs to be rephrased from active to passive or vice versa).</li> <li>3. Adjust the word order and grammatical structure in the target language to match the sentence's meaning, while maintaining clarity.</li> <li>4. Translate the sentence, ensuring that subjectverb-object relationships and other syntactic elements align with target language norms.</li> <li>5. After the translation, check the sentence's grammar and overall flow in the target language, making sure it is clear and fluid.</li> <li>6. If the sentence feels awkward or unnatural, refine the structure by adjusting word choice or reordering components.</li> </ol> <p>[source text] {source_text}</p>	
---------------	---	--

### 7.3 前後処理を工夫した翻訳

「公開情報調査(前後処理)」(3章)において選定した前後処理を行った上で、翻訳を実行する。翻訳の実行にあたっては、最適なモデルをそれぞれ1つ選定するとともに、当該前後処理が有効と考えられる言語方向・項目を前後処理毎に少なくとも8つ選定し、選定した

言語方向・項目の500文について翻訳を行った。

なお、前後処理の選定に際しては、特許庁担当者と選定方針について相談の上、本調査分野に専門的な知見を有する者により選定を行い、特許庁担当者の承認を得た上で、実施した。

### 7.3.1 翻訳対象とする言語方向、項目について

前後処理を工夫した翻訳で対象とする言語方向、項目は前後処理の内容を考慮した上で以下を対象とした(表7-4)。

表 7-4 翻訳対象言語方向、項目（前後処理）

前後処理	言語方向、項目	理由
前後処理①	言語方向：英・中・韓・日 項目：特許請求の範囲・明細書	当該手法は翻訳対象の文中に含まれる重要語を抽出するものであり、当社保有コーパスをより活かすことが可能な特許請求の範囲・明細書を対象とする。
前後処理②	言語方向：英・日 項目：英は特許請求の範囲・明細書・要約・発明の名称 日は拒絶理由通知、意見書、補正書、出願人・発明者の名称	当該手法は翻訳対象文を基に参照用の文対をLLMに作成させるものであり、幅広く適用可能であるため、英日の公報由来の各項目、日英の特許書類、出願人・発明者の名称を対象とする。
前後処理⑤	言語方向：英・中・韓・日 項目：特許請求の範囲・明細書	前後処理①同様、当社保有コーパスをより活かすことが可能な特許請求の範囲・明細書を対象とすることとした。

また、実際に使用するプロンプトの例を以下に示す(表7-5)。

表 7-5 プロンプト例

プロンプト	プロンプト例	備考
前後処理①	You are tasked with translating Chinese to Japanese. You are provided several example translations, and you should refer to the examples to translate the given Chinese patent sentence. No breakdown or explanation needed. Please show me the final translation enclosed in [[[ and ]]]. [example translations in JSON format]	{example}には、原文中の重要用語を含む対訳の例文が設定される。

	{example} [source text] {source_text}	
前後処理②	同上	{example}には、LLMにより作成された参照用文対が設定される。
前後処理⑤	同上	{example}には原文に類似する文が設定される。

## 7.4 追加学習

「公開情報調査（モデル）」（2章）で調査対象としたモデルのうち、オンプレミス環境で動作可能なオープンモデル Gemma に対して、「対訳コーパス又は辞書の準備」（6章）で作成した学習データで追加学習を行った。学習データの文数は「発明の名称」「要約」「特許請求の範囲」「明細書」「拒絶理由通知書」「意見書」「補正書」それぞれ 1,500 文である。作成した学習データでの精度向上が見られない可能性を想定し、当社保有の対訳コーパスをさらに 1,000 文追加し学習したモデル作成も行った<sup>12</sup>。

本事業では学習対象の対訳コーパスが 1,500 又は 2,500 文と少数であることから、特許文書の翻訳ドメイン向けに追加学習する手法として、大規模言語モデルの微修正作業であるファインチューニングを採用した。

大規模言語モデルのファインチューニングには、モデル全体の重みを調整する「Full Fine-tuning」と、一部のパラメータのみを調整する「Parameter-Efficient Fine-Tuning (PEFT)」があるが、数十億規模のパラメータを持つ大規模言語モデルを少量データで Full Fine-tuning すると過学習（学習データの本質的でないノイズのようなデータにまで過剰に適合した学習）となる可能性が高く、また壊滅的忘却（事前学習時に獲得した一般的な言語知識を失う）の可能性もあるため PEFT によるファインチューニングを行うこととした。

<sup>12</sup> さらに追加した対訳コーパスは英日の特許請求の範囲明細書コーパスとなる。

なお、PEFT には、LoRA、Adapter、Prefix Tuning など、さらに細分化された手法があるが、元のモデルの重み  $W$  に対して並列に  $\Delta W$  を追加し、 $\Delta W$  のみを学習する手法である LoRA を用いた。当該手法は Adapter 等他の手法よりもパラメータの更新方式や、モデルへの影響が少なく、性能的にも他の手法より優れていることが多いため、PEFT の主流となっている。

## 7.5 追加学習済みモデルによる翻訳

上記 7.4 で追加学習したモデルにより、「対訳コーパス又は辞書の準備」(6 章) で作成した評価用データを用いて、各言語方向・各項目それぞれ少なくとも 500 文の翻訳を実施した。

## 7.6 ニューラル機械翻訳エンジンによる翻訳

既存のニューラル機械翻訳エンジンにより、「対訳コーパス又は辞書の準備」(6 章) で作成した評価用データを用いて、各言語方向・各項目それぞれ少なくとも 500 文の翻訳を実施した。

なお、使用するニューラル機械翻訳エンジンの選定に際しては、特許庁担当者を選定方針について相談の上、本調査分野に専門的な知見を有する者により選定を行い、特許庁担当者の承認を得た上実施した。具体的には審査書類以外の翻訳では当社保有ニューラル機械翻訳エンジンを、審査書類については特許庁から貸与された対訳コーパス(拒絶理由、意見書)を学習したニューラル機械翻訳エンジンを使用した。

## 8. 翻訳精度の評価

各種モデルによる翻訳（7章）で翻訳した結果に対し、自動評価、人手評価、および大規模言語モデルによる評価にて翻訳精度を評価した。自動評価手法は BLEU、RIBES を使用した。人手評価は、特許庁が公表している「特許文献機械翻訳の品質評価手順（ver1.0）」の品質評価手順に則り「内容伝達レベルの評価」、「重要技術用語の評価」、「流暢さの評価」を実施することに加え、「誤訳カテゴリ別チェック」による評価を実施した。大規模言語モデルによる評価は、公開情報調査(翻訳精度評価手法)(5章)で決定した評価手法による評価を実施した。

### 8.1 自動評価

対訳コーパス又は辞書の準備(6章)で作成した評価用データ（言語方向、項目毎に各 500 文）の機械翻訳結果について、自動評価手法 BLEU、RIBES による評価を行った。言語方向毎の評価結果を表 8.1-1 から 8.1-5 に示す。また、表中「大規模言語モデルによる翻訳」より、同モデルで BLEU 値が 1.5 ポイント以上上がった場合を青字に、1.5 ポイント以上下がった場合を赤字としている。同様に、RIBES 値が 1.5 ポイント以上上がった場合を青字に、1.5 ポイント以上下がった場合を赤字としている。さらに、項目毎の翻訳精度を確認するために、項目別に評価結果を集計した(表 8.1-6, 8.1-7)。そして、これらの評価結果を元に大規模言語モデル 3 モデルの精度比較、前後処理の適用有無による精度比較、プロンプトの工夫有無の翻訳精度を比較した。

## 8.1.1 英日

表 8.1-1 自動評価結果(英日)

項目	発明の名称		要約		請求項		明細書	
翻訳の種類/手法/モデル	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES
<b>大規模言語モデルによる翻訳</b>								
Claude	49.02	88.34	50.13	86.56	52.84	61.19	53.30	88.05
Gemini	50.82	89.20	50.25	86.47	56.31	80.97	54.36	87.87
Gemma	39.74	83.18	41.14	82.75	42.14	58.45	44.17	84.15
<b>前後処理を適用した翻訳</b>								
前後処理①								
Claude					52.53	61.34	55.07	88.08
前後処理②								
Claude	49.86	89.16	50.18	86.09	54.00	69.65	54.41	87.79
前後処理⑤								
Claude					51.36	60.77	51.53	87.78
<b>プロンプトを工夫した翻訳</b>								
prompt1								
Claude	48.04	88.33	50.42	86.16	52.68	61.86	53.37	87.98
f1500_gemma	38.47	82.93	41.26	83.17	42.13	58.57	45.77	84.16
f2500_gemma					42.52	58.36	46.01	84.29
prompt2								
Claude	49.42	88.76						
f1500_gemma	40.49	83.28						
prompt3								
Claude					52.87	62.28		
f1500_gemma					41.74	58.56		
f2500_gemma					41.35	58.10		
prompt4								
Claude						51.00	61.88	
f1500_gemma						32.50	56.97	
f2500_gemma						32.99	57.58	
prompt8								

Claude							52.21	87.31
f1500_gemma							40.05	80.50
f2500_gemma							39.67	80.50
prompt9								
Claude							51.63	87.51
f1500_gemma							40.72	81.12
f2500_gemma							40.19	81.51
<b>追加学習済大規模言語モデル</b>								
f1500_gemma	40.98	83.59	41.09	83.02	41.34	57.89	45.54	84.19
f2500_gemma					41.03	58.12	44.88	84.05
<b>ニューラル機械翻訳</b>								
nmt	55.20	91.05	53.09	87.97	61.29	86.07	57.26	89.11

## 8.1.2 中日

表 8.1-2 自動評価結果(中日)

項目	発明の名称		要約		請求項		明細書	
翻訳の種類/手法/モデル	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES
<b>大規模言語モデルによる翻訳</b>								
Claude	49.48	87.16	47.89	86.13	52.42	65.04	52.73	87.67
Gemini	52.60	88.90	52.39	88.26	56.87	72.61	56.06	88.68
Gemma	44.33	85.09	43.45	83.41	47.15	62.60	46.69	85.79
<b>前後処理を適用した翻訳</b>								
前後処理①								
Claude					54.78	66.39	58.26	89.00
前後処理⑤								
Claude					53.30	66.83	54.32	88.18
<b>プロンプトを工夫した翻訳</b>								
prompt1								
Claude	49.97	87.60	48.59	86.21	52.39	66.17	52.85	87.81
f1500_gemma	43.87	83.86	43.22	83.25	46.99	62.49	46.76	85.28
prompt2								
Claude	49.67	86.99						
f1500_gemma	43.61	84.81						
prompt3								
Claude					52.80	66.31		
f1500_gemma					46.50	62.39		
prompt4								
Claude					52.28	67.00		
f1500_gemma					42.85	60.12		
prompt8								
Claude							50.99	87.03
f1500_gemma							41.85	82.89
prompt9								
Claude							50.73	87.13
f1500_gemma							44.21	83.83
<b>追加学習済大規模言語モデル</b>								

f1500_gemma	44.58	84.43	43.35	83.35	47.30	62.33	47.22	85.48
<b>ニューラル機械翻訳</b>								
nmf	57.98	91.31	55.63	89.48	55.77	80.34	61.41	90.93

### 8.1.3 韓日

表 8.1-3 自動評価結果(韓日)

項目	発明の名称		要約		請求項		明細書	
	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES
<b>大規模言語モデルによる翻訳</b>								
Claude	49.72	88.60	54.13	87.78	48.15	75.03	53.96	89.79
Gemini	50.20	88.26	56.20	88.16	55.02	73.84	55.11	89.87
Gemma	47.03	88.09	52.14	87.17	46.65	74.98	49.40	88.34
<b>前後処理を適用した翻訳</b>								
前後処理①								
Claude					53.83	74.05	61.06	91.08
前後処理⑤								
Claude					57.94	82.37	62.59	91.61
<b>プロンプトを工夫した翻訳</b>								
prompt1								
Claude	49.39	88.05	55.12	88.03	49.05	74.79	54.56	89.91
f1500_gemma	46.86	86.81	51.63	86.80	47.46	72.39	49.53	88.32
prompt2								
Claude	49.62	88.63						
f1500_gemma	46.77	87.43						
prompt3								
Claude					49.34	74.69		
f1500_gemma					47.56	73.16		
prompt4								
Claude					51.06	73.47		
f1500_gemma					41.91	71.01		

prompt8									
Claude							54.46	89.85	
f1500_gemma							43.48	86.13	
prompt9									
Claude							54.44	89.70	
f1500_gemma							45.20	86.46	
<b>追加学習済大規模言語モデル</b>									
f1500_gemma	47.84	87.89	51.81	87.05	46.47	72.47	49.71	88.38	
<b>ニューラル機械翻訳</b>									
nmt	55.78	92.04	60.01	89.92	59.06	88.54	66.00	92.68	

#### 8.1.4 日英(発明の名称、要約、請求項、明細書)

表 8.1-4 自動評価結果(日英(発明の名称、要約、請求項、明細書))

項目	発明の名称		要約		請求項		明細書	
	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES
<b>大規模言語モデルによる翻訳</b>								
Claude	46.26	86.33	49.60	87.96	57.95	90.68	53.78	89.22
Gemini	44.84	84.71	52.21	88.44	59.27	91.77	55.64	88.93
Gemma	39.40	80.61	46.43	86.76	51.79	87.52	48.05	86.90
<b>前後処理を適用した翻訳</b>								
前後処理①								
Claude					58.95	90.68	54.90	89.59
前後処理⑤								
Claude					56.56	90.37	52.41	88.62
<b>プロンプトを工夫した翻訳</b>								
prompt1								
Claude	45.31	86.62	50.02	88.11	58.86	91.23	54.10	89.01
f1500_gemma	38.59	81.42	46.61	86.62	51.66	85.15	48.36	87.15
prompt2								

Claude	45.02	85.98						
f1500_gemma	35.68	80.57						
prompt3								
Claude					58.36	91.09		
f1500_gemma					51.57	85.35		
prompt4								
Claude					56.48	91.01		
f1500_gemma					43.88	84.18		
prompt8								
Claude						49.55	88.28	
f1500_gemma						41.30	84.58	
prompt9								
Claude						51.05	88.61	
f1500_gemma						40.83	85.08	
追加学習済大規模言語モデル								
f1500_gemma	36.39	79.95	46.12	86.14	50.80	82.77	48.19	87.28
ニューラル機械翻訳								
nmt	44.83	84.21	52.04	88.48	61.90	92.35	58.02	89.83

### 8.1.5 日英(拒絶理由、意見書、補正書、出願人名/発明者名)

表 8.1-5 自動評価結果(日英(拒絶理由、意見書、補正書、出願人名/発明者名))

項目	拒絶理由		意見書		補正書		出願人名/発明者名	
	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES
大規模言語モデルによる翻訳								
Claude	48.19	85.18	46.66	85.53	61.93	90.57	7.79	50.71
Gemini	47.72	84.01	48.11	84.86	64.31	91.78	8.52	40.17
Gemma	40.23	79.84	40.41	81.22	56.56	87.19	4.74	31.72
前後処理を適用した翻訳								
前後処理②								

Claude	47.76	84.75	47.26	85.32	63.09	91.56	6.92	48.16
プロンプトを工夫した翻訳								
prompt1								
Claude	48.99	85.07	46.93	85.52	62.72	90.84		
f1500_gemma	41.20	80.52	41.36	83.28	57.58	85.30		
<b>追加学習済大規模言語モデル</b>								
f1500_gemma	41.02	80.77	42.02	83.52	56.21	83.64	4.30	21.35
<b>ニューラル機械翻訳</b>								
nmt	51.36	84.89	51.12	85.78	57.82	89.15	3.70	37.73

## 8.1.6 項目別

8.1.1～3 に示した言語別の各集計結果を項目毎に集計した結果を示す(表 8.1-6)。前後処理②及び f2500\_gemma は英日のみを対象としており、集計母数が異なることから、集計対象外とした。

表 8.1-6 自動評価結果(項目(発明の名称、要約、請求項、明細書))

翻訳の種類/手法/モデル	発明の名称		要約		請求項		明細書	
	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES
<b>大規模言語モデルによる翻訳</b>								
Claude	48.62	87.61	50.44	87.11	52.84	72.99	53.44	88.68
Gemini	49.62	87.77	52.76	87.83	56.87	79.80	55.29	88.84
Gemma	42.63	84.24	45.79	85.02	46.93	70.89	47.08	86.29
<b>前後処理を適用した翻訳</b>								
<b>前後処理①</b>								
Claude					55.02	73.12	57.32	89.44
<b>前後処理⑤</b>								
Claude					54.79	75.09	55.21	89.04
<b>プロンプトを工夫した翻訳</b>								
<b>prompt1</b>								
Claude	48.18	87.65	51.04	87.13	53.25	73.51	53.72	88.68
f1500_gemma	41.95	83.76	45.68	84.96	47.06	69.65	47.61	86.23
<b>prompt2</b>								
Claude	48.43	87.59						
f1500_gemma	41.64	84.02						
<b>prompt3</b>								

Claude	53.34	73.60						
f1500_gemma	46.84	69.87						
<b>prompt4</b>								
Claude	52.71	73.34						
f1500_gemma	40.29	68.07						
<b>prompt8</b>								
Claude						51.80	88.12	
f1500_gemma						41.67	83.53	
<b>prompt9</b>								
Claude						51.96	88.23	
f1500_gemma						42.74	84.12	
<b>追加学習済大規模言語モデル</b>								
f1500_gemma	42.45	83.97	45.59	84.89	46.48	68.87	47.67	86.33
<b>ニューラル機械翻訳</b>								
nmt	53.45	89.65	55.19	88.96	59.51	86.82	60.67	90.64

表 8.1-7 自動評価結果(項目(拒絶理由、意見書、補正書、出願人名/発明者名))

翻訳の種類/手法/モデル	拒絶理由		意見書		補正書		出願人名/発明者名	
	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES
<b>大規模言語モデルによる翻訳</b>								
Claude	48.19	85.18	46.66	85.53	61.93	90.57	7.79	50.71
Gemini	47.72	84.01	48.11	84.86	64.31	91.78	8.52	40.17
Gemma	40.23	79.84	40.41	81.22	56.56	87.19	4.74	31.72
<b>前後処理を適用した翻訳</b>								
<b>前後処理②</b>								
Claude	47.76	84.75	47.26	85.32	63.09	91.56	6.92	48.16
<b>プロンプトを工夫した翻訳</b>								
<b>prompt1</b>								
Claude	48.99	85.07	46.93	85.52	62.72	90.84		
f1500_gemma	41.20	80.52	41.36	83.28	57.58	85.30		
<b>追加学習済大規模言語モデル</b>								
f1500_gemma	41.02	80.77	42.02	83.52	56.21	83.64	4.30	21.35
<b>ニューラル機械翻訳</b>								
<b>none</b>								
nmt	51.36	84.89	51.12	85.78	57.82	89.15	3.70	37.73

### 8.1.7 大規模言語モデル 3 モデル及び NMT モデルの比較

3 種の大規模言語モデル (Claude、Gemini、Gemma) と、比較のためのニューラル翻訳モデル (nmt) の全 4 モデルについて自動評価のスコアを比較した(表 8.1.7-1, 表 8.1.7-2)。

これら 4 モデルは、まず特許文献として各言語 (英日、中日、韓日、日英) とともに「発明の名称」「要約」「特許請求の範囲」「明細書」の 4 項目から 500 文ずつ、計 2,000 文を評価した。本項に示す各言語の評価結果は、これら 4 項目・全 2,000 文の集計結果となる。また、これとは別途に日英審査書類として「拒絶理由通知書」「意見書」「補正書」からそれぞれ 500 文、計 1,500 文についても評価しており、集計結果を表中「審査」欄に示した。

表 8.1.7-1 BLEU

モデル	英日	中日	韓日	日英	審査	全言語
Claude	51.32	50.63	51.49	51.90	52.26	51.52
Gemini	52.94	54.48	54.13	52.99	53.38	53.58
Gemma	41.80	45.41	48.81	46.42	45.73	45.63
nmt	56.71	57.70	60.21	54.20	53.43	56.45

表 8.1.7-2 RIBES

モデル	英日	中日	韓日	日英	審査	全言語
Claude	81.04	81.50	85.30	88.55	87.10	84.70
Gemini	86.13	84.61	85.03	88.46	86.88	86.22
Gemma	77.13	79.22	84.64	85.45	82.75	81.84
nmt	88.55	88.01	90.80	88.72	86.60	88.54

大規模言語モデルの自動評価 BLEU、RIBES は、どちらも評価の高い順に nmt、Gemini、Claude、Gemma となった。言語別に見ると韓日の nmt の BLEU が 60、RIBES が 90 を超える高い評価結果となった。

以下、自動評価結果について機械翻訳において広く使われている BLEU の値を基に考察する(表 8.1.7-3)。

考察に先立ち、参考までに BLEU の計算式や、関連する各種パラメータについて概要

を説明する。

- BLEU

最終的な機械評価のスコア。以下の式で算出される。

$$BLEU = BP \times \exp \left( \frac{1}{4} \sum_{n=1}^4 \log p_n \right)$$

ここで、

$p_n$  = n-gram 精度 (score1~4)

BP = brevity penalty (短すぎる翻訳への罰則)

である。

- score1

1-gram precision (単語一致率) であり、単語単位でどれだけ一致しているかを意味する。

- score2

2-gram precision (連続2語一致率) であり、「I like」「like apples」などの2語連続一致率を意味する。

- score3

3-gram precision であり、より文構造の一致を評価するものである。

- score4

4-gram precision であり、文全体の自然さ・語順の正確さを反映する。

- BP (Brevity Penalty)

短すぎる翻訳へのペナルティであり、1に近いほど長さは適切で、例えば値が1より小さすぎる場合は「短すぎる」といえる。

$$BP = \begin{cases} 1 & (hyp\_len > ref\_len) \\ e^{(1-ref\_len/hyp\_len)} & (hyp\_len \leq ref\_len) \end{cases}$$

- ratio

$hyp\_len/ref\_len$  であり、1だと機械翻訳文と参照文(正解文)とが、ちょうど同じ語数といえる。

・ hyp\_len

機械翻訳文の総単語数。

・ ref\_len

参照文（正解文）の総単語数。

続いて、大規模言語モデル3モデル及びNMTモデルの比較結果をBLEUの各パラメータの面から詳述する。なお、他の項目と比べて機械翻訳向けとは言い難い出願人発明者の名称については考察対象外としている。

Geminiは、LLMの中では最も高スコアであった。score1が高く、score4も比較的高く、語彙選択と文脈の両方が比較的良好。また、ratioが1.00~1.06程度のやや長めのものが多く、過剰生成傾向がある。

Claudeは、Geminiよりやや劣るが、Gemmaよりは明確に高スコアであった。score1は高いがscore4が伸びにくいことから語彙は良いが、文脈の連続性が弱い傾向があるといえる。ratioは0.95~1.03と比較的安定していること、BPが1.0に近いことから、長さ調整は良好といえる。

Gemmaは全モデル中で最もBLEUが低く、短めの出力でBPが下がりやすい。score1~4が全体的に低く、BPも0.89~0.99程度の低めのものが多い。また、ratioについても0.89~0.99と、短めのものが多く、hyp\_lenがref\_lenより明確に短くなっている。このように、Gemmaは短く訳す傾向が強く、その結果、BLEUが大きく下がっており、語彙・文脈の両面で他モデルに劣っている。

NMTは、score1~4が全体的に高い。特にscore4(4-gram)が他モデルより明確に高く、文脈的な連続性・語順の正確性が強いといえる。また、BPが0.97~1.00と安定しており、過不足の少ない長さ調整ができています。そして、ratioが0.97~1.00付近であることから過剰生成が少なく、参照文長に近い。

NMTは長さ・語順・文脈の安定性が高く、BLEU指標では最高である。

表 8.1.7-3 BLEU 詳細

方向	項目	タスク	BLEU	score1	score2	score3	score4	BP	ratio	hyp_len	ref_len
cj	abs	2-2_none_gemini	52.39	77.4	58.5	45.9	36.3	1	1.055	34118	32352
cj	abs	2-2_none_gemma	43.45	72.6	50.2	36.5	26.8	1	1.025	33160	32352
cj	abs	2-2_none_claude	47.89	74.8	54.2	41.1	31.5	1	1.032	33391	32352
cj	abs	2-4_none_nmt	55.63	78.6	61.1	49.4	40.3	1	1.009	32653	32352
cj	clm	2-2_none_gemini	56.87	78.8	63	51.1	41.3	1	1.062	45082	42433
cj	clm	2-2_none_gemma	47.15	74.9	54	40.4	30.2	1	1.024	43467	42433

cj	clm	2-2_none_claude	52.42	77.3	58.7	46	36.2	1	1.038	44030	42433
cj	clm	2-4_none_nmt	55.77	78.6	61.5	49.6	40.3	1	1.009	42802	42433
cj	des	2-2_none_gemini	56.06	79.3	61.6	49.8	40.6	1	1.004	19746	19663
cj	des	2-2_none_gemma	46.69	74.6	53.8	40.7	31.2	0.983	0.983	19330	19663
cj	des	2-2_none_claude	52.73	77.6	58.8	46.2	36.7	1	1.002	19711	19663
cj	des	2-4_none_nmt	61.41	82.3	66.8	56	47.4	0.994	0.994	19547	19663
cj	tit	2-2_none_gemini	52.6	77.3	59	46.5	37.2	0.993	0.993	5602	5643
cj	tit	2-2_none_gemma	44.33	72.1	51.4	38.5	28.7	0.986	0.986	5563	5643
cj	tit	2-2_none_claude	49.48	75.5	56.1	43.3	33.7	0.993	0.993	5602	5643
cj	tit	2-4_none_nmt	57.98	80.6	64.3	53	43.9	0.984	0.984	5552	5643
ej	abs	2-2_none_gemini	50.25	76.3	56.6	43.7	34	0.999	0.999	19886	19912
ej	abs	2-2_none_gemma	41.14	72.1	49	35.4	25.9	0.97	0.97	19321	19912
ej	abs	2-2_none_claude	50.13	77.5	57.4	44.2	34.5	0.982	0.982	19558	19912
ej	abs	2-4_none_nmt	53.09	78.4	59.7	47.3	37.9	0.987	0.987	19647	19912
ej	clm	2-2_none_gemini	56.31	79.4	62.6	50.3	40.8	0.996	0.996	23993	24080
ej	clm	2-2_none_gemma	42.14	75.1	51.7	37.2	26.8	0.949	0.951	22892	24080
ej	clm	2-2_none_claude	52.84	77.9	59.5	46.3	36.3	1	1.014	24406	24080
ej	clm	2-4_none_nmt	61.29	82.8	68	56.5	47.3	0.984	0.984	23704	24080
ej	des	2-2_none_gemini	54.36	78.9	60.7	48.2	38.8	0.993	0.993	19585	19717
ej	des	2-2_none_gemma	44.17	73.6	52.2	38.5	28.7	0.973	0.974	19197	19717
ej	des	2-2_none_claude	53.3	78.5	60.2	47.2	37.6	0.99	0.99	19526	19717
ej	des	2-4_none_nmt	57.26	80.4	63.6	51.7	42.4	0.99	0.99	19517	19717
ej	tit	2-2_none_gemini	50.82	78.1	58.7	45.9	36.7	0.964	0.964	5136	5326
ej	tit	2-2_none_gemma	39.74	72.1	49	34.6	24.9	0.952	0.953	5076	5326
ej	tit	2-2_none_claude	49.02	78.3	57.5	44.1	34.7	0.957	0.958	5100	5326
ej	tit	2-4_none_nmt	55.2	80.6	62.7	50.4	41	0.971	0.971	5173	5326
je	abs	2-2_none_gemini	52.21	77.8	58.1	45.5	36.1	1	1.019	16760	16445
je	abs	2-2_none_gemma	46.43	76.2	54	40.6	31	0.973	0.974	16014	16445
je	abs	2-2_none_claude	49.6	77.8	56.9	43.8	34.2	0.977	0.977	16069	16445
je	abs	2-4_none_nmt	52.04	78.5	58.5	45.6	35.9	0.994	0.994	16341	16445
je	clm	2-2_none_gemini	59.27	81.3	64.7	53.1	44.2	1	1.012	19482	19242
je	clm	2-2_none_gemma	51.79	78.3	58.7	45.7	36.2	0.986	0.986	18980	19242

je	clm	2-2_none_claude	57.95	80.9	63.6	51.7	42.6	0.999	0.999	19216	19242
je	clm	2-4_none_nmt	61.9	83.8	67.9	56.9	48.3	0.984	0.984	18938	19242
je	des	2-2_none_gemini	55.64	79.8	61.4	49.1	39.9	1	1.006	16634	16542
je	des	2-2_none_gemma	48.05	78	56.9	43.5	34	0.949	0.951	15725	16542
je	des	2-2_none_claude	53.78	79.9	60.9	48.1	38.7	0.98	0.981	16221	16542
je	des	2-4_none_nmt	58.02	81.8	64.2	52.3	43.4	0.988	0.988	16338	16542
je	hos	2-2_none_gemini	64.31	83.9	69.3	58.6	50.2	1	1.012	23066	22787
je	hos	2-2_none_gemma	56.56	82.8	64.7	52.5	43.3	0.958	0.959	21846	22787
je	hos	2-2_none_claude	61.93	84.1	68.1	57	48.3	0.983	0.983	22392	22787
je	hos	2-4_none_nmt2	57.82	81.6	64.3	52.4	43.5	0.983	0.983	22404	22787
je	ike	2-2_none_gemini	48.11	76.9	55.2	41.6	31.9	0.987	0.987	20955	21226
je	ike	2-2_none_gemma	40.41	76.6	51.9	37.7	28.2	0.891	0.897	19030	21226
je	ike	2-2_none_claude	46.66	78.2	55.9	42	32.1	0.947	0.949	20133	21226
je	ike	2-4_none_nmt2	51.12	78.9	58.2	45.1	35.7	0.98	0.98	20805	21226
je	kyo	2-2_none_gemini	47.72	76.7	54.6	41.2	31.7	0.987	0.987	20156	20412
je	kyo	2-2_none_gemma	40.23	75.9	51.1	36.9	27.2	0.906	0.91	18579	20412
je	kyo	2-2_none_claude	48.19	78.7	56.8	43.6	33.8	0.952	0.953	19447	20412
je	kyo	2-4_none_nmt2	51.36	79.9	59.3	45.9	36.1	0.97	0.97	19805	20412
je	tit	2-2_none_gemini	44.84	72	51.7	38.2	28.4	1	1.067	4721	4423
je	tit	2-2_none_gemma	39.4	68.7	46.3	32.5	23.3	1	1.055	4665	4423
je	tit	2-2_none_claude	46.26	75.8	54.9	40.3	29.7	0.979	0.98	4333	4423
je	tit	2-4_none_nmt	44.83	73.5	52.2	37.6	28	1	1.022	4521	4423
kj	abs	2-2_none_gemini	56.2	78.9	61.8	50	40.9	1	1.008	26156	25944
kj	abs	2-2_none_gemma	52.14	77	58.5	46.2	36.9	0.99	0.99	25677	25944
kj	abs	2-2_none_claude	54.13	78.4	60.5	48.5	39.4	0.986	0.986	25593	25944
kj	abs	2-4_none_nmt	60.01	83	67.1	56.3	47.8	0.965	0.965	25040	25944
kj	clm	2-2_none_gemini	55.02	80.6	62.4	49	38.8	0.989	0.99	38616	39025
kj	clm	2-2_none_gemma	46.65	77.2	55.6	41.9	32.2	0.951	0.952	37163	39025
kj	clm	2-2_none_claude	48.15	78.6	57.4	43.8	34.2	0.944	0.946	36910	39025
kj	clm	2-4_none_nmt	59.06	84.1	67.8	55.7	46.1	0.955	0.956	37303	39025
kj	des	2-2_none_gemini	55.11	78.9	60.5	48.8	39.6	1	1.036	21043	20309
kj	des	2-2_none_gemma	49.4	75.3	55.2	42.7	33.5	1	1.021	20729	20309

kj	des	2-2_none_claude	53.96	78.2	59.4	47.5	38.5	1	1.028	20876	20309
kj	des	2-4_none_nmt	66	85.2	71.6	61.8	53.8	0.983	0.984	19975	20309
kj	tit	2-2_none_gemini	50.2	75.1	55.8	43.9	35	0.996	0.996	5232	5251
kj	tit	2-2_none_gemma	47.03	74.1	53.9	41.1	32.2	0.981	0.982	5154	5251
kj	tit	2-2_none_claude	49.72	75.4	56	44.1	35.5	0.981	0.981	5150	5251
kj	tit	2-4_none_nmt	55.78	80.9	63.5	51.2	41.4	0.971	0.972	5102	5251

### 8.1.8 前後処理を導入したモデル

大規模言語モデルによる機械翻訳を行う際に、特許翻訳に適すると考えられる特定の前後処理を導入した場合の翻訳品質を評価し、ベースモデルからの変化の度合いを確認した(表 8.1.8-1, 表 8.1.8-2)。

#### 8.1.8.1 前後処理①

前後処理①モデルについて、英日、中日、韓日、日英それぞれ「特許請求の範囲」と「明細書」の2項目、1,000文を評価した。下の各表に、前後処理①モデルの各評価の集計結果を、ベースモデル Claude との対比で示す。比較条件を合わせるため、ベースモデルの評価結果も「特許請求の範囲」と「明細書」の2項目のみの合算値とした。

表 8.1.8-1 BLEU

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	53.07	52.58	51.06	55.87	53.14
前後処理①	53.80	56.52	57.45	56.93	56.17

表 8.1.8-2 RIBES

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	74.62	76.35	82.41	89.95	80.83
前後処理①	74.71	77.70	82.57	90.13	81.28

前後処理①と Claude ベースモデルの比較では、BLEU、RIBES とともに全言語でスコアの上昇が見られた。特に韓日の BLEU は6ポイント以上の上昇が見られた。

BLEU について詳述する(表 8.1.8-3)。

N-gram 適合率 (score1~score4)については、特に kj, des において、score4 (4-gram 一致率) が 38.5 から 47.0 へ、cj, des では 36.7 から 43.3 へと大きな伸びを見せている。これは、単語単位の正解だけでなく、特許文書特有の長い定型句や複雑な構文が、参照訳と極めて高いレベルで一致していることを示している。また、score1 (単語レベル) も一律に 2~3 ポイント向上しており、前後処理①によって用語選択面での改善が見て取れる。

文長制御 (ratio/hyp\_len) とペナルティ (BP)については、cj, des では ratio が 0.999、ej, des では 0.997 と、参照訳と近い長さの翻訳がされていることがわかる。また、オリジナルで「短すぎる」傾向があった kj, clm (ratio 0.946, BP 0.944) が、前後処理 1 では ratio 0.971, BP 0.970 まで改善した。

前後処理①は、韓国語から日本語への翻訳において、BLEU が大幅に上昇している。唯一微減した ej, clm についても、適合率は上がっている (score1: 77.9→78.2) ため、実質的な品質は向上している可能性がある。

表 8.1.8-3 BLEU 詳細

方向	項目	タスク	BLEU	score1	score2	score3	score4	BP	ratio	hyp_len	ref_len
cj	clm	2-2_none_claude	52.42	77.3	58.7	46	36.2	1	1.038	44030	42433
cj	des	2-2_none_claude	52.73	77.6	58.8	46.2	36.7	1	1.002	19711	19663
ej	clm	2-2_none_claude	52.84	77.9	59.5	46.3	36.3	1	1.014	24406	24080
ej	des	2-2_none_claude	53.3	78.5	60.2	47.2	37.6	0.99	0.99	19526	19717
je	clm	2-2_none_claude	57.95	80.9	63.6	51.7	42.6	0.999	0.999	19216	19242
je	des	2-2_none_claude	53.78	79.9	60.9	48.1	38.7	0.98	0.981	16221	16542
kj	clm	2-2_none_claude	48.15	78.6	57.4	43.8	34.2	0.944	0.946	36910	39025
kj	des	2-2_none_claude	53.96	78.2	59.4	47.5	38.5	1	1.028	20876	20309
cj	clm	2-2-2_method1	54.78	78.6	60.9	48.5	38.8	1	1.038	44061	42433
cj	des	2-2-2_method1	58.26	80.4	63.6	52.2	43.3	0.999	0.999	19647	19663
ej	clm	2-2-2_method1	52.53	78.2	59.3	45.9	35.8	1	1.002	24139	24080
ej	des	2-2-2_method1	55.07	79.1	61.3	48.8	39.4	0.997	0.997	19651	19717
je	clm	2-2-2_method1	58.95	81.8	64.9	53.3	44.2	0.991	0.991	19064	19242
je	des	2-2-2_method1	54.9	80.9	62.3	49.8	40.7	0.971	0.971	16068	16542
kj	clm	2-2-2_method1	53.83	80.6	61.8	48.7	39	0.97	0.971	37886	39025
kj	des	2-2-2_method1	61.06	81.6	65.7	55.2	47	1	1.007	20442	20309

### 8.1.8.2 前後処理②

前後処理②モデルについて、英日の全4項目（「発明の名称」「要約」「特許請求の範囲」「明細書」）、及び審査書類（日英）の全3種（「拒絶理由通知」「意見書」「補正書」）を対象に各評価を実施した。下表に、その結果をベースモデル Claude と対比して示す(表 8.1.8-4, 表 8.1.8-5)。

表 8.1.8-4 BLEU

モデル	英日	日英(審査)	全言語
Claude (ベース)	51.32	52.26	51.79
前後処理②	52.11	52.70	52.41

表 8.1.8-5 RIBES

モデル	英日	日英(審査)	全言語
Claude (ベース)	81.04	87.10	84.07
前後処理②	83.17	87.21	85.19

前後処理②と Claude ベースモデルの比較では、英日、日英ともにスコアの上昇は見られたが、1ポイント未満の上昇に留まった。

BLEU について詳述する(表 8.1.8-6)。

N-gram 適合率 (score1~score4)については、特 ej, clm では score4 が 36.3 から 38.2 へ向上、je, hos では 48.3 から 49.5 へ微増している。ej, abs や ej, clm では score1 がオリジナルより低下している。それにもかかわらず BLEU スコアが向上（または維持）しているのは、単語の表面的な一致よりも、フレーズとしての繋がり (score3, score4) が参照訳に近づいたためと考えられる。

文長制御 (ratio / hyp\_len) とペナルティ (BP)については、ej, tit では ratio が 0.958 から 0.970 に改善され、これに伴い BP も向上している。ej, des においては ratio が 0.999 となり、ほぼ参照訳と同一の文長となっている。

前後処理②では、BLEUが増加している ej, clm、des や je, hos (補正書) であってもその増加幅は 1.1~1.2 程度であり、N-gram 適合率、文長制御、ペナルティについてもそれほど改善していない。

表 8.1.8-6 BLEU 詳細

方向	項目	タスク	BLEU	score1	score2	score3	score4	BP	ratio	hyp_len	ref_len
ej	abs	2-2_none_claude	50.13	77.5	57.4	44.2	34.5	0.982	0.982	19558	19912
ej	clm	2-2_none_claude	52.84	77.9	59.5	46.3	36.3	1	1.014	24406	24080
ej	des	2-2_none_claude	53.3	78.5	60.2	47.2	37.6	0.99	0.99	19526	19717
ej	tit	2-2_none_claude	49.02	78.3	57.5	44.1	34.7	0.957	0.958	5100	5326
je	app	2-2_none_claude	7.79	43.6	20.2	4.4	1	1	1.368	1935	1414
je	hos	2-2_none_claude	61.93	84.1	68.1	57	48.3	0.983	0.983	22392	22787
je	ike	2-2_none_claude	46.66	78.2	55.9	42	32.1	0.947	0.949	20133	21226
je	kyo	2-2_none_claude	48.19	78.7	56.8	43.6	33.8	0.952	0.953	19447	20412
ej	abs	2-2-2_method2	50.18	77	57.3	44.1	34.2	0.988	0.988	19681	19912
ej	clm	2-2-2_method2	54	77.1	60.3	47.8	38.2	1	1.042	25094	24080
ej	des	2-2-2_method2	54.41	78.6	60.6	48	38.5	0.999	0.999	19692	19717
ej	tit	2-2-2_method2	49.86	77.9	57.7	44.5	35	0.97	0.97	5167	5326
je	app	2-2-2_method2	6.92	40.7	12.2	4.1	1.1	1	1.516	2144	1414
je	hos	2-2-2_method2	63.09	84.4	68.9	58	49.5	0.987	0.987	22489	22787
je	ike	2-2-2_method2	47.26	78.3	56.3	42.3	32.6	0.952	0.953	20232	21226
je	kyo	2-2-2_method2	47.76	78.2	56.1	42.8	33	0.957	0.958	19552	20412

8.1.8.3 前後処理⑤

前後処理⑤モデルは、前後処理①と同様、英日、中日、韓日、日英それぞれで「特許請求の範囲」と「明細書」の2項目、1,000文を評価対象とした。下の各表に、前後処理⑤モデルの各評価の集計結果を、ベースモデル Claude の評価結果との対比で示す(表 8.1.8-7, 表 8.1.8-8)。

表 8.1.8-7 BLEU

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	53.07	52.58	51.06	55.87	53.14
前後処理⑤	51.45	53.81	60.27	54.49	55.00

表 8.1.8-8 RIBES

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	74.62	76.35	82.41	89.95	80.83

前後処理⑤	74.27	77.50	86.99	89.50	82.07
-------	-------	-------	-------	-------	-------

前後処理⑤と Claude ベースモデルの比較では、BLEU は英日と日英で 2 ポイント未満の低下が見られ、中日で 1 ポイント強、韓日で 9 ポイント強の上昇が見られた。RIBES も同様に英日、日英の低下、中日、韓日でスコアの上昇が見られた。

BLEU について詳述する(表 8.1.8-9)。

N-gram 適合率 (score1~score4)については、kj, clm の score4 は 34.2 から 44.1 へ、kj, des では 38.5 から 49.4 へと上昇した。これは、専門用語の配置や特許文書としての構文が、参照訳と極めて高いレベルで一致していることを意味する。

文長制御 (ratio / hyp\_len) とペナルティ (BP)については、cj, des の ratio が 1.002 と、長さを制御できているといえる一方、ej, des や je, des では ratio が 0.975 まで低下している。これらは、前後処理によって文章が簡略化されすぎ、必要なキーワードやフレーズが脱落したことが BLEU 低下 (BP の影響を含む) を招いたと考えられる。

前後処理⑤は、韓国語から日本語への翻訳において、BLEU が大幅に上昇している。その一方、英語が絡むタスクでは、BLEU が軒並み低下している。特に score1~score4 が一律に下がっている点から、語彙の選択自体が参照訳と乖離している様子が見て取れる。

表 8.1.8-9 BLEU 詳細

方向	項目	タスク	BLEU	score1	score2	score3	score4	BP	ratio	hyp_len	ref_len
cj	clm	2-2_none_claude	52.42	77.3	58.7	46	36.2	1	1.038	44030	42433
cj	des	2-2_none_claude	52.73	77.6	58.8	46.2	36.7	1	1.002	19711	19663
ej	clm	2-2_none_claude	52.84	77.9	59.5	46.3	36.3	1	1.014	24406	24080
ej	des	2-2_none_claude	53.3	78.5	60.2	47.2	37.6	0.99	0.99	19526	19717
je	clm	2-2_none_claude	57.95	80.9	63.6	51.7	42.6	0.999	0.999	19216	19242
je	des	2-2_none_claude	53.78	79.9	60.9	48.1	38.7	0.98	0.981	16221	16542
kj	clm	2-2_none_claude	48.15	78.6	57.4	43.8	34.2	0.944	0.946	36910	39025
kj	des	2-2_none_claude	53.96	78.2	59.4	47.5	38.5	1	1.028	20876	20309
cj	clm	2-2-2_method5	53.3	77.4	59.5	47.1	37.2	1	1.044	44298	42433
cj	des	2-2-2_method5	54.32	78.5	60.1	47.9	38.5	1	1.002	19694	19663
ej	clm	2-2-2_method5	51.36	77.2	58	44.7	34.8	1	1.004	24172	24080
ej	des	2-2-2_method5	51.53	78.4	59.4	46.2	36.3	0.975	0.975	19227	19717
je	clm	2-2-2_method5	56.56	80.1	62.3	50.1	40.9	1	1.002	19281	19242
je	des	2-2-2_method5	52.41	79.6	59.9	46.9	37.4	0.975	0.975	16131	16542
kj	clm	2-2-2_method5	57.94	82.6	65.9	53.5	44.1	0.968	0.968	37788	39025

kj	des	2-2-2_method5	62.59	82.9	68	57.7	49.4	0.988	0.988	20074	20309
----	-----	---------------	-------	------	----	------	------	-------	-------	-------	-------

### 8.1.9 プロンプトに工夫を施したモデル

大規模言語モデルによる機械翻訳を行う際に、特許翻訳に適すると考えられるプロンプトに工夫を施した6種のモデルについて、ベースモデルからの変化の度合いを確認した。

#### 8.1.9.1 プロンプト①

プロンプト①では、翻訳対象の項目（発明の名称、要約、請求の範囲、明細書等）および、想定読者に「特許のサーチャー」を指定した。以下、このプロンプトを Claude に用いて生成させた機械翻訳文の評価結果を示す(表 8.1.9-1, 表 8.1.9-2)。なお、評価対象項目は各言語とも「発明の名称」「特許請求の範囲」「明細書」の3項目である。

表 8.1.9-1 BLEU

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	51.32	50.63	51.49	51.90	51.34
プロンプト①	51.13	50.95	52.03	52.07	51.55

表 8.1.9-2 RIBES

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	81.04	81.50	85.30	88.55	84.10
プロンプト①	81.08	81.95	85.20	88.74	84.24

プロンプト①と Claude ベースモデルの比較では、プロンプト①の BLEU が英日で微減、その他の言語で微増が見られた。RIBES は中日で 0.4 ポイント増加、英日、日英で微増、韓日で微減が見られた。

BLEU について詳述する(表 8.1.9-3)。

N-gram 適合率 (score1~score4)については、ほとんど変化がない。  
 文長制御 (ratio / hyp\_len) とペナルティ (BP)については、プロンプト適用後、  
 hyp\_len (生成された翻訳の長さ) が全体的に増加する傾向にある。

表 8.1.9-3 BLEU 詳細

方向	項目	タスク	BLEU	score1	score2	score3	score4	BP	ratio	hyp_len	ref_len
cj	abs	2-2_none_claude	47.89	74.8	54.2	41.1	31.5	1	1.032	33391	32352
cj	clm	2-2_none_claude	52.42	77.3	58.7	46	36.2	1	1.038	44030	42433
cj	des	2-2_none_claude	52.73	77.6	58.8	46.2	36.7	1	1.002	19711	19663
cj	tit	2-2_none_claude	49.48	75.5	56.1	43.3	33.7	0.993	0.993	5602	5643
ej	abs	2-2_none_claude	50.13	77.5	57.4	44.2	34.5	0.982	0.982	19558	19912
ej	clm	2-2_none_claude	52.84	77.9	59.5	46.3	36.3	1	1.014	24406	24080
ej	des	2-2_none_claude	53.3	78.5	60.2	47.2	37.6	0.99	0.99	19526	19717
ej	tit	2-2_none_claude	49.02	78.3	57.5	44.1	34.7	0.957	0.958	5100	5326
je	abs	2-2_none_claude	49.6	77.8	56.9	43.8	34.2	0.977	0.977	16069	16445
je	app	2-2_none_claude	7.79	43.6	20.2	4.4	1	1	1.368	1935	1414
je	clm	2-2_none_claude	57.95	80.9	63.6	51.7	42.6	0.999	0.999	19216	19242
je	des	2-2_none_claude	53.78	79.9	60.9	48.1	38.7	0.98	0.981	16221	16542
je	hos	2-2_none_claude	61.93	84.1	68.1	57	48.3	0.983	0.983	22392	22787
je	ike	2-2_none_claude	46.66	78.2	55.9	42	32.1	0.947	0.949	20133	21226
je	kyo	2-2_none_claude	48.19	78.7	56.8	43.6	33.8	0.952	0.953	19447	20412
je	tit	2-2_none_claude	46.26	75.8	54.9	40.3	29.7	0.979	0.98	4333	4423
kj	abs	2-2_none_claude	54.13	78.4	60.5	48.5	39.4	0.986	0.986	25593	25944
kj	clm	2-2_none_claude	48.15	78.6	57.4	43.8	34.2	0.944	0.946	36910	39025
kj	des	2-2_none_claude	53.96	78.2	59.4	47.5	38.5	1	1.028	20876	20309
kj	tit	2-2_none_claude	49.72	75.4	56	44.1	35.5	0.981	0.981	5150	5251
cj	abs	2-2-1_prompt1	48.59	75.3	55	41.8	32.2	1	1.033	33435	32352
cj	clm	2-2-1_prompt1	52.39	76.4	58.7	46.2	36.4	1	1.062	45061	42433
cj	des	2-2-1_prompt1	52.85	77.5	58.8	46.4	36.9	1	1.007	19809	19663
cj	tit	2-2-1_prompt1	49.97	75.8	56.7	44.2	34.8	0.986	0.986	5562	5643
ej	abs	2-2-1_prompt1	50.42	77.5	57.5	44.3	34.6	0.987	0.987	19655	19912
ej	clm	2-2-1_prompt1	52.68	76.8	59.5	46.4	36.3	1	1.045	25173	24080
ej	des	2-2-1_prompt1	53.37	78.3	59.9	47.1	37.5	0.995	0.995	19627	19717
ej	tit	2-2-1_prompt1	48.04	78.2	57.2	43.5	33.6	0.95	0.951	5066	5326
je	abs	2-2-1_prompt1	50.02	78.3	57.6	44.5	34.8	0.973	0.973	16006	16445

je	clm	2-2-1_prompt1	58.86	81.4	64.4	52.6	43.6	1	1.001	19256	19242
je	des	2-2-1_prompt1	54.1	80.2	61.2	48.4	39	0.981	0.981	16228	16542
je	hos	2-2-1_prompt1	62.72	84.1	68.6	57.5	48.8	0.989	0.989	22529	22787
je	ike	2-2-1_prompt1	46.93	78.2	56.1	42.2	32.4	0.948	0.95	20158	21226
je	kyo	2-2-1_prompt1	48.99	79.1	57.3	44.2	34.5	0.956	0.957	19531	20412
je	tit	2-2-1_prompt1	45.31	75.9	54.3	39.6	29.1	0.97	0.971	4293	4423
kj	abs	2-2-1_prompt1	55.12	79.2	61.5	49.6	40.5	0.986	0.986	25577	25944
kj	clm	2-2-1_prompt1	49.05	78.9	58.1	44.4	34.7	0.952	0.953	37178	39025
kj	des	2-2-1_prompt1	54.56	78.5	59.9	48.2	39.1	1	1.03	20927	20309
kj	tit	2-2-1_prompt1	49.39	75.5	55.9	43.8	35.1	0.979	0.979	5140	5251

### 8.1.9.2 プロンプト②

プロンプト②では、入力文の技術分野の情報として、その文を採取した文献の筆頭 IPC サブクラスの説明文を示した。以下、Claude にこのプロンプトを用いた機械翻訳文の評価結果を示す(表 8.1.9-4, 表 8.1.9-5)。なお、各言語とも対象項目は「発明の名称」のみである。

表 8.1.9-4 BLEU

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	49.02	49.48	49.72	46.26	48.62
プロンプト②	49.42	49.67	49.62	45.02	48.43

表 8.1.9-5 RIBES

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	88.34	87.16	88.60	86.33	87.61
プロンプト②	88.76	86.99	88.63	85.98	87.59

プロンプト②と Claude ベースモデルの比較では、プロンプト②の BLEU が日英で 1.2 ポイント減少し、英日、中日で微増、韓日で微減が見られた。RIBES は英日、韓日で微増、中日、日英で微減が見られた。

BLEU について詳述する(表 8.1.9-6)。

N-gram 適合率 (score1~score4)については、je を除きほとんど変化がない。

文長制御 (ratio / hyp\_len) とペナルティ (BP)についても、ほとんど変化がない。

日英翻訳については、すべてのスコア (特に score2 以降) が低下しており、このプロンプトによって語彙選択やフレーズ構築をかえって制限してしまった可能性がある。

表 8.1.9-6 BLEU 詳細

方向	項目	タスク	BLEU	score1	score2	score3	score4	BP	ratio	hyp_len	ref_len
cj	tit	2-2_none_claude	49.48	75.5	56.1	43.3	33.7	0.993	0.993	5602	5643
ej	tit	2-2_none_claude	49.02	78.3	57.5	44.1	34.7	0.957	0.958	5100	5326
je	tit	2-2_none_claude	46.26	75.8	54.9	40.3	29.7	0.979	0.98	4333	4423
kj	tit	2-2_none_claude	49.72	75.4	56	44.1	35.5	0.981	0.981	5150	5251
cj	tit	2-2-1_prompt2	49.67	75.1	55.9	43.4	33.8	0.997	0.997	5628	5643
ej	tit	2-2-1_prompt2	49.42	78.3	57.7	44.3	35.2	0.96	0.961	5117	5326
je	tit	2-2-1_prompt2	45.02	75.2	53.5	39	28.7	0.977	0.977	4323	4423
kj	tit	2-2-1_prompt2	49.62	75.7	56.1	44	35.2	0.98	0.98	5148	5251

### 8.1.9.3 プロンプト③&プロンプト④

プロンプト③は、プロンプト①「翻訳対象の項目 (発明の名称、要約、請求の範囲、明細書等) および、想定読者に「特許のサーチャー」を指定」とプロンプト②「技術分野の情報として、IPC の説明文 (サブクラス) を指定」の双方を組み合わせて使用したモデルである。一方、プロンプト④は、プロンプト③に対して更に「文構造を分析し、主要要素 (主語、動詞、目的語) を特定する」という CoT(Chain of Thought)を追加したモデルである。両モデルとも各言語の「特許請求の範囲」が対象項目であるため、本項でまとめて論じる(表 8.1.9-7, 表 8.1.9-8)。

表 8.1.9-7 BLEU

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	52.84	52.42	48.15	57.95	52.84
プロンプト③	52.87	52.80	49.34	58.36	53.34
プロンプト④	51.00	52.28	51.06	56.48	52.71

表 8.1.9-8 RIBES

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	61.19	65.04	75.03	90.68	72.99
プロンプト③	62.28	66.31	74.69	91.09	73.60
プロンプト④	61.88	67.00	73.47	91.01	73.34

プロンプト③と Claude ベースモデルの比較では、BLEU が英日、中日、日英で 1 ポイント未満の、韓日で 1 ポイント強の上昇が見られた。RIBES は韓日以外でスコアの上昇が、韓日で 1 ポイント未満の低下が見られた。

プロンプト④と Claude ベースモデルの比較では、BLEU は日英、英日で 1 ポイント強の低下、中日で 1 ポイント未満の低下、韓日で 3 ポイント弱の向上が見られた。RIBES は韓日が 1.5 ポイント低下、中日は 1.9 ポイントの向上、英日、日英は 1 ポイント未満の向上が見られた。

BLEU について詳述する(表 8.1.9-9)。

N-gram 適合率 (score1~score4)については、韓日において score1 (単語一致) が 78.6 から 79.0、79.8、score4 (長いフレーズの一致) が 34.2 から 35.1、36.2 へと向上している。これにより、韓国語特有の技術用語や請求項の定型句がより正確に出力されていることがわかる。

文長制御 (ratio / hyp\_len) とペナルティ (BP)の値からは全般的に翻訳出力が長文化する傾向があることがわかる。

表 8.1.9-9 BLEU 詳細

方向	項目	タスク	BLEU	score1	score2	score3	score4	BP	ratio	hyp_len	ref_len
cj	clm	2-2_none_claude	52.42	77.3	58.7	46	36.2	1	1.038	44030	42433
ej	clm	2-2_none_claude	52.84	77.9	59.5	46.3	36.3	1	1.014	24406	24080
je	clm	2-2_none_claude	57.95	80.9	63.6	51.7	42.6	0.999	0.999	19216	19242
kj	clm	2-2_none_claude	48.15	78.6	57.4	43.8	34.2	0.944	0.946	36910	39025
cj	clm	2-2-1_prompt3	52.8	76.7	59.1	46.6	36.8	1	1.058	44911	42433
ej	clm	2-2-1_prompt3	52.87	76.7	59.6	46.6	36.7	1	1.048	25244	24080
je	clm	2-2-1_prompt3	58.36	81	64	52	43	1	1.002	19287	19242
kj	clm	2-2-1_prompt3	49.34	79	58.3	44.7	35.1	0.952	0.953	37202	39025
cj	clm	2-2-1_prompt4	52.28	75.8	58.5	46.2	36.5	1	1.072	45468	42433
ej	clm	2-2-1_prompt4	51	75.5	57.8	44.7	34.7	1	1.05	25294	24080
je	clm	2-2-1_prompt4	56.48	80.1	62.2	50	40.8	1	1	19246	19242
kj	clm	2-2-1_prompt4	51.06	79.8	60.3	46.3	36.2	0.958	0.959	37428	39025

#### 8.1.9.4 プロンプト⑧&プロンプト⑨

プロンプト⑧とプロンプト⑨も、評価対象項目が各言語の「明細書」で共通するため、本項でまとめて結果を示す。どちらもプロンプト③に CoT を追加したモデルである(表 8.1.9-10, 表 8.1.9-11)。

表 8.1.9-10 BLEU

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	53.30	52.73	53.96	53.78	53.44
プロンプト⑧	52.21	50.99	54.46	49.55	51.80
プロンプト⑨	51.63	50.73	54.44	51.05	51.96

表 8.1.9-11 RIBES

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	88.05	87.67	89.79	89.22	88.68
プロンプト⑧	87.31	87.03	89.85	88.28	88.12
プロンプト⑨	87.51	87.13	89.70	88.61	88.23

プロンプト⑧と Claude ベースモデルの比較では、韓日の BLEU、RIBES が微増し、それ以外の言語で 1 ポイント未満の低下が見られた。

プロンプト⑨と Claude ベースモデルの比較では、BLEU は韓日が微増し、その他の言語でスコアの低下が見られた。RIBES は全ての言語で 1 ポイント未満の低下が見られた。

BLEU について詳述する(表 8.1.9-12)。

N-gram 適合率 (score1~score4)については、プロンプト⑧の日英翻訳で score1 (79.9→78.7) から score4 (38.7→34.5) まで、全ての N-gram 一致率が大幅に下がっている。特に 4-gram (score4) の低下が激しく、日本語から英語への翻訳において、このプロンプトが文構造やフレーズの自然さを大きく損ねていることがわかる。

文長制御 (ratio / hyp\_len) とペナルティ (BP)については、プロンプト⑧の日英翻訳で ratio が 0.981 から 0.966 へ低下し、hyp\_len が減少している。その結果、BP (0.980→0.966) によるペナルティが強まり、スコアをさらに押し下げている。プロンプトが情報を削りすぎてしまった可能性がある。

日英翻訳について、プロンプト⑨は⑧よりは低下幅が低いがいずれにせよプロンプトによる劣化がみえる。

唯一 BLEU が向上している韓日翻訳については、N-gram 適合率が軒並み微増しており、その結果 BLEU の向上につながっているといえる。

表 8.1.9-12 BLEU 詳細

方向	項目	タスク	BLEU	score1	score2	score3	score4	BP	ratio	hyp_len	ref_len
cj	des	2-2_none_claude	52.73	77.6	58.8	46.2	36.7	1	1.002	19711	19663
ej	des	2-2_none_claude	53.3	78.5	60.2	47.2	37.6	0.99	0.99	19526	19717
je	des	2-2_none_claude	53.78	79.9	60.9	48.1	38.7	0.98	0.981	16221	16542
kj	des	2-2_none_claude	53.96	78.2	59.4	47.5	38.5	1	1.028	20876	20309
cj	des	2-2-1_prompt8	50.99	76.5	57.1	44.4	34.9	1	1.009	19841	19663
ej	des	2-2-1_prompt8	52.21	77.6	58.8	45.7	36	0.997	0.997	19659	19717
je	des	2-2-1_prompt8	49.55	78.7	57.7	44.2	34.5	0.966	0.966	15981	16542
kj	des	2-2-1_prompt8	54.46	78.6	60	48.1	38.8	1	1.026	20844	20309
cj	des	2-2-1_prompt9	50.73	76.2	56.9	44.2	34.6	1	1.005	19761	19663
ej	des	2-2-1_prompt9	51.63	77.6	58.6	45.4	35.5	0.992	0.992	19562	19717
je	des	2-2-1_prompt9	51.05	79.6	59.2	45.9	36.2	0.965	0.966	15975	16542
kj	des	2-2-1_prompt9	54.44	78.6	59.9	48	38.8	1	1.024	20797	20309

### 8.1.10 特許由来の学習データを追加学習させたモデル

Gemma に特許由来の対訳コーパスを学習させたモデルについて、学習前のモデルと比較した。学習させたモデルは「発明の名称」「要約」「特許請求の範囲」「明細書」「拒絶理由通知書」「意見書」「補正書」それぞれ 1,500 文対ずつ学習させたモデル(f1500\_gemma)と「特許請求の範囲」「明細書」についてさらに 1,000 文追加して学習したモデル(f2500\_gemma)<sup>13</sup>、これら f1500-gemma、f2500-gemma に前述のプロンプトを組み合わせたモデル(f1500\_gemma+p1~p9、f2500\_gemma+p1~p9)を評価した(表 8.1.10-1、表 8.1.10-2)。

表 8.1.10-1 BLEU

モデル	英日	中日	韓日	日英	審査	全言語
Gemma(ベース)	41.80	45.41	48.81	46.42	45.73	45.63
f1500_gemma	42.24	45.61	48.96	45.38	46.42	45.72
f1500_gemma+p1	41.91	45.21	48.87	46.31	46.71	45.80

<sup>13</sup> f2500\_gemma は特許請求の範囲、明細書について実施、f2500\_gemma+p1, p8, p9 は明細書に対して実施、f2500\_gemma+p1, p3, p4 は特許請求の範囲に対して実施した。

f1500_gemma+p2	40.49	43.61	46.77	35.68	-	41.64
f1500_gemma+p3	41.74	46.50	47.56	51.57	-	46.84
f1500_gemma+p4	32.50	42.85	41.91	43.88	-	40.29
f1500_gemma+p8	40.05	41.85	43.48	41.30	-	41.67
f1500_gemma+p9	40.72	44.21	45.20	40.83	-	42.74
f2500_gemma	42.96	-	-	-	-	42.96
f2500_gemma+p1	44.27	-	-	-	-	44.27
f2500_gemma+p3	41.35	-	-	-	-	41.35
f2500_gemma+p4	32.99	-	-	-	-	32.99
f2500_gemma+p8	39.67	-	-	-	-	39.67
f2500_gemma+p9	40.19	-	-	-	-	40.19

表 8.1.10-2 RIBES

モデル	英日	中日	韓日	日英	審査	全言語
Gemma(ベース)	77.13	79.22	84.64	85.45	82.75	81.84
f1500_gemma	77.17	78.90	83.95	84.03	82.64	81.34
f1500_gemma+p1	77.21	78.72	83.58	85.08	83.03	81.52
f1500_gemma+p2	83.28	84.81	87.43	80.57	-	84.02
f1500_gemma+p3	58.56	62.39	73.16	85.35	-	69.87
f1500_gemma+p4	56.97	60.12	71.01	84.18	-	68.07
f1500_gemma+p8	80.50	82.89	86.13	84.58	-	83.53
f1500_gemma+p9	81.12	83.83	86.46	85.08	-	84.12
f2500_gemma	71.08	-	-	-	-	71.08
f2500_gemma+p1	71.33	-	-	-	-	71.33
f2500_gemma+p3	58.10	-	-	-	-	58.10
f2500_gemma+p4	57.58	-	-	-	-	57.58
f2500_gemma+p8	80.50	-	-	-	-	80.50
f2500_gemma+p9	81.51	-	-	-	-	81.51

学習前のモデル Gemma との比較では、学習前よりスコアが上昇したモデルが存在する一方で、スコアが低下するモデルも一定数見られた。また、学習したモデルとプロンプトの工夫を組み合わせた翻訳では、ベースモデルよりスコアが低下するものが見られた。

英日については、学習データをさらに 1000 文追加したモデルの評価を行ったが、ベースのモデルより低下するモデルも存在する結果となった。

BLEU について詳述する(表 8.1.10-3)。

ej, tit では score4 (4-gram 一致) が 24.9 から 26.5 へ向上しているものの、大幅にスコアを下げた je, tit では score1 が 68.7 から 64.9 へと大きく低下している。また、ratio が 1.125 まで膨れ上がっており、ファインチューニングによって「過度に説明的なタイトル」を出力する傾向がついてしまった可能性がある。

また英日翻訳についてみると、学習データ件数を 1500 件から 2500 件へと増やしたことで、モデルが翻訳の「正解」をより深く学んだというよりは、学習データ内のノイズや特定のクセまで学習してしまった可能性がある。しかし、追加の学習データを品質の高いコーパスとすることで訳質の向上につながる可能性は十分ある。

表 8.1.10-3 BLEU 詳細

方向	項目	タスク	BLEU	score1	score2	score3	score4	BP	ratio	hyp_len	ref_len
cj	abs	2-2_none_gemma	43.45	72.6	50.2	36.5	26.8	1	1.025	33160	32352
cj	clm	2-2_none_gemma	47.15	74.9	54	40.4	30.2	1	1.024	43467	42433
cj	des	2-2_none_gemma	46.69	74.6	53.8	40.7	31.2	0.983	0.983	19330	19663
cj	tit	2-2_none_gemma	44.33	72.1	51.4	38.5	28.7	0.986	0.986	5563	5643
ej	abs	2-2_none_gemma	41.14	72.1	49	35.4	25.9	0.97	0.97	19321	19912
ej	clm	2-2_none_gemma	42.14	75.1	51.7	37.2	26.8	0.949	0.951	22892	24080
ej	des	2-2_none_gemma	44.17	73.6	52.2	38.5	28.7	0.973	0.974	19197	19717
ej	tit	2-2_none_gemma	39.74	72.1	49	34.6	24.9	0.952	0.953	5076	5326
je	abs	2-2_none_gemma	46.43	76.2	54	40.6	31	0.973	0.974	16014	16445
je	app	2-2_none_gemma	4.74	33.6	10.3	3.1	0.5	1	1.49	2107	1414
je	clm	2-2_none_gemma	51.79	78.3	58.7	45.7	36.2	0.986	0.986	18980	19242
je	des	2-2_none_gemma	48.05	78	56.9	43.5	34	0.949	0.951	15725	16542
je	hos	2-2_none_gemma	56.56	82.8	64.7	52.5	43.3	0.958	0.959	21846	22787
je	ike	2-2_none_gemma	40.41	76.6	51.9	37.7	28.2	0.891	0.897	19030	21226
je	kyo	2-2_none_gemma	40.23	75.9	51.1	36.9	27.2	0.906	0.91	18579	20412
je	tit	2-2_none_gemma	39.4	68.7	46.3	32.5	23.3	1	1.055	4665	4423
kj	abs	2-2_none_gemma	52.14	77	58.5	46.2	36.9	0.99	0.99	25677	25944
kj	clm	2-2_none_gemma	46.65	77.2	55.6	41.9	32.2	0.951	0.952	37163	39025
kj	des	2-2_none_gemma	49.4	75.3	55.2	42.7	33.5	1	1.021	20729	20309
kj	tit	2-2_none_gemma	47.03	74.1	53.9	41.1	32.2	0.981	0.982	5154	5251
cj	abs	2-3_f1500_cj_abs	43.35	72.5	50	36.4	26.8	1	1.028	33254	32352

cj	clm	2-3_f1500_cj_clm	47.3	74.8	54.1	40.6	30.5	1	1.032	43802	42433
cj	des	2-3_f1500_cj_des	47.22	74.2	53.4	40.4	31	1	1.001	19690	19663
cj	tit	2-3_f1500_cj_tit	44.58	71.2	50.6	38	28.8	1	1.005	5674	5643
ej	abs	2-3_f1500_ej_abs	41.09	71.5	48.5	35	25.7	0.978	0.978	19471	19912
ej	clm	2-3_f1500_ej_clm	41.34	74.6	50.7	36.4	26.1	0.949	0.951	22891	24080
ej	des	2-3_f1500_ej_des	45.54	74	52.7	39.2	29.7	0.987	0.987	19453	19717
ej	tit	2-3_f1500_ej_tit	40.98	71.9	49.1	35.6	26.5	0.96	0.961	5117	5326
je	abs	2-3_f1500_je_abs	46.12	75.9	53.4	40	30.2	0.98	0.98	16117	16445
je	app	2-3_f1500_je_app	4.3	35	7.5	2.6	0.5	1	1.453	2055	1414
je	clm	2-3_f1500_je_clm	50.8	77.6	57.1	44.1	34.5	0.997	0.997	19186	19242
je	des	2-3_f1500_je_des	48.19	77.4	56.4	43.1	33.4	0.962	0.963	15932	16542
je	hos	2-3_f1500_je_hos	56.21	81.6	63.3	51.1	41.9	0.974	0.975	22213	22787
je	ike	2-3_f1500_je_ike	42.02	75.6	51.8	37.8	28.4	0.928	0.931	19755	21226
je	kyo	2-3_f1500_je_kyo	41.02	75.4	50.8	36.7	27.2	0.928	0.93	18988	20412
je	tit	2-3_f1500_je_tit	36.39	64.9	43.2	29.7	21.1	1	1.125	4976	4423
kj	abs	2-3_f1500_kj_abs	51.81	77	58.3	46	36.8	0.987	0.987	25613	25944
kj	clm	2-3_f1500_kj_clm	46.47	77.3	56	41.6	31.3	0.954	0.955	37265	39025
kj	des	2-3_f1500_kj_des	49.71	76.1	55.8	43	33.5	1	1.024	20797	20309
kj	tit	2-3_f1500_kj_tit	47.84	74.3	54.4	42.1	33	0.983	0.983	5161	5251
ej	clm	2-3_f2500_ej_clm	41.03	74.6	50.6	35.8	25.6	0.951	0.952	22932	24080
ej	des	2-3_f2500_ej_des	44.88	73.8	52.3	38.8	29.1	0.982	0.982	19364	19717

## 8.2 人手評価

大規模言語モデル（LLM モデル）の特許文献の機械翻訳への適性を測るべく、Claude、Gemini、Gemma の 3 モデルに対して、多角的な観点による人手評価を実施した。本項で、その結果をまとめる。

人手評価では、LLM モデルと比較する対照として、特許由来の対訳コーパスを大量に学習させたニューラル機械翻訳モデル（nmt）に対しても同じ条件で評価した。また、LLM モデルに前後処理を導入したモデル（3 種）、プロンプトに工夫を加えたモデル（6 種）、特許由来の学習データの追加学習を施したモデル（1 種）についても評価を行い、これらの有効性を調査した。

なお、人手評価対象文は、自動評価に用いた各言語・項目あたり 500 文のテストセット（6 章）から 100 文ずつを選定している。

### 8.2.1 人手評価の観点

人手評価は、「内容伝達レベルの評価」、「流暢さの評価」、「重要技術用語の訳質評価」、「誤訳のカテゴリ別カウント」及び「連続文での技術用語の訳ゆれの評価」の 5 つの観点で実施した。各観点の評価基準を以下に示す。

#### 8.2.1.1 内容伝達レベルの評価

内容伝達レベルの評価では、評価対象文に対し、機械翻訳結果が原文の実質的な内容をどの程度正確に伝達しているかを、参考訳文の内容に照らして、下記 5 段階の評価基準で主観的に評価した。

- |   |
|---|
| 5：すべての重要情報が正確に伝達されている。(100%)                    |
| 4：ほとんどの重要情報は正確に伝達されている。(80%～)                   |
| 3：半分以上の重要情報は正確に伝達されている。(50%～)                   |
| 2：いくつかの重要情報は正確に伝達されている。(20%～)                   |
| 1：文意がわからない、もしくは正確に伝達されている重要情報がほとんどない。<br>(～20%) |

本評価は、本事業で作成した対訳コーパスの原文を評価対象文としており、参考訳文には

その原文と対をなす翻訳文を用いた。

※「内容の伝達レベルの評価」の詳細は、特許庁「特許文献機械翻訳の品質評価手順 (ver1.0)<sup>14</sup>」の「3.1.1. 内容の伝達レベルの評価 (P.5～)」を参照のこと。

#### 8.2.1.2 流暢さの評価

流暢さの評価では、機械翻訳文の「文としての読みやすさ、理解しやすさ」のみを、下記 5 段階の基準で評価した。機械翻訳結果のみを独立した文として扱い、原文との整合性は考慮せずに日本語として自然かどうかのみを評価している。

- 5：文意が明解で、人間が書いた日本語に近い。
- 4：日本語として不自然な箇所を若干含むが、文意は明解である。
- 3：日本語として不自然な箇所があり、文意がわかりにくい。
- 2：日本語の文法規則に反する表現をかなり含む。文意がわからない。
- 1：日本語として成立していない。

※「流暢さの評価」の詳細は、特許庁「特許文献機械翻訳の品質評価手順 (ver1.0)」の「3.1.3. 流暢さの評価 (P.10～)」を参照のこと。

#### 8.2.1.3 重要技術用語の訳質評価

重要技術用語の訳質評価では、原則として全ての評価対象文について<sup>15</sup>、文中に存在する技術用語 1 語を事前に指定し、この用語が機械翻訳結果で適切に翻訳されているかを下記の評価基準 A～D で評価した。

- A (適訳語)：参考訳に照らし、技術的に同義かつ一般的に用いられる訳語である。
- B (可訳語)：技術用語として一般的に用いられる訳語ではないが、意味はおおむね正しい。
- C (誤訳語)：誤訳である。
- D (不訳語)：訳漏れ、もしくは原語のままである。

「適訳語 (評価 A)」は、技術的に妥当であれば必ずしも参考訳 (= 参考訳文で使用された訳語) と同一でなくてもよい。具体的には、検索用途に有用な訳語、すなわち検索の際に同義語展開の範疇に含められる語が該当する。これに対して「可訳語 (評価 B)」は、一般

<sup>14</sup> [https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/tokkyohonyaku\\_hyouka.html](https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/tokkyohonyaku_hyouka.html)

<sup>15</sup> 評価すべき技術用語・法律用語を含まない文は対象語なしとした。

的な訳語でないため検索用途には向かないものの、意味はおおむね通じ、照会（粗読）用途には有用である訳語を指す。

※「重要技術用語の訳質評価」の詳細は、特許庁「特許文献機械翻訳の品質評価手順（ver1.0）」の「3.1.2. 重要技術用語の翻訳精度の評価（P.8～）」を参照のこと。

#### 8.2.1.4 誤訳のカテゴリ別カウント

誤訳のカテゴリ別カウントは、機械翻訳で発生しやすい典型的な誤訳をあらかじめ類型化し、各評価対象文について、カテゴリごとの発生の有無<sup>16</sup>をカウントする評価である。具体的には、下記9種の誤訳カテゴリを設定した。

- ① 技術用語の誤訳
- ② 一般用語の誤訳
- ③ 係り受けの誤り
- ④ 訳抜け（原文中の語句が翻訳されず、欠落している。）
- ⑤ 未知語（原文中の語句（またはその一部）が翻訳されず、原語のままである。）
- ⑥ 湧き出し（原文に存在しない語句が出現している。）
- ⑦ 数値・記号エラー（原文中の数値や記号が正しく反映されていない。）
- ⑧ 訳ゆれ（文中に複数回出現する技術用語の訳語が統一されていない）
- ⑨ その他

同一のテストセットにおける各モデルの誤訳の発生頻度を類型別に集計し比較することで、モデルごとの誤訳の傾向を把握する狙いがある。

#### 8.2.1.5 連続文での技術用語の訳ゆれの評価

人手評価の一環として、同一文献から抽出した連続する2文における技術用語の訳ゆれの発生の有無をカウントする評価を実施した。この評価は、「要約」「特許請求の範囲」「明細書」の3項目について、テストセット100文の中に「同一文献から、同一の技術用語を含む、連続した2文」を9～10ペア含めて実施した。

前項に示した「誤訳のカテゴリ別カウント」にも「⑧訳ゆれ」という項目が設けられているが、誤訳のカテゴリ別カウントは個々の評価対象文単位の評価であるため、一文中に同じ技術用語が複数回使用されている文のみが評価の対象となる。だが、技術用語の訳語統一は

---

<sup>16</sup> カウントは文単位で、該当する誤訳が1つでも発生していればカウント1、発生していなければ0となる。文中で同じカテゴリの誤訳が複数回発生していてもカウントは1である。

一文中だけでなく、同一文献の異なる文間でも等しく必須である。このため、連続文を用いた本評価により、同一文献由来の異なる2文における訳ゆれの発生の有無を把握する。

#### 8.2.1.6 出願人名・発明者名の評価

人手評価では、出願人名、発明者名も評価対象とした。ただし、た個人名や法人名には「内容伝達レベル」や「流暢さ」といった文に対する評価観点はフィットしない。このため、出願人名、発明者名の評価は、個人名、法人名それぞれ以下の基準で評価した。

##### (1) 個人名

個人名の多くは日本語(漢字)名称の読みをローマ字で英字化した内容が多くみられる。ただし、漢字名称の読みは同一表記であっても複数の有力なバリエーションが存在する場  
合があり(例「宮川」⇒「Miyagawa」「Miyakawa」)、表記からはどちらが正解訳<sup>17</sup>か判断できない。このため個人名の評価では以下の基準で機械翻訳結果を評価した。

- A: 姓名とも完全一致。なお、姓名の順序や、大文字小文字、姓名間のカンマの有無などは不問とした。
- B: 不一致が一般的な名称の読みのバリエーションの範囲内であるもの。
- C: 姓名の一方は一致したが、もう一方が誤読。
- D: 姓名の双方が誤読。または、人名として訳されていない。

##### (2) 法人名

法人名に関しては、個人名のような読みのバリエーションは考慮せず、正解訳との完全一致を基本とした。ただし、法人格部分の訳語にバリエーションが想定されるため、この点を考慮して以下の評価基準で評価した。

- A: 完全一致。ただし大文字小文字、「株式会社」部分のピリオドやカンマの有無は不問とした。
- B: 「株式会社」部分の相違 (Co.Ltd、Inc、KK、Corp) などごく軽微な不一致。
- C: 正式英名がローマ字読みのところ、英訳しているケースなど、ある程度の類推が可能な不一致。
- D: 類推が不可能と思われる不一致。

---

<sup>17</sup> 個人名・法人名の正解訳は、特許庁保有の出願人発明者対応表(日英)に掲載された英名とした。

## 8.2.2 英日

英日翻訳について、5つの評価観点それぞれの評価値の4項目(発明の名称、要約、請求項、明細書)別の集計結果を各表に示す(表 8.2.2-1, 表 8.2.2-5)。

表 8.2.2-1 内容伝達レベルの評価 (英日) 集計結果

翻訳の種類/手法/モデル	発明の名称			要約			請求項			明細書			
	5	4	3	5	4	3	5	4	3	5	4	3	2
<b>大規模言語モデルによる翻訳</b>													
Claude	97	3		86	13	1	95	5		68	23	8	1
Gemini	97	3		84	15	1	89	11		72	17	11	
Gemma	88	9	3	61	38	1	73	25	2	37	31	22	10
<b>前後処理を適用した翻訳</b>													
<b>前後処理①</b>													
Claude							91	9		65	25	6	4
<b>前後処理②</b>													
Claude	97	3		83	16	1	95	5		71	20	9	
<b>前後処理⑤</b>													
Claude							93	7		66	24	9	1
<b>プロンプトを工夫した翻訳</b>													
<b>prompt1</b>													
Claude	95	5					95	5		70	23	6	1
<b>prompt2</b>													
Claude	98	2											
<b>prompt3</b>													
Claude							95	5					
<b>prompt4</b>													
Claude							97	3					
<b>prompt8</b>													
Claude										73	20	5	2
<b>prompt9</b>													
Claude										70	22	6	2
<b>追加学習済大規模言語モデル</b>													
f1500_gemma	89	8	3	63	35	2	73	23	4	36	34	17	13
<b>ニューラル機械翻訳</b>													
nmt	93	5	2	65	32	3	89	10	1	54	29	13	4

表 8.2.2-2 流暢さの評価（英日）集計結果

翻訳の種類/手法/モデル	発明の名称		要約		請求項		明細書		
	5	4	5	4	5	4	5	4	3
<b>大規模言語モデルによる翻訳</b>									
Claude	99	1	97	3	12	88	95	5	
Gemini	100		98	2	71	29	99	1	
Gemma	100		92	8	7	93	92	7	1
<b>前後処理を適用した翻訳</b>									
<b>前後処理①</b>									
Claude					11	89	95	5	
<b>前後処理②</b>									
Claude	99	1	99	1	35	65	97	3	
<b>前後処理⑤</b>									
Claude					13	87	94	6	
<b>プロンプトを工夫した翻訳</b>									
<b>prompt1</b>									
Claude	98	2			11	89	95	5	
<b>prompt2</b>									
Claude	99	1							
<b>prompt3</b>									
Claude					11	89			
<b>prompt4</b>									
Claude					12	88			
<b>prompt8</b>									
Claude							97	3	
<b>prompt9</b>									
Claude							96	4	
<b>追加学習済大規模言語モデル</b>									
f1500_gemma	98	2	88	12	7	93	95	4	1
<b>ニューラル機械翻訳</b>									
nmt	100		92	8	76	24	94	5	1

表 8.2.2-3 重要技術用語の訳質評価（英日）集計結果

翻訳の種類/手法/モデル	発明の名称			要約		請求項		明細書		
	A	B	C	A	B	A	B	A	B	C
<b>大規模言語モデルによる翻訳</b>										
Claude	94	3		89	8	99		85	8	3
Gemini	93	4		91	6	98	1	88	4	4
Gemma	88	9		82	15	95	4	79	12	5
<b>前後処理を適用した翻訳</b>										
<b>前後処理①</b>										
Claude						97	2	85	8	3
<b>前後処理②</b>										
Claude	94	3		94	3	99		83	9	4
<b>前後処理⑤</b>										
Claude						98	1	86	8	2
<b>プロンプトを工夫した翻訳</b>										
<b>prompt1</b>										
Claude	92	5				99		86	8	2
<b>prompt2</b>										
Claude	95	2								
<b>prompt3</b>										
Claude						98	1			
<b>prompt4</b>										
Claude						98	1			
<b>prompt8</b>										
Claude								88	7	1
<b>prompt9</b>										
Claude								86	9	1
<b>追加学習済大規模言語モデル</b>										
f1500_gemma	89	7	1	79	18	93	6	79	10	7
<b>ニューラル機械翻訳</b>										
nmt	94	2	1	84	13	97	2	80	11	5

表 8.2.2-4 誤訳のカテゴリ別カウント（英日）集計結果

翻訳の種類/手法/モデル/項目	①	②	③	④	⑤	⑥	⑦	⑧	⑨
<b>大規模言語モデルによる翻訳</b>									
Claude									

発明の名称			2					1	
要約	2		4		1	1			
請求項			3				1	49	
明細書	4	1	2	1	1		1	1	3
Gemini									
発明の名称	1		1	1					
要約	1	2	3	3		1	2		
請求項			3	3	1		2	11	
明細書	5	1	3	1		1	1	3	
Gemma									
発明の名称	1	1	6	1		1		3	
要約	5		7	7	2	4	1		
請求項	2	1	8	13		1	4	73	
明細書	13	2	5	8	1	3	2	1	13

### 前後処理を適用した翻訳

#### 前後処理①

Claude

請求項			2	1	1	1	1	48	
明細書	4		3	2			1	1	3

#### 前後処理②

Claude

発明の名称	1		2						
要約	1		7			1	3		
請求項	1		2				2	19	
明細書	5		1	1			1	1	2

#### 前後処理⑤

Claude

請求項	1		1	1		1	4	49
明細書	5		1	3	1	1	2	3

### プロンプトを工夫した翻訳

#### prompt1

Claude

発明の名称			2					2	
請求項			3				2	21	
明細書	3		2		1		1	1	3

#### prompt2

Claude									
発明の名称			2						
<b>prompt3</b>									
Claude									
請求項			2			2			22
<b>prompt4</b>									
Claude									
請求項	2		1				2		17
<b>prompt8</b>									
Claude									
明細書	2		2	2			1	1	1
<b>prompt9</b>									
Claude									
明細書	3		3			1	1	1	3
<b>追加学習済大規模言語モデル</b>									
f1500_gemma									
発明の名称	2		7	1		1			1
要約	7		6	5	2	6		2	
請求項	2	1	14	4		1	4		66
明細書	12	3	5	8	1	8	2	1	10
<b>ニューラル機械翻訳</b>									
nmt									
発明の名称	3	1	1	1					
要約	4		9	2	3	5			
請求項		1	5	2			1		1
明細書	6		4	3			3	2	6

表 8.2.2-5 連続文での技術用語の訳ゆれの評価（英日）集計結果

翻訳の種類/手法/モデル/項目	要約		請求項		明細書	
	訳ゆれあり	訳ゆれなし	訳ゆれあり	訳ゆれなし	訳ゆれあり	訳ゆれなし
<b>大規模言語モデルによる翻訳</b>						
Claude	2	18			18	4
Gemini	4	16	2		16	6
Gemma	2	18	4		14	8
<b>前後処理を適用した翻訳</b>						

<b>前後処理①</b>							
Claude				2	16	6	14
<b>前後処理②</b>							
Claude		2	18		18	4	16
<b>前後処理⑤</b>							
Claude					18	2	18
<b>プロンプトを工夫した翻訳</b>							
<b>prompt1</b>							
Claude					18	2	18
<b>prompt3</b>							
Claude					18		
<b>prompt4</b>							
Claude					18		
<b>prompt8</b>							
Claude						4	16
<b>prompt9</b>							
Claude						6	14
<b>追加学習済大規模言語モデル</b>							
f1500_gemma		8	12	2	16	12	8
<b>ニューラル機械翻訳</b>							
nmt		2	18	2	16	6	14

### 8.2.3 中日

中日翻訳について、5つの観点それぞれの評価値の4項目別の集計結果を各表に示す(表 8.2.3-1, 表 8.2.3-5)。

表 8.2.3-1 内容伝達レベルの評価（中日）集計結果

翻訳の種類/手法/モデル	発明の名称				要約				請求項				明細書				
	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1		
<b>大規模言語モデルによる翻訳</b>																	
Claude	86	5	6	3		86	13	1			74	18	7	1	83	15	2
Gemini	91	8		1		89	10	1			86	7	6	1	85	13	1
Gemma	73	14	7	4	2	56	31	12	1		42	36	21	1	71	21	8

**前後処理を適用した翻訳**

**前後処理①**

Claude 77 15 6 2 83 15 1 1

**前後処理⑤**

Claude 64 20 15 1 82 16 2

**プロンプトを工夫した翻訳**

**prompt1**

Claude 86 6 6 2 71 19 8 2 79 19 2

**prompt2**

Claude 90 5 3 2

**prompt3**

Claude 78 16 5 1

**prompt4**

Claude 74 15 10 1

**prompt8**

Claude 82 14 2 2

**prompt9**

Claude 78 17 3 2

**追加学習済大規模言語モデル**

f1500\_gemma 77 12 6 2 3 61 27 11 1 51 24 22 3 70 23 6 1

**ニューラル機械翻訳**

nmt 91 6 1 2 79 18 3 74 11 11 3 1 77 18 4 1

表 8.2.3-2 流暢さの評価（中日）集計結果

翻訳の種類/手法/モデル	発明の名称					要約					請求項					明細書			
	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1	5	4	3	2
<b>大規模言語モデルによる翻訳</b>																			
Claude	81	13	2	3	1	70	28	2			72	21	6	1		80	17	3	
Gemini	90	6	3	1		79	20	1			89	10	1			91	6	3	
Gemma	79	13	2	4	2	54	39	4	1	2	49	42	7	1	1	76	15	9	
<b>前後処理を適用した翻訳</b>																			
<b>前後処理①</b>																			
Claude											74	21	5			82	14	4	
<b>前後処理⑤</b>																			
Claude											66	28	5	1		84	14	2	

プロンプトを工夫した翻訳																	
<b>prompt1</b>																	
Claude	81	11	3	4	1					76	19	5		78	17	5	
<b>prompt2</b>																	
Claude	83	13	2	2													
<b>prompt3</b>																	
Claude										75	21	3	1				
<b>prompt4</b>																	
Claude										76	20	3	1				
<b>prompt8</b>																	
Claude														77	19	4	
<b>prompt9</b>																	
Claude														78	20	2	
追加学習済大規模言語モデル																	
f1500_gemma	80	11	1	4	4	56	39	4	1	54	35	10	1	79	16	4	1
ニューラル機械翻訳																	
nmt	84	12	1	2	1	61	36	2	1	64	22	10	3	1	76	19	5

表 8.2.3-3 重要技術用語の訳質評価（中日）集計結果

翻訳の種類/手法/モデル	発明の名称				要約				請求項				明細書			
	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
大規模言語モデルによる翻訳																
Claude	85	6	7	1	91	6	2		78	13	9		77	16	2	
Gemini	91	6		2	92	3	4		81	10	9		82	11	2	
Gemma	80	8	7	4	80	9	8	2	71	16	11	2	69	17	6	3
前後処理を適用した翻訳																
<b>前後処理①</b>																
Claude									82	7	11		83	10	1	1
<b>前後処理⑤</b>																
Claude									73	12	15		84	10	1	
プロンプトを工夫した翻訳																
<b>prompt1</b>																
Claude	86	7	4	2					77	14	9		76	16	2	1
<b>prompt2</b>																
Claude	88	7	4													



請求項	20		1	2		1	69
明細書	9	2	2	2			1
<b>前後処理⑤</b>							
Claude							
請求項	27			3	3		61
明細書	5	3	1	1	1	1	1
<b>プロンプトを工夫した翻訳</b>							
<b>prompt1</b>							
Claude							
発明の名称	11		1	1	1		4
請求項	18			1			57
明細書	7	3		2		1	2
<b>prompt2</b>							
Claude							
発明の名称	10						2
<b>prompt3</b>							
Claude							
請求項	16			1			54
<b>prompt4</b>							
Claude							
請求項	21	1	1	2	6		30
<b>prompt8</b>							
Claude							
明細書	7	5	1		1	1	2
<b>prompt9</b>							
Claude							
明細書	6	3	3	1	1	1	4
<b>追加学習済大規模言語モデル</b>							
f1500_gemma							
発明の名称	20	1	3	1	4		3
要約	25	4	10	2	1	1	7
請求項	35		4	9	5	4	74
明細書	13	3	3	1	5	1	2
<b>ニューラル機械翻訳</b>							
nmt							
発明の名称	8		1				2

要約	11	2	2	12	1	1	1	2
請求項	14		2	2	1	1	3	52
明細書	8	3		2		1		1

表 8.2.3-5 連続文での技術用語の訳ゆれの評価（中日）集計結果

翻訳の種類/手法/モデル/項目	要約		請求項		明細書	
	訳ゆれあり	訳ゆれなし	訳ゆれあり	訳ゆれなし	訳ゆれあり	訳ゆれなし
<b>大規模言語モデルによる翻訳</b>						
Claude	4	16	4	16	2	18
Gemini	4	16	4	16	2	18
Gemma		20	6	14		20
<b>前後処理を適用した翻訳</b>						
<b>前後処理①</b>						
Claude			2	18	4	16
<b>前後処理⑤</b>						
Claude			6	14	8	12
<b>プロンプトを工夫した翻訳</b>						
<b>prompt1</b>						
Claude			4	16	4	16
<b>prompt3</b>						
Claude			2	18		
<b>prompt4</b>						
Claude				20		
<b>prompt8</b>						
Claude					6	14
<b>prompt9</b>						
Claude					6	14
<b>追加学習済大規模言語モデル</b>						
f1500_gemma	2	18	4	16	2	18
<b>ニューラル機械翻訳</b>						
nmt	4	16	4	16	2	18

## 8.2.4 韓日

韓日翻訳について、5つの観点それぞれの評価値の4項目別の集計結果を各表に示す(表 8.2.4-1, 表 8.2.4-5)。

表 8.2.4-1 内容伝達レベルの評価（韓日）集計結果

翻訳の種類/手法/モデル	発明の名称			要約			請求項			明細書		
	5	4	3	5	4	3	5	4	3	5	4	3
<b>大規模言語モデルによる翻訳</b>												
Claude	98	2		94	5	1	88	10	2	96	4	
Gemini	98	2		94	6		90	9	1	95	5	
Gemma	88	9	3	77	19	4	77	19	4	69	26	5
<b>前後処理を適用した翻訳</b>												
<b>前後処理①</b>												
Claude							90	9	1	95	5	
<b>前後処理⑤</b>												
Claude							86	13	1	92	8	
<b>プロンプトを工夫した翻訳</b>												
<b>prompt1</b>												
Claude	97	2	1				91	8	1	96	3	1
<b>prompt2</b>												
Claude	97	3										
<b>prompt3</b>												
Claude							89	10	1			
<b>prompt4</b>												
Claude							91	8	1			
<b>prompt8</b>												
Claude										93	5	2
<b>prompt9</b>												
Claude										95	3	2
<b>追加学習済大規模言語モデル</b>												
f1500_gemma	88	10	2	75	20	5	72	23	5	77	19	4
<b>ニューラル機械翻訳</b>												
nmt	90	9	1	84	15	1	76	17	7	91	9	

表 8.2.4-2 流暢さの評価（韓日）集計結果

翻訳の種類/手法/モデル	発明の名称		要約		請求項			明細書	
	5	4	5	4	5	4	3	5	4
<b>大規模言語モデルによる翻訳</b>									
Claude	100		95	5	92	8		93	7
Gemini	100		98	2	94	6		98	2
Gemma	100		97	3	92	7	1	93	7
<b>前後処理を適用した翻訳</b>									
<b>前後処理①</b>									
Claude					92	8		92	8
<b>前後処理⑤</b>									
Claude					92	8		95	5
<b>プロンプトを工夫した翻訳</b>									
<b>prompt1</b>									
Claude	100				94	6		93	7
<b>prompt2</b>									
Claude	100								
<b>prompt3</b>									
Claude					93	7			
<b>prompt4</b>									
Claude					97	3			
<b>prompt8</b>									
Claude								96	4
<b>prompt9</b>									
Claude								97	3
<b>追加学習済大規模言語モデル</b>									
f1500_gemma	99	1	98	2	91	8	1	94	6
<b>ニューラル機械翻訳</b>									
nmt	100		95	5	90	10		96	4

表 8.2.4-3 重要技術用語の訳質評価（韓日）集計結果

翻訳の種類/手法/モデル	発明の名称			要約			請求項				明細書		
	A	B	C	A	B	C	A	B	C	D	A	B	C
<b>大規模言語モデルによる翻訳</b>													
Claude	91	6		90	8	1	94	4	2		91	3	1

Gemini	91	6		90	7	2	95	4	1		91	3	1
Gemma	88	6	3	84	12	3	82	11	6	1	80	6	9
<b>前後処理を適用した翻訳</b>													
<b>前後処理①</b>													
Claude							92	6	2		92	2	1
<b>前後処理⑤</b>													
Claude							96	1	3		91	2	2
<b>プロンプトを工夫した翻訳</b>													
<b>prompt1</b>													
Claude	92	5					94	4	2		91	3	1
<b>prompt2</b>													
Claude	93	4											
<b>prompt3</b>													
Claude							94	4	2				
<b>prompt4</b>													
Claude							93	3	4				
<b>prompt8</b>													
Claude											89	3	3
<b>prompt9</b>													
Claude											91	3	1
<b>追加学習済大規模言語モデル</b>													
f1500_gemma	85	9	3	84	12	3	83	10	7		82	6	7
<b>ニューラル機械翻訳</b>													
nmt	90	6	1	93	5	1	91	6	2	1	93	1	1

表 8.2.4-4 誤訳のカテゴリ別カウント（韓日）集計結果

翻訳の種類/手法/モデル/項目	①	②	③	④	⑤	⑥	⑦	⑧	⑨
<b>大規模言語モデルによる翻訳</b>									
Claude									
発明の名称	1		1					1	
要約	3	2	1			1	1		
請求項	8	3			1				
明細書	2	2							
Gemini									
発明の名称			1					1	

要約	4	1	2			2		
請求項	6	3	1					
明細書	3		1			1		
Gemma								
発明の名称	5		2	3				
要約	16	3	2	6	1	1	1	
請求項	16	1	4	2	1		1	5
明細書	23		1	3	2		1	1

### 前後処理を適用した翻訳

#### 前後処理①

Claude

請求項	9	1						
明細書	4	1				2		

#### 前後処理⑤

Claude

請求項	7	2	2		1			3
明細書	4	3	1			2		

### プロンプトを工夫した翻訳

#### prompt1

Claude

発明の名称	1		2					
請求項	8				1			
明細書	2	2						

#### prompt2

Claude

発明の名称	2		1				1	
-------	---	--	---	--	--	--	---	--

#### prompt3

Claude

請求項	9	1			1			1
-----	---	---	--	--	---	--	--	---

#### prompt4

Claude

請求項	9	2		1				1
-----	---	---	--	---	--	--	--	---

#### prompt8

Claude

明細書	5		1					1
-----	---	--	---	--	--	--	--	---

#### prompt9

Claude								
明細書	2			1			1	1
<b>追加学習済大規模言語モデル</b>								
f1500_gemma								
発明の名称	6		3	1	1			
要約	17	3	2	4		1	2	
請求項	17	2	9	3				1 11
明細書	14	2	1	1			1	2
<b>ニューラル機械翻訳</b>								
nmt								
発明の名称	5		1	4				
要約	6	2	3	3			3	
請求項	13	1	7	7		1	2	1
明細書	6	1		3			4	1

表 8.2.4-5 連続文での技術用語の訳ゆれの評価（韓日）集計結果

翻訳の種類/手法/モデル/項目	要約		請求項		明細書	
	訳ゆれあり	訳ゆれなし	訳ゆれあり	訳ゆれなし	訳ゆれあり	訳ゆれなし
<b>大規模言語モデルによる翻訳</b>						
Claude	2	18	2	18		20
Gemini	2	18		20	2	18
Gemma	6	14	6	14	2	18
<b>前後処理を適用した翻訳</b>						
<b>前後処理①</b>						
Claude			4	16		20
<b>前後処理⑤</b>						
Claude			2	18	4	16
<b>プロンプトを工夫した翻訳</b>						
<b>prompt1</b>						
Claude			2	18		20
<b>prompt3</b>						
Claude			4	16		
<b>prompt4</b>						
Claude			2	18		
<b>prompt8</b>						

Claude							20
<b>prompt9</b>							
Claude							20
<b>追加学習済大規模言語モデル</b>							
f1500_gemma	6	14	10	10	2		18
<b>ニューラル機械翻訳</b>							
nmt	2	18	6	14	4		16

## 8.2.5 日英

日英翻訳について、5つの観点それぞれの評価値の4項目別の集計結果を各表に示す(表 8.2.5-1, 表 8.2.5-5)。

表 8.2.5-1 内容伝達レベルの評価（日英）集計結果

翻訳の種類/手法/モデル	発明の名称			要約			請求項			明細書			
	5	4	3	5	4	3	5	4	3	5	4	3	2
<b>大規模言語モデルによる翻訳</b>													
Claude	99	1		87	13		89	9	2	81	13	5	1
Gemini	94	6		90	10		90	9	1	70	24	5	1
Gemma	85	12	3	75	25		65	22	13	59	27	6	8
<b>前後処理を適用した翻訳</b>													
<b>前後処理①</b>													
Claude							89	9	2	77	17	5	1
<b>前後処理⑤</b>													
Claude							95	5		78	18	4	
<b>プロンプトを工夫した翻訳</b>													
<b>prompt1</b>													
Claude	97	3					93	7		85	10	3	2
<b>prompt2</b>													
Claude	97	2	1										
<b>prompt3</b>													
Claude							95	5					
<b>prompt4</b>													
Claude							94	6					

<b>prompt8</b>													
Claude										70	23	6	1
<b>prompt9</b>													
Claude										75	18	6	1
<b>追加学習済大規模言語モデル</b>													
f1500_gemma	84	11	5	69	31	61	23	16	61	21	13	5	
<b>ニューラル機械翻訳</b>													
nmt	93	7		75	24	1	87	11	2	58	28	13	1

表 8.2.5-2 流暢さの評価（日英）集計結果

翻訳の種類/手法/モデル	発明の名称		要約		請求項			明細書				
	5	4	5	4	5	4	3	5	4	3	2	
<b>大規模言語モデルによる翻訳</b>												
Claude	99	1	99	1	90	8	2	94	6			
Gemini	100		98	2	99	1		93	7			
Gemma	99	1	97	3	76	21	3	87	11	2		
<b>前後処理を適用した翻訳</b>												
<b>前後処理①</b>												
Claude					91	8	1	91	9			
<b>前後処理⑤</b>												
Claude					94	6		89	11			
<b>プロンプトを工夫した翻訳</b>												
<b>prompt1</b>												
Claude	99	1			94	5	1	95	5			
<b>prompt2</b>												
Claude	99	1										
<b>prompt3</b>												
Claude					95	5						
<b>prompt4</b>												
Claude					96	4						
<b>prompt8</b>												
Claude								89	11			
<b>prompt9</b>												
Claude								92	8			
<b>追加学習済大規模言語モデル</b>												

f1500_gemma	100	98	2	69	25	6	88	11	1	
<b>ニューラル機械翻訳</b>										
nmt	99	1	95	5	98	2	78	19	2	1

表 8.2.5-3 重要技術用語の訳質評価（日英）集計結果

翻訳の種類/手法/モデル	発明の名称				要約			請求項		明細書				
	A	B	C	D	A	B	C	A	B	A	B	C	D	
<b>大規模言語モデルによる翻訳</b>														
Claude	95	2			95	1	1	99		88		8		
Gemini	92	4		1	97			98	1	89		6	1	
Gemma	86	8	3		90	6	1	96	3	84		8	3	1
<b>前後処理を適用した翻訳</b>														
<b>前後処理①</b>														
Claude								99		84		10	2	
<b>前後処理⑤</b>														
Claude								99		88		8		
<b>プロンプトを工夫した翻訳</b>														
<b>prompt1</b>														
Claude	94	3						99		89		7		
<b>prompt2</b>														
Claude	94	3												
<b>prompt3</b>														
Claude								99						
<b>prompt4</b>														
Claude								99						
<b>prompt8</b>														
Claude										89		7		
<b>prompt9</b>														
Claude										90		6		
<b>追加学習済大規模言語モデル</b>														
f1500_gemma	86	8	2	1	90	6	1	95	4	83		8	5	
<b>ニューラル機械翻訳</b>														
nmt	92	4	1		90	7		99		90		4	2	

表 8.2.5-4 誤訳のカテゴリ別カウント（日英）集計結果

翻訳の種類/手法/モデル/項目	①	②	③	④	⑤	⑥	⑦	⑧	⑨
<b>大規模言語モデルによる翻訳</b>									
Claude									
発明の名称			1						1
要約	2	1	3				1		
請求項	1		6				2		
明細書	1	2	2				1		2
Gemini									
発明の名称	1		2	1		12			3
要約			2				1		
請求項	2		2	1		2	1		
明細書	3		2				1		1
Gemma									
発明の名称	6		3	3		23			1
要約	4	2	6	5		1			
請求項	5		16	11		2	2		1
明細書	5	2	1	9	1		1		5
<b>前後処理を適用した翻訳</b>									
<b>前後処理①</b>									
Claude									
請求項			9	1			1		1
明細書	2	1	2				1		1
<b>前後処理⑤</b>									
Claude									
請求項	1		3				2		1
明細書	2	1	1				1		
<b>プロンプトを工夫した翻訳</b>									
<b>prompt1</b>									
Claude									
発明の名称			2						1
請求項	1		4	1			1		
明細書	1	1	2				1		2
<b>prompt2</b>									
Claude									
発明の名称			2	1					1

**prompt3**

Claude

請求項	1	2	1	1
-----	---	---	---	---

**prompt4**

Claude

請求項	1	3	1	2
-----	---	---	---	---

**prompt8**

Claude

明細書	2	1	3	2	2
-----	---	---	---	---	---

**prompt9**

Claude

明細書	2	1	2	1	1	2
-----	---	---	---	---	---	---

**追加学習済大規模言語モデル**

f1500\_gemma

発明の名称	5	6	3	33			
要約	3	1	5	5	2	1	
請求項	1	25	3	1	3	1	2
明細書	9	3	2	6	1	2	7

**ニューラル機械翻訳**

nmt

発明の名称	1	1	1			
要約	1	1	7	2	3	2
請求項	1	1	1	5	1	
明細書	5	2	2	1	4	5

表 8.2.5-5 連続文での技術用語の訳ゆれの評価（日英）集計結果

翻訳の種類/手法/モデル/項目	要約		請求項		明細書	
	訳ゆれあり	訳ゆれなし	訳ゆれあり	訳ゆれなし	訳ゆれあり	訳ゆれなし
<b>大規模言語モデルによる翻訳</b>						
Claude		20	4	14	2	18
Gemini		20		18	4	16
Gemma	2	18		18	4	16
<b>前後処理を適用した翻訳</b>						
<b>前後処理①</b>						
Claude			6	12	4	16

<b>前後処理⑤</b>						
Claude			2	16	2	18
<b>プロンプトを工夫した翻訳</b>						
<b>prompt1</b>						
Claude			4	14	2	18
<b>prompt3</b>						
Claude			4	14		
<b>prompt4</b>						
Claude			2	16		
<b>prompt8</b>						
Claude					4	16
<b>prompt9</b>						
Claude					2	18
<b>追加学習済大規模言語モデル</b>						
f1500_gemma		2	18	2	16	6 14
<b>ニューラル機械翻訳</b>						
nmt		6	14	4	14	2 18

### 8.2.6 日英(拒絶理由、意見書、補正書、出願人名/発明者名)

日英翻訳については、前項に示した特許文献由来の4項目とは別途に、審査書類3種(拒絶理由通知書、意見書、補正書)各100文、及び出願人/発明者名100件についても評価対象とした。本項に、審査書類及び出願人/発明者名に対して実施した各評価の集計結果を示す(表 8.2.6-1, 表 8.2.6-6)。

表 8.2.6-1 内容伝達レベルの評価(日英審査書類)集計結果

翻訳の種類/手法/モデル	拒絶理由				意見書			補正書			
	5	4	3	2	5	4	3	5	4	3	2
<b>大規模言語モデルによる翻訳</b>											
Claude	83	16	1		89	11		84	8		8
Gemini	77	22	1		93	7		88	10		2
Gemma	59	26	14	1	61	36	3	52	29		19
<b>前後処理を適用した翻訳</b>											
<b>前後処理②</b>											

Claude	86	13	1	91	9	87	8	5		
<b>追加学習済大規模言語モデル</b>										
f1500_gemma	52	29	17	2	70	26	4	45	24	31
<b>ニューラル機械翻訳</b>										
nmt	77	19	4	81	18	1	57	30	12	1

表 8.2.6-2 流暢さの評価（日英審査書類）集計結果

翻訳の種類/手法/モデル	拒絶理由			意見書			補正書		
	5	4	3	5	4	5	4	3	
<b>大規模言語モデルによる翻訳</b>									
Claude	80	19	1	93	7	68	28	4	
Gemini	77	21	2	94	6	77	23		
Gemma	76	20	4	87	13	44	50	6	
<b>前後処理を適用した翻訳</b>									
<b>前後処理②</b>									
Claude	80	18	2	93	7	73	26	1	
<b>追加学習済大規模言語モデル</b>									
f1500_gemma	73	23	4	87	13	33	55	12	
<b>ニューラル機械翻訳</b>									
nmt	80	16	4	86	14	48	46	6	

表 8.2.6-3 重要技術用語の訳質評価（日英審査書類）集計結果

翻訳の種類/手法/モデル	拒絶理由				意見書				補正書		
	A	B	C	D	A	B	C	D	A	B	D
<b>大規模言語モデルによる翻訳</b>											
Claude	86	12			88	7	1		95	1	
Gemini	90	7	1		92	3	1		95	1	
Gemma	85	12	1		81	11	3	1	89	7	
<b>前後処理を適用した翻訳</b>											
<b>前後処理②</b>											
Claude	89	9			90	5	1		95	1	
<b>追加学習済大規模言語モデル</b>											
f1500_gemma	81	15	1	1	80	11	5		90	6	
<b>ニューラル機械翻訳</b>											

nmt 90 8 91 5 90 4 2

表 8.2.6-4 誤訳のカテゴリ別カウント（日英審査書類）集計結果

翻訳の種類/手法/モデル/項目	①	②	③	④	⑤	⑥	⑦	⑧	⑨
<b>大規模言語モデルによる翻訳</b>									
Claude									
拒絶理由	10	2		1			4		3
意見書	13		3				2		2
補正書	3	6	5	4			1		6
Gemini									
拒絶理由	10	4	2	2		1	1		4
意見書	8	1	1	1			1		1
補正書	2	8	2	2			1		2
Gemma									
拒絶理由	10	11	7	8		2	5		1
意見書	15	8	10	18		2	2	1	2
補正書	9	14	7	20		3	2		12
<b>前後処理を適用した翻訳</b>									
前後処理②									
Claude									
拒絶理由	11	3	3	2			1		3
意見書	8		3	2			1		1
補正書	3	5	3	5		1			3
<b>追加学習済大規模言語モデル</b>									
f1500_gemma									
拒絶理由	13	7	10	6		3	6		2
意見書	24	10	2	12			2		1
補正書	10	15	12	16		4		1	23
<b>ニューラル機械翻訳</b>									
nmt									
拒絶理由	3	7	3	1					3
意見書	4	3	4	2		2	2		7
補正書	3	9	8	15		2	2	1	3

表 8.2.6-5 連続文での技術用語の訳ゆれの評価（日英審査書類）集計結果

翻訳の種類/手法/モデル/項目	補正書		意見書		拒絶理由	
	訳ゆれあり	訳ゆれなし	訳ゆれあり	訳ゆれなし	訳ゆれあり	訳ゆれなし
<b>大規模言語モデルによる翻訳</b>						
Claude	4	16		20	2	18
Gemini	4	16	2	18		20
Gemma	6	14	2	18		20
<b>前後処理を適用した翻訳</b>						
<b>前後処理②</b>						
Claude	4	16		20	4	16
<b>追加学習済大規模言語モデル</b>						
f1500_gemma	6	14	4	16	2	18
<b>ニューラル機械翻訳</b>						
nmt	4	16		20	2	18

表 8.2.6-6 出願人名/発明者の評価（日英）集計結果

翻訳の種類/手法/モデル	企業				個人			
	A	B	C	D	A	B	C	D
<b>大規模言語モデルによる翻訳</b>								
Claude	10	16	13	11	35	8	7	
Gemini	10	22	8	10	39	9	2	
Gemma	5	19	7	19	17	5	24	4
<b>前後処理を適用した翻訳</b>								
<b>前後処理②</b>								
Claude	11	13	13	13	37	7	6	
<b>追加学習済大規模言語モデル</b>								
f1500_gemma	7	21	7	15	15	8	23	4
<b>ニューラル機械翻訳</b>								
nmt	2	13	7	28	27	8	4	11

## 8.2.7 特許文献4項目別の集計結果

8.2.2～5 に示した言語別の各評価の集計結果を項目ごと（発明の名称／要約／特許請求

の範囲／明細書）に合算した結果を本項に示す(表 8.2.7-1, 表 8.2.7-5)。

前後処理②は英日のみを対象としており、集計母数が異なることから合算の対象外とした。

表 8.2.7-1 内容伝達レベルの評価（項目別合算値）集計結果

翻訳の種類/手法/モデル	発明の名称					要約					請求項					明細書				
	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1
<b>大規模言語モデルによる翻訳</b>																				
Claude	380	11	6	3		353	44	3			346	42	11	1		328	55	15	2	
Gemini	380	19		1		357	41	2			355	36	8	1		322	59	17	2	
Gemma	334	44	16	4	2	269	113	17		1	257	102	40	1		236	105	41	18	
<b>前後処理を適用した翻訳</b>																				
<b>前後処理①</b>																				
Claude											347	42	9	2		320	62	12	6	
<b>前後処理⑤</b>																				
Claude											338	45	16	1		318	66	15	1	
<b>プロンプトを工夫した翻訳</b>																				
<b>prompt1</b>																				
Claude	375	16	7	2							350	39	9	2		330	55	12	3	
<b>prompt2</b>																				
Claude	382	12	4	2																
<b>prompt3</b>																				
Claude											357	36	6	1						
<b>prompt4</b>																				
Claude											356	32	11	1						
<b>prompt8</b>																				
Claude																318	62	15	5	
<b>prompt9</b>																				
Claude																318	60	17	5	
<b>追加学習済大規模言語モデル</b>																				
f1500_gemma	338	41	16	2	3	268	113	18	1		257	93	47	3		244	97	40	18	1
<b>ニューラル機械翻訳</b>																				
nmt	367	27	4	2		303	89	8			326	49	21	3	1	280	84	30	6	

表 8.2.7-2 流暢さの評価（項目別合算値）集計結果

翻訳の種類/手法/モデル	発明の名称					要約					請求項					明細書				
	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1	5	4	3	2	
<b>大規模言語モデルによる翻訳</b>																				
Claude	379	15	2	3	1	361	37	2			332	59	8	1		362	35	3		
Gemini	390	6	3	1		373	26	1			378	21	1			381	16	3		
Gemma	378	14	2	4	2	340	53	4	1	2	294	93	11	1	1	348	40	12		
<b>前後処理を適用した翻訳</b>																				
<b>前後処理①</b>																				
Claude											338	56	6			360	36	4		
<b>前後処理⑤</b>																				
Claude											335	59	5	1		362	36	2		
<b>プロンプトを工夫した翻訳</b>																				
<b>prompt1</b>																				
Claude	378	14	3	4	1						338	56	6			361	34	5		
<b>prompt2</b>																				
Claude	381	15	2	2																
<b>prompt3</b>																				
Claude											331	65	3	1						
<b>prompt4</b>																				
Claude											343	53	3	1						
<b>prompt8</b>																				
Claude																359	37	4		
<b>prompt9</b>																				
Claude																363	35	2		
<b>追加学習済大規模言語モデル</b>																				
f1500_gemma	377	14	1	4	4	340	55	4	1		290	92	17	1		356	37	6	1	
<b>ニューラル機械翻訳</b>																				
nmt	383	13	1	2	1	343	54	2	1		334	52	10	3	1	344	47	8	1	

表 8.2.7-3 重要技術用語の訳質評価（項目別合算値）集計結果

翻訳の種類/手法/モデル	tit				abs				clm				des			
	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
<b>大規模言語モデルによる翻訳</b>																
Claude	365	17	7	1	365	23	4		370	17	11		341	35	6	
Gemini	367	20		3	370	16	6		372	16	10		350	24	8	

Gemma	342	31	13	4	336	42	12	2	344	34	17	3	312	43	23	4
<b>前後処理を適用した翻訳</b>																
<b>前後処理①</b>																
Claude									370	15	13		344	30	7	1
<b>前後処理⑤</b>																
Claude									366	14	18		349	28	5	
<b>プロンプトを工夫した翻訳</b>																
<b>prompt1</b>																
Claude	364	20	4	2					369	18	11		342	34	5	1
<b>prompt2</b>																
Claude	370	16	4													
<b>prompt3</b>																
Claude									376	15	7					
<b>prompt4</b>																
Claude									369	15	14					
<b>prompt8</b>																
Claude													347	28	7	
<b>prompt9</b>																
Claude													343	34	4	1
<b>追加学習済大規模言語モデル</b>																
f1500_gemma	336	34	17	3	332	46	13	1	340	35	22	1	310	46	25	1
<b>ニューラル機械翻訳</b>																
nmt	370	15	5		360	31	1		371	15	11	1	349	21	12	

表 8.2.7-4 誤訳のカテゴリ別カウント（項目別合算値）集計結果

翻訳の種類/手法/モデル/項目	①	②	③	④	⑤	⑥	⑦	⑧	⑨
<b>大規模言語モデルによる翻訳</b>									
Claude									
発明の名称	13		4	1	1			1	4
要約	15	6	10	1	2	2	2		2
請求項	25	3	9	1	2		3	1	128
明細書	13	8	4	1	1		2	2	7
Gemini									
発明の名称	7		4	3	1	12		1	5
要約	13	4	5	7		1	5		5

請求項	22	3	5	6	2	3	3	37	
明細書	18	1	5	2		1	3	5	
Gemma									
発明の名称	30	1	12	11	3	25		6	
要約	52	7	20	25	6	10	2	6	
請求項	63	5	29	38	6	11	6	2	153
明細書	51	5	8	24	10	7	4	1	22

### 前後処理を適用した翻訳

#### 前後処理①

Claude

請求項	29	1	11	3	3	1	3	118	
明細書	19	4	5	4	2		4	1	5

#### 前後処理⑤

Claude

請求項	36	2	6	1	4	4	6	114
明細書	16	7	3	4	2	2	6	4

### プロンプトを工夫した翻訳

#### prompt1

Claude

発明の名称	12		6	1	1	1		7	
請求項	27		7	1	2		3	78	
明細書	13	6	4		3		2	2	7

#### prompt2

Claude

発明の名称	12		5	1				1	3
-------	----	--	---	---	--	--	--	---	---

#### prompt3

Claude

請求項	26	1	4	1	2		3	77
-----	----	---	---	---	---	--	---	----

#### prompt4

Claude

請求項	32	3	2	5	2	7	4	48
-----	----	---	---	---	---	---	---	----

#### prompt8

Claude

明細書	16	6	6	3		1	4	1	6
-----	----	---	---	---	--	---	---	---	---

#### prompt9

Claude

明細書	13	4	5	5	1	2	4	1	10
<b>追加学習済大規模言語モデル</b>									
f1500_gemma									
発明の名称	33	1	16	8	2	38			4
要約	52	8	13	24	4	10	4	2	7
請求項	55	3	52	19	5	6	7	2	153
明細書	48	11	8	18	2	14	6	1	21
<b>ニューラル機械翻訳</b>									
nmt									
発明の名称	17	1	3	6		1			2
要約	22	5	21	19	4	8	6	1	2
請求項	28	3	15	16	1	3	4	4	53
明細書	25	6	6	9			12	3	12

表 8.2.6-5 連続文での技術用語の訳ゆれの評価（日英審査書類）集計結果

翻訳の種類/手法/モデル/項目	補正書		意見書		拒絶理由	
	訳ゆれあり	訳ゆれなし	訳ゆれあり	訳ゆれなし	訳ゆれあり	訳ゆれなし
<b>大規模言語モデルによる翻訳</b>						
Claude	4	16			20	2
Gemini	4	16	2		18	20
Gemma	6	14	2		18	20
<b>前後処理を適用した翻訳</b>						
<b>前後処理②</b>						
Claude	4	16			20	4
<b>追加学習済大規模言語モデル</b>						
f1500_gemma	6	14	4		16	2
<b>ニューラル機械翻訳</b>						
nmt	4	16			20	2

## 8.2.8 大規模言語モデル 3 モデル及び対照 NMT モデルの比較

人手評価を実施した 3 種の大規模言語モデル（Claude、Gemini、Gemma）と、比較のためのニューラル翻訳モデル（nmt）の全 4 モデルについて、本項で評価観点ごとに評価

結果を比較する。

これら4モデルは、まず特許文献として各言語（英日、中日、韓日、日英）とも「発明の名称」「要約」「特許請求の範囲」「明細書」の4項目から100文ずつ、計400文の機械翻訳結果を評価した。本項に示す各言語の評価結果は、これら4項目、全400文の集計結果となる。また、これとは別途に日英審査書類として「拒絶理由通知書」「意見書」「補正書」からそれぞれ100文ずつ評価しており、これら全300文の集計結果も表中「審査」欄に示した。

なお、日英出願人・発明者として個人名50件、法人名50件の計100件も人手評価の対象としているが、評価基準が他と異なるため一覧表には含めず、後述の第8.2.8.5項にて単独で取り上げる。

#### 8.2.8.1 内容伝達レベルの評価結果の比較

下表に、各モデルの内容伝達レベルの評価結果について、言語ごとに評価対象全件の平均レベルを算出した結果をまとめた(表8.2.8-1)。

表 8.2.8-1 [内容伝達レベルの評価] 言語別の全項目平均スコア

モデル	英日	中日	韓日	日英	審査	全言語
Claude	4.84	4.76	4.93	4.87	4.82	4.84
Gemini	4.83	4.84	4.94	4.84	4.85	4.86
Gemma	4.53	4.44	4.74	4.62	4.45	4.56
nmt	4.69	4.71	4.74	4.74	4.65	4.71

内容伝達レベルの評価は、本調査で実施した人手評価においては「総合評価」の位置づけとなる。高評価であるほど、「外国語文献の内容把握」という目的に対する実用性が高い機械翻訳文と見なせる。

内容伝達レベルは5段階評価であるが、LLM3モデルの平均レベルはいずれもレベル4を大きく上回っており、ベースモデルの状態でも高精度な特許翻訳文を生成できることが示された。特に Claude と Gemini に関しては、特許由来の大量の対訳コーパスを学習させたニューラル機械翻訳モデル (nmt) を上回る高評価を得ており、特許翻訳用エンジンとして十分な実用性を有すると判断できる。Claude と Gemini の相対順位は言語により入れ替わるが、各言語とも僅差であり、内容伝達の観点からはほぼ同等と見なせる。

Gemma も各言語でレベル 4 を大きく上回っており品質は良好である。ただし、他のモデルと比較すると各言語とも Gemini や Claude とはやや差があり、また多くの言語で nmt より下位となった。この結果から、特許翻訳用途への適合性においては、Claude、Gemini のほうが優ると判断される。

英日、中日、韓日、日英の比較では、各モデルとも韓日の平均レベルが最も高く、次いで Claude と Gemma では日英、英日、中日と続く。Gemini では中日と日英が同スコアで英日が最下位となったが、その差は 0.01 ポイントと僅差である。nmt を含め、各モデルとも各言語の平均スコアはごく狭い範囲に収まっており、言語間の差は最大でも Gemma の 0.30 ポイント（韓日 4.74、中日 4.44）であった。各言語のテストセットは当然ながら内容が異なり、翻訳難易度は完全に均一ではない。このため言語間の相対比較では評価値に一定の幅を想定するべきであり、本評価に見られる言語間の評価値の差はその範囲内と見なすべき小差である。つまり、本評価結果からは、各モデルとも（評価対象とした）全ての言語について翻訳品質の差が見いだされることはなかった。

前掲の 8.2.2～6 の各表に示したとおり、Claude と Gemini ではレベル 2 以下を付された低品質の機械翻訳文はほとんど検出されず、Claude では 400 文中レベル 2 が 6 文（英日 1、中日 4、日英 1）、Gemini もレベル 2 が 4 文（中日 3、日英 1）とごく少数であった。これに対し、Gemma はレベル 2 以下が 26 文（英日 10、中日 8、日英 8）あり、うち中日の 3 文にはレベル 1 が付された。また nmt はレベル 2 以下が 12 文（英日 4、中日 7、日英 1）あった。

Gemma でレベル 1 が付された 3 文はいずれも文の主題となる技術用語が中国語のまま（つまり「未知語」として）出力されたことが理由であった。以下、一例を示す。

[CNA112352046\_JPA22024026241\_tit]

原文	麦角硫因、抗坏血酸 2-葡萄糖苷、抗坏血酸和它们的组合的成肌分化促进作用
正解訳文	エルゴチオネイン、アスコルビン酸 2-グルコシド、アスコルビン酸、およびそれらの組み合わせの筋分化促進作用
Claude レベル：5	エルゴチオネイン、アスコルビン酸 2-グルコシド、アスコルビン酸およびそれらの組み合わせの筋分化促進作用
Gemini レベル：5	エルゴチオネイン、アスコルビン酸 2-グルコシド、アスコルビン酸及びそれらの組み合わせの筋分化促進作用
Gemma レベル：1	麦角硫因、抗坏血酸 2-グルコシド、抗坏血酸およびそれらの組み合わせの筋分化促進作用

nmt レベル：5	エルゴチオネイン、アスコルビン酸 2-グルコシド、アスコルビン酸及びこれらの組み合わせの筋分化促進作用
--------------	---

上例では、Gemma の機械翻訳文のみ「**麦角硫因**」「**抗坏血**」と中国語がそのまま出力されており、他の 2 モデル（及び nmt モデル）がレベル 5 を付されるなか、レベル 1 の低評価が付された。レベル 1 が付された他の 2 文も同様に未知語が発生しており、Gemma と他の 2 モデルとの評価差の一因がこうした「技術用語のカバー範囲の違い」による可能性を示唆する結果である。

一方、複数のモデルで低評価が付された文も見られた。このような文は、原文自体の翻訳難易度が高かったと見なせる。以下、一例を示す。

[CNA114318617\_JPA22024543675\_clm]

原文	根据权利要求 1~4 中任一项所述的一种 <b>网络复合丝</b> 的应用，其特征是：将所述网络复合丝作为经纱，制成织物。
正解訳文	請求項 1~4 のいずれか一項に記載の <b>インターレース複合糸</b> の使用であって、前記インターレース複合糸を経糸として織物を作製する、ことを特徴とするインターレース複合糸の使用。
Claude レベル：2	請求項 1~4 のいずれか一項に記載の <b>ネットワーク複合糸</b> の応用であって、前記ネットワーク複合糸を経糸として織物を製造することを特徴とする。
Gemini レベル：2	前記 <b>ネットワーク複合糸</b> を経糸として織物を製織することを特徴とする、請求項 1~4 のいずれか一項に記載のネットワーク複合糸の使用。
Gemma レベル：2	権利請求項 1~4 のいずれか一項に記載の <b>ネットワーク複合糸</b> の用途であり、前記ネットワーク複合糸を経糸として織物を作製することを特徴とする。
nmt レベル：2	請求項 1~4 のいずれか一項に記載の <b>ネットワーク複合糸</b> の応用であって、その特徴は以下のとおりである：このネットワーク複合糸を経糸として織物とした。

本例では全モデルにレベル 2 が付された。各モデルとも発明の主題に関わる「**网络复合丝/インターレース複合糸**」を「**ネットワーク複合糸**」と意味不明に誤訳しており、これが重大視された結果である。LLM モデルだけでなく、膨大な学習データを用いる nmt でも同じ形で誤訳されていることから、この語自体が特許文献での使用例が乏しい難訳語とわかる。つまり本例は、原文の翻訳難易度の高かったことが各モデルの低評価の原因といえる。

日英人手評価では、審査書類についても「拒絶理由通知書」「意見書」「補正書」の3種それぞれ100文ずつ、全300文を評価した。審査書類全件の平均レベルは前表に示したとおりであり、nmtを含めた4モデルいずれもレベル4を大きく上回る高評価であり、特に Gemini と Claude は4.8以上で nmt (4.65) より高評価、一方 Gemma は4.45で nmt よりやや低評価となった。こうした状況は各言語の特許文献400文の評価結果と同じであり、各LLMモデル、特に Gemini と Claude は審査書類の翻訳にも十分に実用可能といえる。

### 8.2.8.2 流暢さの評価結果の比較

流暢さの評価では、生成された機械翻訳文が日本語文として自然で読み易いか否かのみを評価している。このため、原文の内容が正しく反映されているか（つまり内容の伝達レベル）は問われない。結果を下表に示す(表 8.2.8-2)。

表 8.2.8-2 [流暢さの評価] 言語別の全項目平均スコア

モデル	英日	中日	韓日	日英	審査	全言語
Claude	4.92	4.70	4.95	4.95	4.79	4.86
Gemini	4.98	4.85	4.98	4.98	4.82	4.92
Gemma	4.90	4.52	4.95	4.89	4.66	4.78
nmt	4.92	4.73	4.94	4.92	4.68	4.84

モデル間の比較では全言語で Gemini が最上位となった。ただし、各モデル、各言語とも平均スコアは非常に高い水準にあり、極端に低スコアのモデルや言語は見られない。対照の nmt モデルと比べても遜色はなく、翻訳文の流暢さに関しては、各モデルとも十分に実用に足ると判断される。

### 8.2.8.3 重要技術用語の訳質評価結果の比較

重要技術用語の訳質評価では、テストセットの原文から1文につき1語、評価対象の技術用語をあらかじめ選定し、これが適切に翻訳されているかを評価した。判定はA（適訳語）、B（可訳語）、C（誤訳語）、D（不訳語）のいずれかとなる。各モデルの評価結果の詳細は前掲の8.2.2～6の各表に示したとおりである。本項では、各モデルのA判定率すなわち「重要技術用語が適訳語に訳された文の比率」を重視し、その結果を言語別に算出した。下表に示す(表 8.2.8-3)。

表 8.2.8-3 [重要技術用語の評価] 言語別のA判定率

モデル	英日	中日	韓日	日英	審査	全言語
Claude	94.3%	84.2%	93.6%	96.9%	92.8%	92.4%
Gemini	95.1%	88.0%	93.9%	96.7%	95.5%	93.8%
Gemma	88.4%	76.3%	85.4%	91.5%	87.9%	85.9%
nmt	91.3%	90.8%	93.9%	95.4%	93.4%	93.0%

A 判定率は、特許文献に用いられる技術用語を適訳語で翻訳できるカバー率を示す数値といえる。特許由来の大量の学習データを用いる nmt が得意とするところであり、各言語とも 90%を超えた。これに対し、LLM モデルの Gemini と Claude は、中日のみ nmt に及ばなかったものの、他の 4 言語は nmt と同水準もしくはやや上回る数値となり、全言語平均も nmt と同等となった。よってこれら 2 モデルは、技術用語の訳質の観点においても十分に実用に足るレベルと判断される。

一方、Gemma は各言語で A 判定率が 5~10 ポイントほど低かった。内容伝達レベルの評価では、Gemma のみ他のモデルが正しく翻訳できる技術用語が未知語となったケースを例示し、技術用語のカバー範囲が他のモデルに比べて狭い可能性を述べたが、本評価もこの推論を肯定する結果となった。

#### 8.2.8.4 誤訳のカテゴリ別カウント／連続文での技術用語の訳ゆれの評価結果の比較

表 8.2.8-4 [誤訳のカテゴリ別カウント] 及び [連続文の訳ゆれ] の言語別カウント数

モデル	言語	誤訳のカテゴリ別カウント									連続文 訳ゆれ
		①	②	③	④	⑤	⑥	⑦	⑧	⑨	
Claude	英日	6	1	11	1	2	1	2	1	53	3/29
	中日	42	6	2	3	3	0	0	2	85	5/30
	韓日	14	7	2	0	1	1	1	1	0	2/30
	日英	4	3	12	0	0	0	4	0	3	3/29
	審査	26	8	8	5	0	0	7	0	11	3/30
	合計	92	25	35	9	6	2	14	4	152	16/148
Gemini	英日	7	3	10	8	1	2	5	0	14	6/29
	中日	34	1	0	4	2	1	0	0	34	5/30
	韓日	13	4	1	4	0	0	3	1	0	2/30
	日英	6	0	8	2	0	14	3	0	4	2/29
	審査	20	13	5	5	0	1	3	0	7	3/30
	合計	80	21	24	23	3	18	14	1	59	18/148

Gemma	英日	21	4	26	29	3	9	7	1	89	7/29
	中日	95	6	8	27	17	16	0	1	85	3/30
	韓日	60	4	9	14	4	1	2	1	6	7/30
	日英	20	4	26	28	1	26	3	0	7	3/29
	審査	34	33	24	46	0	7	9	1	15	4/30
	合計	230	51	93	144	25	59	21	4	202	24/148
nmt	英日	13	2	19	8	3	5	4	2	7	5/29
	中日	41	5	4	17	2	1	3	4	57	5/30
	韓日	30	4	11	17	0	1	9	2	0	6/30
	日英	8	4	13	8	0	5	6	0	5	6/29
	審査	10	19	15	18	0	4	4	1	13	3/30
	合計	102	34	62	68	5	16	26	9	82	25/148

#### [誤訳のカテゴリ別カウント]

誤訳のカテゴリ別カウント(表 8.2.8-4)では、基本的には「① 技術用語の誤訳」のカウント数が各モデル、各言語とも多くなった。高度な技術用語が多数使われる特許文の評価では技術用語の誤訳のカウントが多くなるのは一般的な現象であり、事実、nmt モデルでも同じ傾向が見られる。各モデルの「① 技術用語の誤訳」のカウント数を比べると、全言語を通じて Gemini のカウント数が最も少なく、Claude も小差で続くが、Gemma のみカウント数が明らかに多い。この結果からも、3 種の LLM モデルの技術用語のカバレッジに差があることが見て取れる。なお、Claude、Gemini とも「技術用語の誤訳」のカウント数は nmt よりやや少なく、前述の「重要技術用語の評価」の状況と整合する結果となった。

誤訳カテゴリ②～⑧についても、概して Claude と Gemini は nmt よりカウントが少なく、Gemma は nmt よりややカウントが多い傾向が見て取れる。この結果は、内容伝達レベルにおける LLM の 3 モデルと nmt モデルの相対順位とも整合する。特に Gemma は英日・日英の「③ 係り受け誤り」と「④ 訳抜け」、中日の「⑤ 未知語」と「⑥ 湧き出し」のカウントが nmt より大幅に多く、このモデルの誤訳の傾向と見なせる。

Claude と Gemini のカウント数は各カテゴリともおおむね少差であるが、強いて挙げれば Gemini の「④ 訳抜け」のカウント数が各言語で Claude より多い。この結果からは、Gemini のほうがやや訳抜けが発生する確率が高いと推測される。特にカウント差が大きい英日から Gemini の訳抜け判定事例を二例、Claude と対比して示す。

[JPA2020509323\_USAA2024115284\_abs]

原文	Baffles are connected to one upper edge of the troughs, while an air flow space is positioned over the other <b>upper trough edge</b> .
正解訳文	バッフルはトラフの1つの上縁に接続され、一方、気流空間は他の <b>上側トラフ端縁</b> の上方に位置付けられている。
Gemini レベル:5	バッフルがトラフの一方の上縁に接続されている一方、空気流路が他方の <b>上縁</b> の上方に配置されている。
Claude レベル:5	バッフルはトラフの一方の上端縁に接続され、一方、エアフロー空間は他方の <b>上部トラフ縁</b> の上方に位置している。

[JPA2022088181\_USAA2024003095\_abs]

原文	A wet sheet which is hard to be fractured <b>during production into a molded body</b> , stable in shape, and easy to handle in processing, and a method for producing a molded body from the wet sheet.
正解訳文	<b>成形体を製造する上で</b> 破断し難く、形状が安定しており、加工する上での取り扱いが容易な湿潤シート及び、当該湿潤シートから製造される成形体の製造方法を提供する。
Gemini レベル:4	<b>成形時に</b> 破断し難く、形状安定性に優れ、加工時の取扱いが容易な湿潤シート、及び該湿潤シートから成形体を製造する方法。
Claude レベル:5	<b>成形体への製造時に</b> 破断しにくく、形状が安定し、加工時の取り扱いが容易な湿潤シートと、該湿潤シートから成形体を製造する方法。

Gemini の翻訳文は、1 例目は原文に「**upper trough edge**／**上部トラフ縁**」とあるところを「**上縁**」、2 例目は「**during production into a molded body**／**成形体を製造する上で**」とあるところ「**成形時に**」と、一部文言が欠落している。ただし、1 例目については冒頭に「one upper edge of the troughs／トラフの一方の上縁」とあり、ここは不足なく訳されているため、「**上縁**」だけでもトラフの上縁であることは読み取れる（このため内容伝達レベルは5が付されている）。2 例目も「**成形体を製造する**」と「**成形時に**」は（本例では）実質的に同じ意味といえ、両例とも訳すべき語の欠落というよりは、やや極端な省略の結果である可能性が高い。Gemini で訳抜けと判定された文には他にも同様のケースが見られ、このモデルは他に比して文言を積極的に省略をする傾向があると見なせる。これが、原文に忠実な逐語訳をする Claude とのカウント差に表れている。

なお、各モデルとも英日及び中日の「⑨ その他」が極端に多くカウントされた。これは「特許請求の範囲」において「機械翻訳文が体言止めの文体となっていない（日本の記載様

式に沿っていない)」ことに対する指摘が大半であった。例えば上表に示した各モデルの英日の「⑨ その他」の全カウントのうち「特許請求の範囲」由来の文でのカウントは、Claude が 53 文中の 49 文、Gemini が 14 文中の 11 文、Gemma が 89 文中の 73 文と大半を占めており、かつこれら全てが「請求項が体言止めになっていない」という指摘であった。中日もほぼ同様の状況であり、「⑨ その他」にカウントされた「特許請求の範囲」由来の文のうち、Claude では 85 文中の 74 文、Gemini は 34 文中の 21 文、Gemma は 85 文中の 63 文がこの事象によるものであった。特に Claude と Gemma で頻発しており、Gemini はこれら 2 モデルに比べて発生数は少ない。また、各モデルとも韓日ではこの事象はほとんど検出されなかった（Gemma で 2 文のみ）。対照である nmt モデルでは、英日は 1 文のみの発生と LLM モデルに比べて少量であったが、中日では 51 文と Claude や Gemma に近い頻度で多発しており、言語による差が激しい。

以下、中日の事例を一例示す。

[CNA107405077\_JPA22024153662\_clm]

原文	如权利要求 5 所述的射频引导线，其还包括套环结构，所述套环结构靠近所述尖端结构定位。
正解訳文	前記チップ構造に近接して位置するカラー構造をさらに含む、請求項 5 に記載の高周波ガイドワイヤ。
Claude レベル：4	請求項 5 に記載の高周波ガイドワイヤであって、前記先端構造の近くに位置するスリーブリング構造をさらに含む。
Gemini レベル：5	請求項 5 に記載の高周波ガイドワイヤであって、さらにカラー構造を備え、前記カラー構造は前記先端構造の近傍に配置される。
Gemma レベル：4	請求項 5 に記載の RF ガイドワイヤであって、先端構造に近接して位置するループ構造を備えることを特徴とする、RF ガイドワイヤ。
nmt レベル：4	請求項 5 に記載の無線周波数ガイドワイヤであって、それはさらにカラー構造を含み、前記カラー構造は前記先端構造に近接して位置決めされる。

本例では、Gemma を除く 3 モデル（Claude、Gemini、nmt）の機械翻訳文が体言止めで訳されておらず、「⑨ その他」にカウントされた。本来、この不備は Gemma で最も多発しており、また Gemini では発生頻度は少ないが、本例のようなケースも存在する。総じて、各モデルにおける本事象の発生はランダム的であった。

なお、この「請求項が体言止めになっていない」という事象は、文体、体裁に対する課題であり、内容理解への悪影響は少ない。事実、上例も Gemma を除く 3 モデル（nmt を含む）で請求項が体言止めとなっていないが、それらの内容伝達レベルも 4、5、4 と良好

である<sup>18</sup>。実用性の観点からは軽微な不備と見なせる<sup>19</sup>。

中日に関しては、「⑨その他」にカテゴリズされた特有の不備として、「訳出不要の『一种』が『一つの』等と訳出される」事象が各モデルで複数件見られた。中国語の「一种」は英語の不定冠詞 a/an と同様、通常は翻訳不要である。こうした翻訳不要の「一种」が「一つの」「一種の」等と訳出されるケースが散見された。一例を示す。

[tit64 : CNA113279221\_JPA22024517392\_tit]

原文	一种洗衣机上盖及洗衣机
正解訳文	洗濯機の上蓋及び洗濯機
機械翻訳文 (Gemini)	一種の洗濯機用上蓋及び洗濯機

上例では原文冒頭の「一种」を正解訳文では訳出していないが、機械翻訳文 (Gemini) では「一種の」と訳出しており、これが不備と判定された。発生数は Claude で 3 件、Gemini で 6 件、Gemma で 8 件と比較的少量であり、また内容理解への悪影響も小さいが、やや不自然な文となる。

[連続文での技術用語の訳ゆれの評価]

人手評価の一環として、同一文献から抽出した連続する 2 文における技術用語の訳ゆれの発生の有無をカウントする評価を実施したことは先に述べた (⇒8.2.1.5)。前表の最右欄には、「連続文訳ゆれ」として、そのカウント結果も掲載した。

この評価は、「要約」「特許請求の範囲」「明細書」の 3 項目について、テストセット 100 文の中に「同一文献から、同一の技術用語を含む、連続した 2 文」を 10 ペア含めて実施した。したがって評価対象は 3 項目×10 ペア=30 ペアとなる<sup>20</sup>。例えばこの欄が「5/30」となっていれば、30 ペア中 5 ペアで技術用語の訳ゆれが検出されたことを示す。

結果は前表のとおりであり、各 LLM モデルとも文間の訳ゆれは少なからず発生していることが判明した。ニューラル機械翻訳で文間の訳ゆれが生じる理由は、用語辞書を用いず

<sup>18</sup> 上例の内容伝達レベルは Claude、Gemma、nmt でそれぞれレベル 4 であるが、これは体言止めへの不対応が原因ではなく、それぞれ別箇所の誤訳 (Claude : スリープリング構造、Gemma : ループ構造、nmt : 無線周波数) による減点である。

<sup>19</sup> 「体言止めで訳文を出力せよ」といったプロンプトを指定することで、改善できる可能性がある。

<sup>20</sup> ただし事情により、英日・日英のみ「特許請求の範囲」が 9 ペア、全項目合計 29 ペアを評価した。

(つまり訳語を固定せず) 学習データからその都度最良と判断した訳語を選択するという方法論と、文単位の翻訳処理であるため先行する文の訳語情報が引き継がれないという事情による。今回評価した各 LLM モデルもこれらの条件は同じであるため、同様に訳ゆれが発生したと考えられる。

#### 8.2.8.5 出願人・発明者の評価結果の比較

出願人／発明者名の評価では、法人名と個人名をそれぞれ 50 件ずつ、計 100 件のデータについて評価した。評価対象とした翻訳方向は日本語⇒英語である。法人名・個人名は文の評価に用いた評価観点がフィットしないため、専用の評価基準を定めた。

法人名については、正解英訳と完全一致したものを A 評価とし、名称の本体部分は一致したが「株式会社」など法人格を示す略語が正解英訳と異なるものは B 評価とした。「株式会社」の略語には「Co.,Ltd.」「Corp」「Inc.」「KK」など有力なもの複数存在するが、意味は同じであるため、本評価では準正解として扱った。

一方、C～D 評価は不正解にあたるが、機械翻訳結果から一定の類推が可能であれば C 評価、不可能であれば D 評価と区別した。具体的には、例えば正解英訳がローマ字表記であるのに英訳されているケース（「日本キッチン工業株式会社／NIPPON KITCHEN KOGYO KK」⇒「Japan Kitchen Industries Co., Ltd.」）や、名称の主要な部分に合致が見られるケース（「平河ヒューテック株式会社／HIRAKAWA HEWTECH CORP」⇒「Hirakawa Hutec Co., Ltd.」）などは C 評価とし、名称の主要部分が合致しないなど類推が困難と見なされるケース（「東日本産業株式会社／HIGASHINIHON-SANGYO CO LTD」⇒「Tohoku Industries Co., Ltd.」）は D 評価とした。

個人名についても、姓・名とも完全一致したものを A 評価とし、姓・名が不一致であっても、一般的な読みのバリエーションの範囲内（「東 直親／HIGASHI NAOCHIKA」⇒「Azuma Naochika」）であれば、準正解とみなして B 評価とした。C～D 評価は不正解に相当するが、姓・名のいずれかが合致していれば C、姓名の双方が不一致のものや、人名として訳されていないものは D とした。

法人名、個人名とも、A 評価（正解）から B 評価（準正解）までが実用レベルの翻訳結果と考えられる。各モデルにおける A 評価+B 評価の比率（A～B 率）は下表のとおりである（表 8.2.8-5）。

表 8.2.8-5 出願人名／発明者の評価における A～B 率

モデル	法人名			個人名			合算 A～B 率
	評価 A	評価 B	AB 率	評価 A	評価 B	AB 率	
Claude	10	16	52.0%	35	8	86.0%	69.0%
Gemini	10	22	64.0%	39	9	96.0%	80.0%
Gemma	5	19	48.0%	17	5	44.0%	46.0%
nmt	2	13	30.0%	27	8	70.0%	50.0%

対照に用いた nmt は、特許文献由来の学習データのみを用いたニューラル機械翻訳モデルである。特許文献には法人名や個人名の記載が少ないため、本モデルは固有名詞の翻訳には不向きである。結果、nmt の法人・個人合算の A～B 率は 50.0%にとどまった。これに対し、大規模翻訳モデルの Gemini は 80.0%、Claude は 69.0%と、いずれも nmt より明らかに高い翻訳精度を示した。唯一、Gemma は AB 率 48.0%と nmt と同水準にとどまった。

各モデルとも、法人名では A 評価より B 評価のほうが多く、個人名では A 評価のほうが B 評価より多かった。これは、法人名は「株式会社」の略称に多数のバリエーションがあり、完全一致（A 評価）の難易度が個人名よりも高かったためと推測される。また、A～B 率も Gemma 以外は法人名よりも個人名のほうが顕著に高い。これは法人名と個人名の翻訳難易度の差によるものと見なせる。日本人個人名の英訳がローマ字表記に限定されるのに対し、日本法人名は必ずしもローマ字表記とは限らず、社名の一部を英訳する場合（「財団法人川村理化学研究所／KAWAMURA INST OF CHEM RES」）や変則的な英名（「T s Z 株式会社／THREE SIX ZERO CO LTD」）である場合があり、法則性に乏しい。このため、難易度が高いものが少なからず存在する。

## 8.2.9 前後処理を導入したモデル

人手評価の一環として、大規模言語モデルによる機械翻訳に、特許翻訳に適すると考えられる前後処理を導入した場合の翻訳品質を評価し、ベースモデルからの変化の度合いを確認した。本項でその結果をまとめる。

本評価では、大規模言語モデル Claude をベースモデルとし、このモデルに下記 3 種の前後処理(表 8.2.9-1)を導入した場合に機械翻訳文の翻訳品質水準がどのように変化するかを調べた。

表 8.2.9-1 人手評価の対象とした前後処理一覧

前後処理	言語方向、項目	理由
前後処理①	言語：英日・中日・韓日・日英 項目：特許請求の範囲・明細書	翻訳対象文から LLM を用いて重要語を抽出、当該重要語を含むコーパスを当社保有のコーパス群から選択し、FewShot の例に含めて翻訳を実施。
前後処理②	言語：英日・日英 項目：英は特許請求の範囲・明細書・要約・発明の名称 日は拒絶理由通知、意見書、補正書、出願人・発明者の名称	翻訳対象文を翻訳する際の参照用文対を LLM を用いて作成する前処理
前後処理⑤	言語：英日・中日・韓日・日英 項目：特許請求の範囲・明細書	翻訳対象文を翻訳する際の参照用文対を当社保有コーパスから類似検索した上で選択し、FewShot の例に含めて翻訳を実施。

前後処理を導入したモデルの人手評価は、上表のとおり、前後処理ごとに特定の言語、項目について実施した。以下に示す各前後処理モデルの評価結果は、それぞれの対象項目の合計であるため、対照するベースモデルの評価結果も、同一の対象項目のみの合計値とした。

#### 8.2.9.1 前後処理①

前後処理①は、翻訳対象文から LLM を用いて重要語を抽出し、当該語を含むコーパスを対応特許由来の対訳コーパス群から選択して、FewShot の例に含めて翻訳を実施するというものである。この処理を導入した前後処理①モデルについて、英日、中日、韓日、日英それぞれ「特許請求の範囲」と「明細書」の2項目、200文を評価した。下の各表に、前後処理①モデルの各評価の集計結果を、ベースモデル Claude との対比で示す(表 8.2.9-2～表 8.2.9-5)。比較条件を合わせるため、ベースモデルの評価結果も「特許請求の範囲」と「明細書」の2項目のみの合算値とした。

表 8.2.9-2 [内容伝達レベルの評価] 言語別の全項目平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.77	4.73	4.91	4.81	4.81
前後処理①	4.71	4.74	4.92	4.79	4.79

表 8.2.9-3 [流暢さの評価] 言語別の全項目平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.87	4.71	4.93	4.91	4.86
前後処理①	4.88	4.74	4.92	4.91	4.86

表 8.2.9-4 [重要技術用語の評価] 言語別の A 判定率

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	94.3%	79.5%	94.9%	95.9%	91.2%
前後処理①	94.4%	84.6%	94.4%	93.8%	91.8%

表 8.2.9-5 [誤訳のカテゴリ別カウント] 及び [連続文の訳ゆれ] の言語別カウント数

モデル	言語	誤訳のカテゴリ別カウント									連続文 訳ゆれ
		①	②	③	④	⑤	⑥	⑦	⑧	⑨	
Claude (ベース)	英日	4	1	5	1	1	0	2	1	52	2/19
	中日	22	3	0	1	1	0	0	2	81	3/20
	韓日	10	5	0	0	1	0	0	0	0	1/20
	日英	2	2	8	0	0	0	3	0	2	3/19
	合計	38	11	13	2	3	0	5	3	135	9/78
前後処理①	英日	4	0	5	3	1	1	2	1	51	4/19
	中日	29	2	0	3	4	0	1	0	70	3/20
	韓日	13	2	0	0	0	0	2	0	0	2/20
	日英	2	1	11	1	0	0	2	0	2	5/19
	合計	48	5	16	7	5	1	7	1	123	14/78

前後処理①モデルと Claude ベースモデルの比較では、各評価において言語や項目ごとに改善（青字）と低下（赤字）が混在し、かつこれらの変動も概して少量にとどまった。前後処理①は LLM の few shot に特許由来の対訳コーパスを追加することで技術用語や特許表現の改善を図る効果を狙ったものであるが、ベースモデル自体の評価値が高いこともあり、各観点の評価において前後処理①を導入した特段の効果が表れている項目はほとんど見られない。ただし、中日の重要技術用語の評価の A 判定率のみは、ベースモデルの 79.5% から 84.6% と 5.1 ポイント向上している。ベースモデルの中日 A 判定率は他の言語に比べて顕著に低く、改善の余地が大きかった。このため、前後処理①で対訳コーパスが追加された効果が顕在化したと考えられる。

前後処理①は、この「中日の重要技術用語の適訳語化」において、翻訳水準（特に技術用語の訳質）の改善に一定の効果を示した。これ以外の項目については特段の改善は見られなかったが、これは Claude はベースモデルの時点で翻訳水準が非常に高く、改善の余地はごく小さかったためとも考えられる。

### 8.2.9.2 前後処理②

前後処理②は、Claude ベースモデルに対し、対象文を翻訳する際の参照用文対を LLM を用いて作成し提供する前処理である。この処理を導入した前後処理②モデルについて、英日の全 4 項目（「発明の名称」「要約」「特許請求の範囲」「明細書」）、及び審査書類（日英）の全 3 種（「拒絶理由通知」「意見書」「補正書」）を対象に各評価を実施した。下表にその結果をベースモデル Claude と対比して示す(表 8.2.9-6～表 8.2.9-9)。

表 8.2.9-6 [内容伝達レベルの評価] 言語別の全項目平均スコア

モデル	英日	審査（日英）	全言語
Claude（ベース）	4.84	4.82	4.83
前後処理②	4.84	4.86	4.85

表 8.2.9-7 [流暢さの評価] 言語別の全項目平均スコア

モデル	英日	審査（日英）	全言語
Claude（ベース）	4.92	4.79	4.86
前後処理②	4.97	4.81	4.89

表 8.2.9-8 [重要技術用語の評価] 言語別の A 判定率

モデル	英日	審査（日英）	全言語
Claude（ベース）	94.3%	92.8%	93.6%
前後処理②	95.1%	94.5%	94.8%

表 8.2.9-9 [誤訳のカテゴリ別カウント] 及び [連続文の訳ゆれ] の言語別カウント数

モデル	言語	誤訳のカテゴリ別カウント									連続文 訳ゆれ
		①	②	③	④	⑤	⑥	⑦	⑧	⑨	
Claude (ベース)	英日	6	1	11	1	2	1	2	1	53	3/19
	審査	26	8	8	5	0	0	7	0	11	3/30
	合計	32	9	19	6	2	1	9	1	64	6/49
前後処理②	英日	8	0	12	1	0	1	6	1	51	3/19
	審査	22	8	9	9	0	1	2	0	7	4/30

	合計	30	8	21	10	0	2	8	1	58	7/49
--	----	----	---	----	----	---	---	---	---	----	------

前後処理②モデルの評価値は、特許文献（英日）、審査書類（日英）とも、大半の項目でベースモデルより数値が向上している。マイナス方向の変化は、誤訳カテゴリによってカウントが増えたものが数種あり、連続文での訳ゆれが若干増えた程度であり、総合的にはベースモデルより翻訳品質は高いと判断される。ただし、評価対象言語や項目が前後処理①とほとんど重ならないため、どちらがより有効かは今回の評価からは確定できない。

なお、前後処理②モデルは 8.2.8.5 項で述べた出願人・発明者 100 件の評価も実施した。下表にその集計結果をベースモデルとの対比で示す(表 8.2.9-10)。

表 8.2.9-10 出願人名／発明者の評価における A～B 率

モデル	法人名			個人名			合算 A～B 率
	評価 A	評価 B	AB 率	評価 A	評価 B	AB 率	
Claude（ベース）	10	16	52.0%	35	8	86.0%	69.0%
前後処理②	11	13	48.0%	37	7	88.0%	68.0%

そもそも前後処理②は出願人・発明者名の訳質改善に寄与する性質のものではなく、評価値も上表のとおり誤差の範囲の微変動のみであった。

### 8.2.9.3 前後処理⑤

前後処理⑤は、対照文を翻訳する際の参照用文対を対応特許由来の対訳コーパスから類似検索して取得し、FewShot の例に含める前処理である。この前処理を導入した前後処理⑤モデルは、前後処理①と同様、英日、中日、韓日、日英それぞれで「特許請求の範囲」と「明細書」の 2 項目、200 文を評価対象とした。下の各表に、前後処理⑤モデルの各評価の集計結果を、ベースモデル Claude の評価結果との対比で示す(表 8.2.9-11～表 8.2.9-14)。

表 8.2.9-11 [内容伝達レベルの評価] 言語別の全項目平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude（ベース）	4.77	4.73	4.91	4.81	4.81
前後処理⑤	4.74	4.64	4.89	4.85	4.78

表 8.2.9-12 [流暢さの評価] 言語別の全項目平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.87	4.71	4.93	4.91	4.86
前後処理⑤	4.89	4.71	4.94	4.92	4.87

表 8.2.9-13 [重要技術用語の評価] 言語別の A 判定率

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	94.3%	79.5%	94.9%	95.9%	91.2%
前後処理⑤	94.4%	80.5%	95.9%	95.9%	91.7%

表 8.2.9-14 [誤訳のカテゴリ別カウント] 及び [連続文の訳ゆれ] の言語別カウント数

モデル	言語	誤訳のカテゴリ別カウント									連続文 訳ゆれ
		①	②	③	④	⑤	⑥	⑦	⑧	⑨	
Claude (ベース)	英日	4	1	5	1	1	0	2	1	52	2/19
	中日	22	3	0	1	1	0	0	2	81	3/20
	韓日	10	5	0	0	1	0	0	0	0	1/20
	日英	2	2	8	0	0	0	3	0	2	3/19
	合計	38	11	13	2	3	0	5	3	135	9/78
前後処理⑤	英日	6	0	2	4	1	2	6	0	52	1/19
	中日	32	3	0	1	4	4	1	0	62	7/20
	韓日	11	5	3	0	1	0	2	0	3	3/20
	日英	3	1	4	0	0	0	3	0	1	2/19
	合計	52	9	9	5	6	6	12	0	118	13/78

前後処理⑤の評価結果は、前後処理①と同様、評価観点や言語、誤訳カテゴリによりベースモデルから数値が上昇したものと下降したものが混在し、効果の判断が難しい。例えば、前処理の導入により「重要技術用語の A 判定率」が 4 言語中 3 言語で向上しているが、その反面、誤訳のカテゴリ別カウントの「① 技術用語の誤訳」は各言語で増加し、内容伝達レベルも 3 言語で低下するなど、関連の深い評価観点の間でも効果と逆効果が混在する状況である。ただし、いずれも少量の変動であり、前後処理①の中日技術用語のように顕著に改善した項目は見られない。

## 8.2.10 プロンプトに工夫を施したモデル

LLM モデルでの機械翻訳時に投入するプロンプトに工夫を施すことで、より良い翻訳結果を得られる可能性がある。人手評価では、Claude をベースモデルとし、プロンプトに工

夫を施した下記6種のモデルについて評価した(表 8.2.10-1)。

表 8.2.10-1 人手評価の対象としたプロンプト一覧

モデル名	言語方向、項目	プロンプトの概要
プロンプト①	言語：英日・中日・韓日・日英 項目：特許請求の範囲・明細書・ 発明の名称	翻訳対象の項目(発明の名称、要約、 請求の範囲、明細書等)および、想 定読者に「特許のサーチャー」を指 定。
プロンプト②	言語：英日・中日・韓日・日英 項目：発明の名称	技術分野の情報として、IPC の説明 文(サブクラス)を指定。
プロンプト③	言語：英日・中日・韓日・日英 項目：特許請求の範囲	プロンプト①+プロンプト②
プロンプト④	言語：英日・中日・韓日・日英 項目：特許請求の範囲	プロンプト③に Hierarchical Translation(文を解析して主語、動 詞、目的語という中核構造を把握 し、まずはその関係性を崩さぬよう 翻訳し、そのうえで改良すべき箇所 や流暢さなどを整えていく段階的 な翻訳)を行う CoT(Chain of Thought)を追加。
プロンプト⑧	言語：英日・中日・韓日・日英 項目：明細書	プロンプト③に Translation Explanation(文の中核構造を把握 し、それにフィットする文体や訳語 を選び、その理由を説明することで 妥当性を担保)を行う CoT を追加。
プロンプト⑨	言語：英日・中日・韓日・日英 項目：明細書	プロンプト③に Structural Transformation(入力文の構造を 解析し、最適の文構造や統語構造 に適宜言い換えて翻訳)を行う CoT を追加。

プロンプトに工夫を施した各モデルの人手評価は、それぞれ上表に示した言語と項目に  
対して実施した。対照するベースモデルの評価結果も同じ項目のみの合算値とした。

### 8.2.10.1 プロンプト①

プロンプト①では、翻訳対象の項目（発明の名称、要約、請求の範囲、明細書等）および、想定読者に「特許のサーチャー」を指定した。このプロンプトを Claude に用いた機械翻訳文の評価結果を各表に示す(表 8.2.10-2～表 8.2.10-5)。評価対象項目は各言語とも「発明の名称」「特許請求の範囲」「明細書」の3項目である。

表 8.2.10-2 [内容伝達レベルの評価] 言語別の全項目平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.83	4.73	4.93	4.87	4.84
プロンプト①	4.84	4.71	4.94	4.89	4.85

表 8.2.10-3 [流暢さの評価] 言語別の全項目平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.91	4.70	4.95	4.94	4.88
プロンプト①	4.89	4.70	4.96	4.96	4.88

表 8.2.10-4 [重要技術用語の評価] 言語別の A 判定率

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	95.2%	81.6%	94.5%	96.6%	92.0%
プロンプト①	94.9%	81.3%	94.9%	96.6%	91.9%

表 8.2.10-5 [誤訳のカテゴリ別カウント] 及び [連続文の訳ゆれ] の言語別カウント数

モデル	言語	誤訳のカテゴリ別カウント									連続文 訳ゆれ
		①	②	③	④	⑤	⑥	⑦	⑧	⑨	
Claude (ベース)	英日	4	1	7	1	1	0	2	1	53	2/19
	中日	34	3	0	2	2	0	0	2	83	3/20
	韓日	11	5	1	0	1	0	0	1	0	1/20
	日英	2	2	9	0	0	0	3	0	3	3/19
	合計	51	11	17	3	4	0	5	4	139	9/78
プロンプト ①	英日	3	0	7	0	1	0	3	1	26	2/19
	中日	36	3	0	1	4	1	0	1	63	4/20
	韓日	11	2	2	0	1	0	0	0	0	1/20
	日英	2	1	8	1	0	0	2	0	3	3/19
	合計	52	6	17	2	6	1	5	2	92	10/78

プロンプト①により入力文のジャンル(発明の名称、要約など)と読者(特許サーチャー)が明示されることで、より特許文の体裁に沿った翻訳が得られる効果が期待された。結果、英日において請求項の用言止めが顕著に改善され、誤訳カテゴリ別カウントの「⑨その他」のカウント数がベースモデルの53文から26文に半減した。一方、ベースモデルで同じくこの不備が多発した中日に関しては、83文→63文と若干の改善にとどまり、発生数は依然として多い。それ以外の項目には目立った変動はなく、言語、項目によっては数値の低下も見られる。つまり、本モデルについても、英日請求項の文体改善を除けば、いくつかの前後処理モデルと同様「Claude ベースモデルの翻訳精度が非常に良好であるためプロンプトの工夫で改善できる余地が少なく、結果、誤差レベルの小変動が生じたのみ」という状況といえる。

### 8.2.10.2 プロンプト②

プロンプト②では、入力文の技術分野の情報として、その文を採取した文献の筆頭IPCサブクラスの説明文を提示した。以下、Claude にこのプロンプトを用いた機械翻訳文の評価結果を示す(表 8.2.10-6～表 8.2.10-9)。各言語とも対象項目は「発明の名称」のみである。

表 8.2.10-6 [内容伝達レベルの評価] 言語別の全項目平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.97	4.74	4.98	4.99	4.92
プロンプト②	4.98	4.83	4.97	4.96	4.94

表 8.2.10-7 [流暢さの評価] 言語別の全項目平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.99	4.70	5.00	4.99	4.92
プロンプト②	4.99	4.77	5.00	4.99	4.94

表 8.2.10-8 [重要技術用語の評価] 言語別の A 判定率

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	96.9%	85.9%	93.8%	97.9%	93.6%
プロンプト②	97.9%	88.9%	95.9%	96.9%	94.9%

表 8.2.10-9 [誤訳のカテゴリ別カウント] 及び [連続文の訳ゆれ] の言語別カウント数

モデル	言語	誤訳のカテゴリ別カウント									連続文
		①	②	③	④	⑤	⑥	⑦	⑧	⑨	

											訳ゆれ <sup>21</sup>
Claude (ベース)	英日	0	0	2	0	0	0	0	0	1	—
	中日	12	0	0	1	1	0	0	0	2	—
	韓日	1	0	1	0	0	0	0	1	0	—
	日英	0	0	1	0	0	0	0	0	1	—
	合計	13	0	4	1	1	0	0	1	4	—
プロンプト ②	英日	0	0	2	0	0	0	0	0	0	—
	中日	10	0	0	0	0	0	0	0	2	—
	韓日	2	0	1	0	0	0	0	1	1	—
	日英	0	0	2	1	0	0	0	0	1	—
	合計	12	0	5	1	0	0	0	1	3	—

プロンプト②は、入力文の属する技術分野を明示することで、分野によって訳し分ける必要がある語句の翻訳精度が向上することが期待された。結果、中日と韓日の重要技術用語の評価でそれぞれ 3.0 ポイント、2.0 ポイントの向上が見られた。また、内容伝達レベルも中日で 0.09 ポイント向上した。

今回の人手評価では、ベースモデル (Claude) の時点で評価値がきわめて高いこともあり、前後処理の導入やプロンプトの工夫による評価値の変動幅が小さいケースが多い。特にプロンプト②で評価対象項目とした「発明の名称」は各言語とも 4 項目の中で最も評価値が高く、改善の余地が小さかった。こうした中、プロンプト②における上記の評価値向上は変動幅が比較的大きく、プロンプトによる改善効果が得られた証左といえる。

具体的な改善効果を示す中日の一例を示す。

[CNA113774509\_JPA22024522118\_tit]

原文	一种连续聚合-干湿法纺丝制备 <b>改性间位芳纶纤维</b> 的方法及装置
正解訳文	連続重合-乾湿式紡糸による <b>改質メタ系アラミド繊維</b> の製造方法及び装置
Claude ベース レベル: 3	<b>改性間位アラミド繊維</b> を製造するための連続重合-乾湿式紡糸方法及び装置
プロンプト② レベル: 5	<b>改質メタ系アラミド繊維</b> を製造するための連続重合-乾湿式紡糸方法及び装置

<sup>21</sup> 「発明の名称」は 1 文献あたり 1 文で「連続文」が得られないため「連続文での訳ゆれ」評価の対象外である。

本例では「**改性間位芳纶纤维**」が重要技術用語に選ばれている。この語の訳語が、ベースモデルの「**改性間位アラミド繊維**」(C判定)からプロンプト②モデルでは「**改質メタ系アラミド繊維**」(A判定)に改善し、結果、内容伝達レベルもレベル3からレベル5に向上した。入力文の技術分野を特定したことと、「**間位**」の訳語が「**間位**」から「**メタ系**」に改善したことの因果関係は解明が難しいが、両モデルの差はプロンプトの文言のみであるため、プロンプトの工夫による変化であることは確実である。

### 8.2.10.3 プロンプト③&プロンプト④

プロンプト③は、プロンプト①「翻訳対象の項目（発明の名称、要約、請求の範囲、明細書等）および、想定読者に「特許のサーチャー」を指定」とプロンプト②「技術分野の情報として、IPCの説明文（サブクラス）を指定」の双方を組み合わせて使用したモデルである。一方、プロンプト④は、プロンプト③に対して更に「文構造を分析し、主要要素（主語、動詞、目的語）を特定する」という CoT(Chain of Thought)を追加したモデルである。両モデルとも各言語の「特許請求の範囲」が対象項目であるため、本項でまとめて論じる(表 8.2.10-10～表 8.2.10-13)。

表 8.2.10-10 [内容伝達レベルの評価] 言語別の全項目平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.95	4.65	4.86	4.87	4.83
プロンプト③	4.95	4.71	4.88	4.95	4.87
プロンプト④	4.97	4.62	4.90	4.94	4.86

表 8.2.10-11 [流暢さの評価] 言語別の全項目平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.78	4.64	4.93	4.88	4.81
プロンプト③	4.68	4.70	4.93	4.95	4.82
プロンプト④	4.74	4.71	4.97	4.96	4.85

表 8.2.10-12 [重要技術用語の評価] 言語別の A 判定率

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	100.0%	78.0%	94.0%	100.0%	93.0%
プロンプト③	99.0%	85.0%	94.0%	100.0%	94.5%
プロンプト④	99.0%	79.0%	93.0%	100.0%	92.8%

表 8.2.10-13 [誤訳のカテゴリ別カウント] 及び [連続文の訳ゆれ] の言語別カウント数

モデル	言語	誤訳のカテゴリ別カウント									連続文 訳ゆれ
		①	②	③	④	⑤	⑥	⑦	⑧	⑨	
Claude (ベース)	英日	0	0	3	0	0	0	1	0	49	0/9
	中日	16	0	0	1	1	0	0	1	79	2/10
	韓日	8	3	0	0	1	0	0	0	0	1/10
	日英	1	0	6	0	0	0	2	0	0	2/9
	合計	25	3	9	1	2	0	3	1	98	4/38
プロンプト ③	英日	0	0	2	0	0	0	2	0	22	0/9
	中日	16	0	0	0	1	0	0	0	54	1/10
	韓日	9	1	0	0	1	0	0	0	1	2/10
	日英	1	0	2	1	0	0	1	0	0	2/9
	合計	26	1	4	1	2	0	3	0	77	5/38
プロンプト ④	英日	2	0	1	0	0	0	2	0	17	0/9
	中日	21	1	0	1	2	6	0	0	30	0/10
	韓日	9	2	0	1	0	0	0	0	1	1/10
	日英	0	0	1	3	0	1	2	0	0	1/9
	合計	32	3	2	5	2	7	4	0	48	2/38

プロンプト③に関しては、他に比べて低い評価値であった中日の重要技術用語の A 判定率が 6.0 ポイントと大幅に向上し、これにより内容伝達レベルも 0.06 ポイント良化した。プロンプト③はプロンプト①とプロンプト②の組み合わせであるが、この変動は前述したプロンプト②モデルの傾向と同じであり、主にプロンプト②（入力文の技術分野を明示）の効果と推測される。

一方、プロンプト④は、人手翻訳者が翻訳文を生成する際の思考の一つのあり方を模した 6 段階の Hierarchical Translation の手順を CoT としてプロンプトに指示したものである。これにより、より人手翻訳に近い翻訳文が得られることが期待された。

実際の評価結果も、ベースモデルに比べて内容伝達レベルが全言語合計で +0.03 ポイント、流暢さの評価が +0.04 ポイントと一定の評価値向上が見られた。CoT を導入していないプロンプト③に比べても各評価のスコアは概して高いが、中日の内容伝達レベルのみベースモデルに比べて -0.03 ポイントと低下しており、その結果、全言語平均ではプロンプト③に 0.01 ポイント及ばなかった。

プロンプト④モデルで、ベースモデル及びプロンプト③モデルに比べて内容伝達レベルが低評価となった事例を示す。

[CNA111093667\_JPA22024107309\_clm]

原文	根据权利要求 1~4 中任一项所述的固体制剂, 其中, 维生素 B <sub>1</sub> 类为 <b>呋喃硫胺</b> 、 <b>呋喃硫胺盐酸盐</b> 或 <b>硝酸硫胺</b> 。
正解訳文	ビタミン B <sub>1</sub> 類が、 <b>フルスルチアミン</b> 、 <b>フルスルチアミン塩酸塩</b> またはチアミン硝化物である、請求項 1~4 のいずれかに記載の固形製剤。
Claude ベース レベル: 5	請求項 1~4 のいずれか一項に記載の固形製剤において、ビタミン B <sub>1</sub> 類が <b>フルスルチアミン</b> 、 <b>フルスルチアミン塩酸塩</b> 又は <b>硝酸チアミン</b> である。
プロンプト② レベル: 5	請求項 1~4 のいずれか一項に記載の固形製剤であって、ビタミン B <sub>1</sub> 類が <b>フルスルチアミン</b> 、 <b>フルスルチアミン塩酸塩</b> 又は <b>硝酸チアミン</b> である、固形製剤。
プロンプト② レベル: 3	請求項 1~4 のいずれか一項に記載の固体制剤において、ビタミン B <sub>1</sub> 類化合物が <b>フルフリルチアミン</b> 、 <b>フルフリルチアミン塩酸塩</b> 又は <b>硝酸チアミン</b> である、固体制剤。

本例では、ベースモデルとプロンプト③モデルが重要技術用語「**呋喃硫胺**」を「**フルスルチアミン**」と正しく訳しているのに対し、プロンプト④モデルのみ「**フルフリルチアミン**」という出所不明の訳語に誤訳された。これにより重要技術用語評価が C 判定、内容伝達レベルも 3 となった。なぜ CoT の導入によってこのような誤訳が発生したかは不明である。

なお、プロンプト①モデルで見られた「英日の請求項の文体の改善」については、プロンプト③、④はいずれもプロンプト①が組み込まれているため、両モデルともプロンプト①と同様に大きく改善している。さらにプロンプト④は中日も 79 文→30 文と、「特許請求の範囲」を評価対象とした 5 モデル（前後処理①、⑤、プロンプト①、③、④）の中で最も大きく改善した。

#### 8.2.10.4 プロンプト⑧&プロンプト⑨

プロンプト⑧とプロンプト⑨も、評価対象項目が各言語の「明細書」で共通するため、本項でまとめて扱う(表 8.2.10-14~表 8.2.10-17)。

プロンプト⑧と⑨は、それぞれプロンプト③に対して「人手翻訳者の思考のあり方を模

した手順」を CoT として追加したものであり、この点において前項のプロンプト④と同系列である。プロンプト⑧では「文の中核構造を把握し、それにフィットする文体や訳語を選び、その理由を説明させて妥当性を担保させる」 Translation Explanation の手順を、プロンプト⑨では「入力文の構造を解析し、最適の文構造や統語構造に適宜言い換えて翻訳する」 Structural Transformation の手順を追加した。以下、両モデルの評価結果をベースモデルと対比して示す。

表 8.2.10-14 [内容伝達レベルの評価] 言語別の全項目平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.56	4.81	4.96	4.74	4.77
プロンプト⑧	4.64	4.76	4.91	4.95	4.82
プロンプト⑨	4.60	4.71	4.93	4.94	4.80

表 8.2.10-15 [流暢さの評価] 言語別の全項目平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.95	4.77	4.92	4.94	4.90
プロンプト⑧	4.97	4.73	4.96	4.89	4.89
プロンプト⑨	4.96	4.76	4.97	4.92	4.90

表 8.2.10-16 [重要技術用語の評価] 言語別の A 判定率

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	88.5%	81.1%	95.8%	91.7%	89.3%
プロンプト⑧	91.7%	85.3%	93.7%	92.7%	90.9%
プロンプト⑨	89.6%	80.0%	95.8%	93.8%	89.8%

表 8.2.10-17 [誤訳のカテゴリ別カウント] 及び [連続文の訳ゆれ] の言語別カウント数

モデル	言語	誤訳のカテゴリ別カウント									連続文 訳ゆれ
		①	②	③	④	⑤	⑥	⑦	⑧	⑨	
Claude (ベース)	英日	4	1	2	1	1	0	1	1	3	2/9
	中日	6	3	0	0	0	0	0	1	2	1/10
	韓日	2	2	0	0	0	0	0	0	0	0/10
	日英	1	2	2	0	0	0	1	0	2	1/9
	合計	13	8	4	1	1	0	2	2	7	4/38
プロンプト ⑧	英日	2	0	2	2	0	0	1	1	1	2/9
	中日	7	5	0	1	0	1	1	0	2	3/10

	韓日	5	0	1	0	0	0	0	0	1	0/10
	日英	2	1	3	0	0	0	2	0	2	2/9
	合計	16	6	6	3	0	1	4	1	6	7/38
プロンプト ⑨	英日	3	0	3	0	0	1	1	1	3	3/9
	中日	6	3	0	3	1	1	1	0	4	3/10
	韓日	2	0	0	1	0	0	1	0	1	0/10
	日英	2	1	2	1	0	0	1	0	2	1/9
	合計	13	4	5	5	1	2	4	1	10	7/38

プロンプト⑧、⑨とも人手翻訳者が翻訳文を生成する際の思考のあり方を模した手順をCoTとしてプロンプトに指示したモデルである。指示した手順はプロンプト⑧がTranlation Explanation（6段階）、⑨がStructural Translation（6段階）と異なるが、狙う効果はいずれも「より人手翻訳に近い翻訳文を得る」ことである。

両モデルの評価結果は、上表に示したとおり、大枠としてはどちらも内容伝達レベルや重要技術用語の訳質などで全言語合算の評価値がベースモデルより向上する傾向が見られた。ただしその一方で、両モデルとも（同系統のプロンプト④と同様に）中日の内容伝達レベルの平均値がベースモデルよりも若干低下している。さらにプロンプト⑧、⑨では韓日の平均レベルも低下し、幾つかの誤訳カテゴリのカウント値や連続文での訳ゆれの発生率も増加している。このように少量の向上と低下が混在するのは各種前後処理／プロンプトを通じて常に見られる傾向ではあるが、両モデルともプロンプト③や④に比べて混在の度合いが強く、その改善効果は不明瞭である。

#### 8.2.11 特許由来の学習データを追加学習させたモデル

LLM ベースモデルに特許由来の対訳コーパスを学習させることで、翻訳品質がさらに改善する可能性が考えられる。このため、本事業では、英日、中日、韓日、日英の「発明の名称」「要約」「特許請求の範囲」「明細書」からそれぞれ1,500文対ずつ、ならびに日英審査書類の「拒絶理由通知書」「意見書」「補正書」からも1,500文対ずつを、LLMモデルへの追加学習用対訳コーパスとして作成した（⇒第6章）。そして、Gemma ベースモデルに対し、これらのコーパスを追加学習させたモデル（f1500\_gemma）を作成し、その機械翻訳文の人手評価結果をベースモデルと比較することで、追加学習の効果を調査した。

評価は各言語とも特許文献の全4項目（「発明の名称」「要約」「特許請求の範囲」「明細書」）、及び日英審査書類の全3種（「拒絶理由通知書」「意見書」「補正書」）を対象とした。

以下、追加学習モデル f1500\_Gemma の各観点による人手評価結果を、ベースモデル Gemma と対比して示す(表 8.2.11-1～表 8.2.11-5)。

表 8.2.11-1 [内容伝達レベルの評価] 言語別の全項目平均スコア

モデル	英日	中日	韓日	日英	審査	全言語
Gemma (ベース)	4.53	4.44	4.74	4.62	4.45	4.56
f1500_gemma	4.52	4.48	4.74	4.62	4.37	4.55

表 8.2.11-2 [流暢さの評価] 言語別の全項目平均スコア

モデル	英日	中日	韓日	日英	審査	全言語
Gemma (ベース)	4.90	4.52	4.95	4.89	4.66	4.78
f1500_gemma	4.89	4.56	4.95	4.89	4.59	4.78

表 8.2.11-3 [重要技術用語の評価] 言語別の A 判定率

モデル	英日	中日	韓日	日英	審査	全言語
Gemma (ベース)	88.4%	76.3%	85.4%	91.5%	87.9%	85.9%
f1500_gemma	87.4%	73.8%	85.4%	91.0%	86.6%	84.8%

表 8.2.11-4 [誤訳のカテゴリ別カウント] 及び [連続文の訳ゆれ] の言語別カウント数

モデル	言語	誤訳のカテゴリ別カウント									連続文 訳ゆれ
		①	②	③	④	⑤	⑥	⑦	⑧	⑨	
Gemma (ベース)	英日	21	4	26	29	3	9	7	1	89	7/29
	中日	95	6	8	27	17	17	0	1	85	3/30
	韓日	60	4	9	14	4	1	2	1	6	7/30
	日英	20	4	26	28	1	26	3	0	7	3/29
	審査	34	33	24	46	0	7	9	1	15	4/30
	合計	230	51	93	144	25	60	21	4	202	24/148
f1500_gemma	英日	23	4	32	18	3	16	6	3	77	11/29
	中日	93	8	4	25	9	14	2	0	86	4/30
	韓日	54	7	15	9	1	1	3	1	13	9/30
	日英	18	4	38	17	0	37	6	1	9	5/29
	審査	47	32	24	34	0	7	8	1	26	6/30
	合計	235	55	113	103	13	75	25	6	211	35/148

表 8.2.11-5 出願人名／発明者の評価における A～B 率

モデル	法人名			個人名			合算 A～B 率
	評価 A	評価 B	AB 率	評価 A	評価 B	AB 率	
Gemma	5	19	48.0%	17	5	44.0%	46.0%
f1500_gemma	7	21	56.0%	15	8	46.0%	51.0%

ベースモデルとの比較では、中日の内容伝達レベルが 0.04 ポイント向上した一方で、英日の内容伝達レベルは-0.01 ポイント低下し、また韓日以外の三言語で重要技術用語の A 判定率がそれぞれ-1.0、-2.5、-0.5 ポイント低下、連続文の訳ゆれが 9 ペア増加するなど品質低下を示す項目も少なくない。審査書類も各観点でベースモデルより数値が低下しており、また誤訳カテゴリ別チェックも全体的にカウントが増えている。本来、特許由来の学習データを追加学習させることにデメリットがあるとは考えにくいものの、今回の評価結果からは、1,500 文の特許由来対訳コーパスの追加学習による翻訳改善効果は乏しいと結論される。

### 8.3 大規模言語モデルによる評価

各モデルによる翻訳 (7 章) で作成した人手評価用データの機械翻訳結果について、公開情報調査(翻訳精度評価手法)(5 章)で調査対象とした手法に使い評価した。手法の概要を表 8.3-1 に示す。評価手法①と評価手法②の評価観点は似ているが、手法①は参照訳 (= 正解訳) を使用せず評価しているのに対し、評価手法②では参照訳を使用した評価に違いがある。

評価結果を 8.3.1 から 8.3.6 に示す。評価結果は、手法①では誤訳カテゴリ毎に、誤訳が出現した延べ数を集計した。評価手法②では、評価観点 (正確性 (文)、正確性 (用語)、流暢さ) 毎に文数を集計した。評価手法③は、文毎の評価値 (0～100) の平均値を集計した。そしてこれら評価結果を元に大規模言語モデル 3 モデルの精度比較、前後処理の適用有無による精度比較、プロンプトの工夫有無の翻訳精度を比較した。また、評価の事例を 8.3.11 に記載した。

表 8.3-1 評価手法の概要

手法	評価内容	評価で使用するデータ	評価で使用するモデル
評価手法①	誤訳は①誤訳カテゴリ、②誤訳サブカテゴリ、③誤訳レベルの観点で分類する。誤訳カテゴリは、Accuracy(文の誤訳)、Fluency(流暢さ)、Style(文体)、Terminology(語の誤訳)、誤訳レベルは Critical, Major, Miner で構成される。	原文 機械翻訳文	Gemini
評価手法②	正確性 (文) (5 段階)、正確性 (用語) (4 段階)、流暢さ (5	原文	Gemini

法②	段階) で評価を行う。	機械翻訳文 参照訳	
評価手法③	機械翻訳文を0～100の値で評価する。(値が大きいほど正確な翻訳を意味する)	原文 機械翻訳文 参照訳	Gemini

### 8.3.1 英日

表 8.3.1-1 大規模言語モデル評価(評価手法①)

項目	発明の名称				要約				請求項				明細書			
評価観点	acc	flu	st	term	acc	flu	st	ter	acc	flu	styl	te	acc	flu	styl	ter
	ura	en	yl	inolo	ura	en	yl	min	ura	en	e	rm	ura	en	e	mi
	cy	cy	e	gy	cy	cy	e	olog	cy	cy		in	cy	cy		no
								y				ol				lo
												og				gy
												y				
<b>大規模言語 による翻訳</b>																
Claude	4	1	5	1	3		14	4	10	8	15	1	8	2	21	5
Gemini	1		2	2	3		9		3	2	4	2	6	1	5	1
Gemma	8	4	2	7	21	3	13	8	28	6	10	4	28	1	19	9
<b>前後処理を 適用した翻 訳</b>																
前後処理																
①																
Claude																
前後処理																
②																
Claude	3		3	2	4	2	19		5	7	13	2	3	3	16	1
前後処理																
⑤																
Claude																
前後処理																
Claude																

プロンプト を工夫した 翻訳																	
prompt1																	
Claude	3	2	2	1					6	13	17	2	6		17	4	
prompt2																	
Claude	2			4													
prompt3																	
Claude									12	13	16	2					
prompt4																	
Claude									6	12	15	1					
prompt8																	
Claude													7		18	4	
prompt9																	
Claude													6	2	19	6	
追加学習済 大規模言語 モデル																	
f1500_																	
gemma	11	3	4	4	24	4	15	13	17	8	14	10	22	1	14	10	
ニューラル 機械翻訳																	
nmt	6		5	8	18	1	25	19	13	7	19	2	15	6	22	8	

表 8.3.1-2 大規模言語モデル評価(評価手法②(正確性(文)))

項目	発明の名称					要約				請求の範囲				明細書								
	5	4	3	2	1	5	4	3	2	5	4	3	2	1	5	4	3	2	1			
大規模言語モデルによる翻訳																						
Claude	96	3	1			94	3	2	1	89	8	2	1	95	3	1	1					
Gemini	93	4	2	1		96	4			96	3	1		97	3							
								1			1				1							
Gemma	85	5	7	3		79	9	0	2	78	2	8	2	73	7	7	2	1				

プロンプトを工夫した翻訳																	
<b>prompt1</b>																	
Claude	94	5	1					96	3	1		92	5	3			
<b>prompt2</b>																	
Claude	95	3	2														
<b>prompt3</b>																	
Claude								92	3	5							
<b>prompt4</b>																	
Claude								96	1	3							
<b>prompt8</b>																	
Claude												94	2	4			
<b>prompt9</b>																	
Claude												95	3	2			
前後処理を適用した翻訳																	
<b>前後処理①</b>																	
Claude								91	5	2	2	89	8	3			
<b>前後処理②</b>																	
Claude	95	5			94	1	4	1	97	2	1	95	4	1			
<b>前後処理⑤</b>																	
Claude								91	4	5		97	2	1			
追加学習済大規模言語モデル																	
f1500_gem								1		1	1			1			
ma	84	6	3	6	1	78	8	2	2	78	0	0	2	81	8	0	1
ニューラル機械翻訳																	
nmt	90	6	1	1	2	84	5	8	3	90	4	5	1	89	6	4	1

表 8.3.1-3 大規模言語モデル評価(評価手法②(正確性(用語)))

項目	発明の名称			要約			請求の範囲			明細書			
	4	3	2	4	3	2	1	4	3	2	4	3	2
大規模言語モデルによる翻訳													

Claude	83	16	1	84	14	2	87	12	1	84	15	1	
Gemini	91	8	1	91	8	1	89	9	2	96	3	1	
Gemma	75	17	8	72	23	5	69	25	6	63	26	11	
<b>前後処理を適用した翻訳</b>													
<b>前後処理①</b>													
Claude							84	14	2	74	25	1	
<b>前後処理②</b>													
Claude	90	10		83	14	3	91	9		85	15		
<b>前後処理⑤</b>													
Claude							81	16	3	82	18		
<b>プロンプトを工夫した翻訳</b>													
<b>prompt1</b>													
Claude	85	15					90	8	2	83	15	2	
<b>prompt2</b>													
Claude	84	15	1										
<b>prompt3</b>													
Claude							89	9	2				
<b>prompt4</b>													
Claude							87	11	2				
<b>prompt8</b>													
Claude										81	17	2	
<b>prompt9</b>													
Claude										82	17	1	
<b>追加学習済大規模言語モデル</b>													
f1500_gem													
ma	78	15	7	59	30	11	67	24	9	64	23	13	
<b>ニューラル機械翻訳</b>													
nmt	78	16	6	74	18	7	1	77	22	1	72	23	5

表 8.3.1-4 大規模言語モデル評価(評価手法②(流暢さ))

項目	発明の名称	要約	請求の範囲	明細書
----	-------	----	-------	-----

	5	4	3	2	5	4	3	2	5	4	3	2	5	4	3
<b>大規模言語モデルによる翻訳</b>															
Claude	91	9			82	16	2		34	59	6	1	82	17	1
Gemini	98	2			88	12			75	22	3		87	13	
Gemma	87	7	6		72	20	5	3	34	55	10	1	74	21	5
<b>前後処理を適用した翻訳</b>															
<b>前後処理①</b>															
Claude									38	54	8		72	27	1
<b>前後処理②</b>															
Claude	96	4			76	22	1	1	61	36	3		76	24	
<b>前後処理⑤</b>															
Claude									50	45	5		82	17	1
<b>プロンプトを工夫した翻訳</b>															
<b>prompt1</b>															
Claude	91	8	1						28	62	9	1	73	26	1
<b>prompt2</b>															
Claude	94	6													
<b>prompt3</b>															
Claude									34	58	8				
<b>prompt4</b>															
Claude									24	67	9				
<b>prompt8</b>															
Claude													83	16	1
<b>prompt9</b>															
Claude													79	19	2
<b>追加学習済大規模言語モデル</b>															
f1500_gemma	87	7	4	2	67	25	8		40	49	11		74	24	2
<b>ニューラル機械翻訳</b>															
nmt	87	10	3		71	24	4	1	58	35	7		68	26	6

表 8.3.1-5 大規模言語モデル評価(評価手法③)

値は評価値(0~100)の平均

項目	発明の名称	要約	請求項	明細書
<b>大規模言語モデルによる翻訳</b>				

Claude	97.790	96.760	90.430	96.350
Gemini	98.200	98.080	97.120	96.790
Gemma	91.690	88.510	85.210	89.250
<b>前後処理を適用した翻訳</b>				
<b>前後処理①</b>				
Claude			90.830	96.390
<b>前後処理②</b>				
Claude	97.220	97.230	94.430	97.230
<b>前後処理⑤</b>				
Claude			92.150	97.570
<b>プロンプトを工夫した翻訳</b>				
<b>prompt1</b>				
Claude	96.660		89.980	96.260
<b>prompt2</b>				
Claude	97.710			
<b>prompt3</b>				
Claude			89.360	
<b>prompt4</b>				
Claude			90.560	
<b>prompt8</b>				
Claude				96.330
<b>prompt9</b>				
Claude				96.830
<b>追加学習済大規模言語モデル</b>				
f1500_gemma	93.020	86.970	83.390	90.620
<b>ニューラル機械翻訳</b>				
nmt	94.830	91.270	92.000	93.880

### 8.3.2 中日

表 8.3.2-1 大規模言語モデル評価(評価手法①)

項目	発明の名称				要約				請求項				明細書			
	acc	flu	st	term	acc	flu	st	term	acc	flu	st	term	acc	flu	st	term
	ura	en	yl	inolo	ura	en	yl	inolo	ura	en	yl	inolo	ura	en	yl	inolo
評価観点	cy	cy	e	gy	cy	cy	e	gy	cy	cy	e	gy	cy	cy	e	gy

**大規模言語  
による翻訳**

																1
Claude	3			8	5	2	7	9	5	4	8	12	5	2	0	3
Gemini	2		1	1	4	1	3	1	2	4	3		2			2
Gemm																
a	13	1	1	8	27	6	6	16	28	10	7	16	15	7	2	8

**前後処理を  
適用した翻  
訳**

前後処理

①

Claude									6	9	8	7	3	2	6	5
--------	--	--	--	--	--	--	--	--	---	---	---	---	---	---	---	---

前後処理

⑤

Claude																1
Claude									13	5	0	4	5	1	5	7

**プロンプト  
を工夫した  
翻訳**

prompt1

Claude																	1
Claude	7		1	7					6	5	1	6	2	3	9	9	

prompt2

Claude	4			4													
--------	---	--	--	---	--	--	--	--	--	--	--	--	--	--	--	--	--

prompt3

Claude									6	7	8	8					
--------	--	--	--	--	--	--	--	--	---	---	---	---	--	--	--	--	--

prompt4

Claude									12	8	4	6					
--------	--	--	--	--	--	--	--	--	----	---	---	---	--	--	--	--	--

prompt8

Claude														5	3	9	
--------	--	--	--	--	--	--	--	--	--	--	--	--	--	---	---	---	--

prompt9

Claude														6	2	4	4
--------	--	--	--	--	--	--	--	--	--	--	--	--	--	---	---	---	---

**追加学習済  
大規模言語  
モデル**

f1500_																			
gemma	17			8	26	8	5	17	22	12	8	14	25	6	6	4			
<b>ニューラル</b>																			
<b>機械翻訳</b>																			
								1				2							
nmt	7	1	4	17	14	3	14	17	37	6	17	7	7	8	7				

表 8.3.2-2 大規模言語モデル評価(評価手法②(正確性(文)))

項目	発明の名称					要約					請求項					明細書				
	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1
<b>大規模言語による翻訳</b>																				
Claude	90	6	4			93	6	1			97	1	2			98	2			
Gemini	98	1	1			99	1				97	3				99	1			
Gemma	80	10	7	3		76	8	13	2	1	72	17	6	5		79	11	9	1	
<b>プロンプトを工夫した翻訳</b>																				
prompt1																				
Claude	91	7	2								97	3				96	4			
prompt2																				
Claude	95	2	3																	
prompt3																				
Claude											96	4								
prompt4																				
Claude											94	4	2							
prompt8																				
Claude																99			1	
prompt9																				
Claude																96	2	2		
<b>前後処理を適用した翻訳</b>																				
前後処理①																				
Claude											90	8	1	1		97	3			
前後処理⑤																				
Claude											93	4	2	1		98	2			
<b>追加学習済大規模言語モデル</b>																				
f1500_gemma	83	9	6	2		76	13	10	1		77	15	7	1		83	10	4	2	1
<b>ニューラル機械翻訳</b>																				
nmt	92	5	1	1	1	84	8	6	2		86	9	2	2	1	96	4			

表 8.3.2-3 大規模言語モデル評価(評価手法②(正確性(用語)))

項目	発明の名称				要約				請求項				明細書			
	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1
<b>大規模言語による翻訳</b>																
Claude	82	9	9		81	16	3		71	25	4		79	17	4	
Gemini	91	6	3		89	10	1		86	12	2		93	7		
Gemma	72	18	10		58	22	19	1	50	32	18		61	20	18	1
<b>前後処理を適用した翻訳</b>																
前後処理①																
Claude									74	20	6		85	12	3	
前後処理⑤																
Claude									77	19	4		84	14	2	
<b>プロンプトを工夫した翻訳</b>																
prompt1																
Claude	83	13	3	1					68	32			78	20	2	
prompt2																
Claude	85	11	4													
prompt3																
Claude									73	25	2					
prompt4																
Claude									71	24	5					
prompt8																
Claude													84	16		
prompt9																
Claude													79	18	3	
<b>追加学習済大規模言語モデル</b>																
f1500_gemma	72	19	9		60	27	13		55	30	15		66	25	9	
<b>ニューラル機械翻訳</b>																
nmt	86	10	4		77	15	8		70	25	5		86	13	1	

表 8.3.2-4 大規模言語モデル評価(評価手法②(流暢さ))

項目名	発明の名称					要約					請求項					明細書				
	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1
<b>大規模言語による翻訳</b>																				
Claude	92	4	4			69	26	5			43	50	6	1		71	27	2		

Gemini	97	1	2	82	18	61	30	9	90	10					
Gemma	90	7	3	64	21	14	1	38	51	10	1	69	22	9	
<b>前後処理を適用した翻訳</b>															
前後処理①															
Claude						45	48	5	2	78	20	1	1		
前後処理⑤															
Claude						45	44	11		73	25	2			
<b>プロンプトを工夫した翻訳</b>															
prompt1															
Claude	92	4	3	1		36	49	14	1	74	22	4			
prompt2															
Claude	95	2	3												
prompt3															
Claude						44	42	14							
prompt4															
Claude						41	50	9							
prompt8															
Claude										74	24	2			
prompt9															
Claude										76	23	1			
<b>追加学習済大規模言語モデル</b>															
f1500_gemma	90	8	2	58	36	6	34	51	14	1	67	30	2	1	
<b>ニューラル機械翻訳</b>															
nmt	93	6	1	70	18	11	1	40	44	13	2	1	79	18	3

表 8.3.2-5 大規模言語モデル評価(評価手法③)

項目	発明の名称	要約	請求項	明細書	
<b>大規模言語による翻訳</b>					
Claude		96.14	93.61	90.65	94.52
Gemini		97.68	97.81	94.12	97.57
Gemma		91.96	85.15	80.78	88.91
<b>前後処理を適用した翻訳</b>					
前後処理①					
Claude			91.24	95.55	
前後処理⑤					

Claude			87.70	96.18
<b>プロンプトを工夫した翻訳</b>				
prompt1				
Claude	95.42		87.84	94.86
prompt2				
Claude	96.71			
prompt3				
Claude			90.19	
prompt4				
Claude			89.28	
prompt8				
Claude				95.17
prompt9				
Claude				95.25
<b>追加学習済大規模言語モデル</b>				
f1500_gemma	91.57	85.47	82.78	89.41
<b>ニューラル機械翻訳</b>				
nmt	97.15	89.87	82.08	96.07

### 8.3.3 韓日

表 8.3.3-1 大規模言語モデル評価(評価手法①)

項目	発明の名称				要約				請求項				明細書			
	acc	flu	st	term	acc	flu	st	term	acc	flu	st	term	acc	flu	st	term
評価観点	ura	en	yl	inolo	ura	en	yl	inolo	ura	en	yl	inolo	ura	en	yl	inolo
	cy	cy	e	gy	cy	cy	e	gy	cy	cy	e	gy	cy	cy	e	gy
<b>大規模言語</b>																
<b>による翻訳</b>																
Claude	2	1		5	5	1	5	6	4	3	6	7	4	2	6	2
Gemini		1	1	3	5		1	1	1	9	5	2	1	2	4	
Gemma																
a	11		1	5	19	1	6	10	18	13	6	10	22	3	5	8
<b>前後処理を適用した翻訳</b>																

前後処理																				
①																				
Claude																				
前後処理																				
⑤																				
Claude																				
<b>プロンプト を工夫した 翻訳</b>																				
prompt1																				
Claude	3		1		5				5	5	8		4	2			7		4	
prompt2																				
Claude	1		2		5															
prompt3																				
Claude									6	3	4		5							
prompt4																				
Claude									5	4	3		7							
prompt8																				
Claude														4	4	6			1	
prompt9																				
Claude														4	2	3			4	
<b>追加学習済 大規模言語 モデル</b>																				
f1500_																				
gemma	9			5	18	3	6	7	24	11	9	7	12	3	5				8	
<b>ニューラル 機械翻訳</b>																				
nmt	7			7	15	8	5	10	25	16	5	11	10	4	4				3	

表 8.3.3-2 大規模言語モデル評価(評価手法②(正確性(文)))

項目	発明の名称					要約					請求項					明細書				
	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1
<b>大規模言語による翻訳</b>																				
Claude	93	3	3	1		94	3		3		91	5	1	3		97	2		1	

Gemini	100					92	2	6			98	1	1		97	2	1		
Gemma	86	6	4	3	1	78	9	7	5	1	74	8	12	5	1	75	9	11	5
<b>プロンプトを工夫した翻訳</b>																			
prompt1																			
Claude	98			1	1						94		2	2	2	98	1		1
prompt2																			
Claude	96	2	1	1															
prompt3																			
Claude											90	4	4	2					
prompt4																			
Claude											92	3	2	1	2				
prompt8																			
Claude																92	3	3	2
prompt9																			
Claude																93	2	4	1
<b>前後処理を適用した翻訳</b>																			
前後処理①																			
Claude											94		1	4	1	94	3	3	
前後処理⑤																			
Claude											92	2	3	2	1	96	2	2	
<b>追加学習済大規模言語モデル</b>																			
f1500_gemma	85	7	5	2	1	84	3	10	1	2	75	10	11	4		83	6	7	4
<b>ニューラル機械翻訳</b>																			
nmt	93	4	1	2		83	5	8	2	2	79	8	7	6		93	2	4	1

表 8.3.3-3 大規模言語モデル評価(評価手法②(正確性(用語)))

項目	発明の名称				要約				請求項				明細書			
	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1
<b>大規模言語による翻訳</b>																
Claude	72	24	4		75	21	4		57	34	9		81	16	3	
Gemini	81	19			78	15	7		80	18	2		85	14	1	
Gemma	73	18	9		66	16	16	2	55	28	17		61	21	18	
<b>前後処理を適用した翻訳</b>																
前後処理①																
Claude									71	26	3		80	17	3	
前後処理⑤																

Claude										76	18	6		86	12	2
<b>プロンプトを工夫した翻訳</b>																
prompt1																
Claude										78	21	1		65	27	8
prompt2																
Claude										79	18	3				
prompt3																
Claude										71	23	6				
prompt4																
Claude										65	30	5				
prompt8																
Claude														85	9	6
prompt9																
Claude														86	9	5
<b>追加学習済大規模言語モデル</b>																
f1500_gemma										72	18	10		66	20	14
<b>ニューラル機械翻訳</b>																
nmt										81	16	3		78	8	14
										78	10	12		86	11	3

表 8.3.3-4 大規模言語モデル評価(評価手法②(流暢さ))

項目名	発明の名称					要約					請求項					明細書							
	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1			
<b>大規模言語による翻訳</b>																							
Claude	82	16	2			68	30	1	1		22	68	8	2		80	18	2					
Gemini	83	17				72	28				49	48	3			85	14	1					
Gemma	86	13	1			63	33	3	1		20	65	11	4		76	18	6					
<b>前後処理を適用した翻訳</b>																							
前後処理①																							
Claude																28	62	9	1		80	19	1
前後処理⑤																							
Claude																50	43	7			82	16	2
<b>プロンプトを工夫した翻訳</b>																							
prompt1																							

Claude	80	18	2					26	68	4	2	80	19	1		
prompt2																
Claude	82	17	1													
prompt3																
Claude								22	73	4	1					
prompt4																
Claude								40	53	7						
prompt8																
Claude												82	15	3		
prompt9																
Claude												85	14	1		
<b>追加学習済大規模言語モデル</b>																
f1500_gemma	91	7	1	1	67	26	7	24	63	13		82	16	2		
<b>ニューラル機械翻訳</b>																
nmt	85	13	2		77	14	7	1	1	65	21	9	5	91	7	2

表 8.3.3-5 大規模言語モデル評価(評価手法③)

項目	発明の名称	要約	請求項	明細書	
<b>大規模言語による翻訳</b>					
Claude		95.44	93.90	86.37	95.96
Gemini		97.65	95.40	92.44	97.69
Gemma		91.78	88.79	80.39	87.95
<b>前後処理を適用した翻訳</b>					
前後処理①					
Claude				86.21	96.99
前後処理⑤					
Claude				91.41	97.17
<b>プロンプトを工夫した翻訳</b>					
prompt1					
Claude		95.93		86.81	96.61
prompt2					
Claude		95.81			
prompt3					

Claude				87.54	
prompt4					
Claude				88.67	
prompt8					
Claude				94.96	
prompt9					
Claude				96.55	
<b>追加学習済大規模言語モデル</b>					
f1500_gemma		93.24	89.33	80.01	91.01
<b>ニューラル機械翻訳</b>					
nmt		95.66	91.31	86.63	96.25

### 8.3.4 日英(発明の名称、要約、請求項、明細書)

表 8.3.4-1 大規模言語モデル評価(評価手法①)

項目	発明の名称				要約				請求項				明細書					
	acc	flu	st	term	acc	flu	st	term	acc	flu	st	term	acc	flu	st	term		
	ura	en	yl	inolo	ura	en	yl	inolo	ura	en	yl	inolo	ura	en	yl	inolo		
評価観点	cy	cy	e	gy	cy	cy	e	gy	cy	cy	e	gy	cy	cy	e	gy		
<b>大規模言語</b>																		
<b>による翻訳</b>																		
																	1	
Claude	3	2	8	3	4	4	9	2	7	5	8	3	5	1	6	1	1	
Gemini	3		2	1	5	3	8	1	8	5	1	2	2	1	0			
Gemm							1										1	
a	5	3	1	2	19	4	5	2	24	3	8	6	21	3	4		3	
<b>前後処理を適用した翻訳</b>																		
前後処理																		
①																		
																	1	2
Claude									10	5	0	2	7	2	2		4	



none																
Claude	96	3	1	99	1	97	2	1	98	2						
Gemini	98	1	1	99	1	97	1	2	98	1	1					
Gemma	90	5	3	2	90	5	2	3	77	12	6	5	89	7	4	
<b>プロンプトを工夫した翻訳</b>																
prompt1																
Claude	94	3	3			99		1	97	1	2					
prompt2																
Claude	93	4	2	1												
prompt3																
Claude						96	2	2								
prompt4																
Claude						95	4	1								
prompt8																
Claude										95	4	1				
prompt9																
Claude										98	1	1				
<b>前後処理を適用した翻訳</b>																
前後処理①																
Claude						96	2	2		95	4	1				
前後処理⑤																
Claude						98	1	1		99		1				
<b>追加学習済大規模言語モデル</b>																
f1500_gemma	88	8	1	3	85	9	4	2	81	7	8	4	92	4	4	
<b>ニューラル機械翻訳</b>																
nmt	90	7	2	1	83	9	5	3	88	7	5		94	3	2	1

表 8.3.4-3 大規模言語モデル評価(評価手法②(正確性(用語)))

項目	発明の名称				要約				請求項				明細書			
	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1
<b>大規模言語による翻訳</b>																
Claude	93	5	2		93	7			94	4	2		94	6		
Gemini	95	4	1		96	4			92	6	2		96	3	1	
Gemma	87	8	5		84	13	3		80	11	8	1	87	8	5	
<b>前後処理を適用した翻訳</b>																
前後処理①																

Claude						91	7	2			91	7	2		
前後処理⑤															
Claude						91	9				91	8	1		
<b>プロンプトを工夫した翻訳</b>															
prompt1															
Claude			95	3	2					96	4		95	4	1
prompt2															
Claude			91	6	3										
prompt3															
Claude									92	6	2				
prompt4															
Claude									89	9	2				
prompt8															
Claude												96	3	1	
prompt9															
Claude												93	6	1	
<b>追加学習済大規模言語モデル</b>															
f1500_gemma			89	8	3	83	13	4	81	13	5	1	81	15	4
<b>ニューラル機械翻訳</b>															
nmt			86	10	4	83	9	8	80	14	6		86	11	3

表 8.3.4-4 大規模言語モデル評価(評価手法②(流暢さ))

項目名	発明の名称					要約					請求項					明細書				
	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1
<b>大規模言語による翻訳</b>																				
Claude	95	5				78	22				85	12	3			84	16			
Gemini	97	3				91	9				91	9				92	8			
Gemma	97	3				78	17	5			65	27	8			81	15	4		
<b>前後処理を適用した翻訳</b>																				
前後処理①																				
Claude											87	10	3			84	16			
前後処理⑤																				
Claude											82	17	1			86	13	1		



Claude	96.85			
prompt3				
Claude			97.45	
prompt4				
Claude			97.62	
prompt8				
Claude				97.44
prompt9				
Claude				98.24
<b>追加学習済大規模言語モデル</b>				
f1500_gemma	94.95	93.87	89.68	96.70
<b>ニューラル機械翻訳</b>				
nmt	96.10	92.52	96.05	96.40

### 8.3.5 日英(拒絶理由、意見書、補正書、出願人名/発明者名)

表 8.3.5-1 大規模言語モデル評価(評価手法①)

項目	拒絶理由				意見書				補正書				出願人名/発明者名				
評価観点	ac	flu	s	ter	ac	flu	s	ter	ac	flu	s	ter	ac	flu	non	s	ter
	cur	en	t	min	cur	en	t	min	cur	en	t	min	cur	en	-	t	min
	ac	cy	yl	olog	ac	cy	yl	olog	ac	cy	yl	olog	ac	cy	tran	yl	olog
	y	e	y	y	e	y	y	y	e	y	y	y	slati	e	y	on <sup>22</sup>	
<b>大規模言語による翻訳</b>																	
Claude			2				1				1						
e	11	3	2	2	6	3	7	7	9	9	8	1	16	3		1	
Gemini			1				1										
ni	5	3	4	6	3		5	2	8	7	4	2	5	3		1	
Gemma			2				2				1						
ma	28	3	1	8	41	4	6	7	34	8	2	9	50	2			

<sup>22</sup> 出願人名発明者名「七澤 茂介」を”XXXXXX”と翻訳したデータに対して、評価手法①が誤訳サブカテゴリである“non-translation”を、誤訳カテゴリとして1件出力された。

**前後処理を適用した翻訳**

前後処理②

Claude	1					2					1				
e	7	2	4	4	10	1	0	9	7	6	0	3	12	1	

**追加学習済大規模言語モデル**

f1500_gemma	1					1					1				
e	32	5	4	7	28	3	7	7	35	16	9	3	46	1	

**ニューラル機械翻訳**

nmt	4					4					1						
e	28	8	1	9	27	11	3	10	41	26	6	7	64	9	1	1	1

表 8.3.5-2 大規模言語モデル評価(評価手法②(正確性(文)))

項目	拒絶理由					意見書					補正書					出願人名/発明者名				
	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1
<b>大規模言語による翻訳</b>																				
Claude	98	2				97	3				100					79	4	11	6	
Gemini	99	1				100					98	1	1			83	5	8	3	1
Gemma	86	8	5	1		74	17	9			85	6	7	2		55	2	19	14	10
<b>前後処理を適用した翻訳</b>																				
前後処理②																				
Claude	97	2			1	95	3	2			95	4	1			80	5	9	4	2
<b>追加学習済大規模言語モデル</b>																				
none																				
f1500_gemma	83	6	9	2		90	3	5	2		86	7	6	1		51	3	23	15	8
<b>ニューラル機械翻訳</b>																				
nmt	85	7	7	1		87	6	5	2		83	10	7			43	9	7	16	25

表 8.3.5-3 大規模言語モデル評価(評価手法②(正確性(用語)))

項目	拒絶理由				意見書				補正書				出願人名/発明者名			
	4	3	2	1	4	3	2	1	4	3	2	1	4	3	2	1
<b>大規模言語による翻訳</b>																
Claude	89	10	1		91	8	1		95	5			79	3	18	
Gemini	97	1	2		96	4			94	5	1		82	4	14	
Gemma	85	11	4		79	17	4		72	22	6		53	4	43	
<b>前後処理を適用した翻訳</b>																
前後処理②																
Claude	94	4	2		93	6	1		95	5			82		18	
<b>追加学習済大規模言語モデル</b>																
f1500_gemma	82	11	7		74	20	6		80	14	6		52	3	45	
<b>ニューラル機械翻訳</b>																
nmt	86	10	3	1	82	13	5		74	22	4		43	6	49	2

表 8.3.5-4 大規模言語モデル評価(評価手法②(流暢さ))

項目名	拒絶理由					意見書					補正書					出願人名/発明者名				
	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1
<b>大規模言語による翻訳</b>																				
Claude	68	31	1			78	21	1			74	22	4			99		1		
Gemini	83	17				86	14				86	14				98	2			
Gemma	66	31	3			73	27				62	30	8			95	1	3	1	
<b>前後処理を適用した翻訳</b>																				
前後処理②																				
Claude	76	23	1			77	21	2			77	21	2			99	1			
<b>追加学習済大規模言語モデル</b>																				
<b>デル</b>																				
f1500_gemma	66	27	7			79	19	2			56	30	13	1		99		1		
<b>ニューラル機械翻訳</b>																				
nmt	64	27	9			50	38	12			61	33	3	3		82	1	10	4	3

表 8.3.5-5 大規模言語モデル評価(評価手法③)

項目	拒絶理由	意見書	補正書	出願人名/発明者名
<b>大規模言語による翻訳</b>				
Claude	97.17	97.60	96.46	87.51

Gemini	98.69	98.80	98.14	87.68
Gemma	92.93	93.04	91.56	72.76
<b>前後処理を適用した翻訳</b>				
前後処理②				
Claude	97.31	97.64	97.36	88.14
<b>追加学習済大規模言語モデル</b>				
f1500_gemma	91.33	94.74	88.92	70.25
<b>ニューラル機械翻訳</b>				
nmt	90.18	89.97	91.00	60.69

### 8.3.6 特許文献4項目別の集計結果

8.3.1～3 に示した言語別の各評価結果を項目別に集計した結果を示す(表 8.3.6-1、8.3.6-2、8.3.6-3)。前後処理②は英日のみを対象としており、集計母数が異なることから集計対象外とした。

表 8.3.6-1 大規模言語モデル評価  
(評価手法①(発明の名称、要約、請求項、明細書))

評価観点	発明の名称				要約				請求項				明細書			
	st		st		st		st		st		st		st			
	accu	flue	yl	termin	accu	flue	yl	termin	accu	flue	yl	termin	accu	flue	yl	termin
	racy	ncy	e	ology	racy	ncy	e	ology	racy	ncy	e	ology	racy	ncy	e	ology
<b>大規模言語による翻訳</b>																
Claude	12	4	13	17	17	7	35	21	26	20	37	23	22	7	53	11
Gemini	6	1	6	7	17	4	21	3	14	20	13	6	11	4	19	3
Gemma	37	8	5	22	86	14	40	36	98	32	31	36	86	14	40	28
<b>前後処理を適用した翻訳</b>																
前後処理①																
Claude									33	30	44	19	27	7	55	19
前後処理⑤																
Claude									33	22	35	14	24	7	42	19
<b>プロンプトを工夫した翻訳</b>																

<b>prompt1</b>																	
Claude	18	3	11	15					21	27	44	14	14	4	54	22	
<b>prompt2</b>																	
Claude	12		8	15													
<b>prompt3</b>																	
Claude									28	25	35	18					
<b>prompt4</b>																	
Claude									32	27	30	17					
<b>prompt8</b>																	
Claude														23	8	53	8
<b>prompt9</b>																	
Claude														23	7	42	15
<b>追加学習済大規模言語モデル</b>																	
<b>f1500_gem</b>																	
ma	47	7	14	19	91	19	36	40	88	42	42	38	80	15	42	28	
<b>ニューラル機械翻訳</b>																	
<b>nmt</b>																	
	38	5	17	25	71	37	59	56	71	62	62	35	50	27	57	25	

表 8.3.6-2 大規模言語モデル評価  
(評価手法②(正確性(文))(発明の名称、要約、請求項、明細書))

	発明の名称					要約					請求項					明細書				
	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1
<b>大規模言語による翻訳</b>																				
Claude	375	15	9	1		380	13	3	4		374	16	6	4		388	9	2	1	
Gemini	389	6	3	2		386	8	6			388	8	4			391	7	2		
Gemma	341	26	21	11	1	323	31	32	12	2	301	49	32	17	1	316	44	31	8	1
<b>プロンプトを工夫した翻訳</b>																				
<b>prompt1</b>																				
Claude	377	15	6	1	1						386	3	6	3	2	383	11	5	1	
<b>prompt2</b>																				
Claude	379	11	8	2																
<b>prompt3</b>																				
Claude											374	13	11	2						

<b>prompt4</b>																											
Claude																			377	12	8	1	2				
<b>prompt8</b>																											
Claude																			380	9	9	2					
<b>prompt9</b>																											
Claude																			382	8	9	1					
<b>前後処理を適用した翻訳</b>																											
<b>前後処理①</b>																											
Claude																			371	15	6	7	1	375	18	7	
<b>前後処理⑤</b>																											
Claude																			374	11	11	3	1	390	6	3	1
<b>追加学習済大規模言語モデル</b>																											
f1500_gemma	340	30	15	13	2	323	33	36	6	2	311	42	36	11	339	28	25	7	1								
<b>ニューラル機械翻訳</b>																											
nmt	365	22	5	5	3	334	27	27	10	2	343	28	19	8	2	372	15	10	3								

表 8.3.6-3 大規模言語モデル評価  
(評価手法②(正確性(用語))(発明の名称、要約、請求項、明細書))

	発明の名称				要約				請求項						
	4	3	2	1	4	3	2	1	4	3	2	1			
<b>大規模言語による翻訳</b>															
<b>none</b>															
Claude	330	54	16		333	58	9		309	75	16				
Gemini	358	37	5		354	37	9		347	45	8				
Gemma	307	61	32		280	74	43	3	254	96	49	1			
<b>前後処理を適用した翻訳</b>															
<b>前後処理①</b>															
Claude													320	67	13
<b>前後処理⑤</b>															
Claude													325	62	13
<b>プロンプトを工夫した翻訳</b>															
<b>prompt1</b>															
Claude	341	52	6	1									319	71	10
<b>prompt2</b>															
Claude	339	50	11												





Claude				96.72
<b>追加学習済大規模言語モデル</b>				
f1500_gemma	93.20	88.91	83.97	91.94
<b>ニューラル機械翻訳</b>				
nmt	95.94	91.24	89.19	95.65

### 8.3.7 大規模言語モデル 3 モデル及び NMT モデルの比較

大規模言語モデルによる評価を実施した 3 種の大規模言語モデル（Claude、Gemini、Gemma）と、比較のためのニューラル翻訳モデル（nmt）の全 4 モデルについて、本項で評価観点ごとに評価結果を比較する。これら 4 モデルは、まず特許文献として各言語（英日、中日、韓日、日英）とも「発明の名称」「要約」「特許請求の範囲」「明細書」の 4 項目から 100 文ずつ、計 400 文を評価した。本項に示す各言語の評価結果は、これら 4 項目・全 400 文の集計結果となる。また、これとは別途に日英審査書類として「拒絶理由通知書」「意見書」「補正書」からそれぞれ 100 文、計 300 文についても評価しており、集計結果を表中「審査」欄に示した。

#### 8.3.7.1 文の誤訳指摘数、正確性（文）の評価結果の比較

下表に、各モデルの評価手法①の文の誤訳指摘数、評価手法②の正確性（文）の評価結果について、評価手法①は、正確性（文）に関する誤訳の指摘数を、評価手法②は 5 段階評価の平均値を集計した。また評価手法③は 0～100 の範囲の評価値の平均値を集計した（表 8.3.7-1～表 8.3.7-3）。

表 8.3.7-1 評価手法①文の誤訳指摘数

モデル	英日	中日	韓日	日英	審査	全言語
Claude	25	18	15	19	26	103
Gemini	13	10	7	18	16	64
Gemma	85	83	70	69	103	410
nmt	52	48	57	73	96	326

評価手法①の文に関する誤訳は、Gemma、nmt、Claude、Gemini の順に多く指摘された。

言語別に見ると審査書類は、Gemini以外のモデルで多くの指摘が見られた。

表 8.3.7-2 評価手法②正確性（文）平均スコア

モデル	英日	中日	韓日	日英	審査	全言語
Claude	4.91	4.93	4.89	4.97	4.98	4.94
Gemini	4.94	4.98	4.95	4.97	4.98	4.96
Gemma	4.66	4.62	4.59	4.78	4.73	4.67
nmt	4.79	4.83	4.75	4.83	4.77	4.79

評価手法②の正確性（文）は、全てのモデルスコアの平均が4.5以上の評価となった。Gemini、Claudeが平均4.9以上でほぼ同様の精度となり、nmt、Gemmaが続く結果となった。

表 8.3.7-3 評価手法③平均スコア

モデル	英日	中日	韓日	日英	審査	全言語
Claude	93.73	95.33	97.83	92.92	97.08	95.38
Gemini	96.80	97.55	98.42	95.80	98.54	97.42
Gemma	86.70	88.67	93.66	87.23	92.51	89.75
nmt	91.29	93.00	95.27	92.46	90.38	92.48

評価手法③のスコアは、Gemmaの英日、中日、日英以外で90以上となり、全体的に高いスコアとなった。

### 8.3.7.2 流暢さの評価結果の比較

各モデルの評価手法①の流暢さ、評価手法②の流暢さの評価結果について、評価手法①は、流暢さに関する誤訳の指摘数を、評価手法②は5段階評価の平均値を集計した(表 8.3.7-4～表 8.3.7-5)。

表 8.3.7-4 評価手法①流暢さ 誤訳指摘数

モデル	英日	中日	韓日	日英	審査	全言語
Claude	11	8	7	12	15	53

Gemini	3	5	12	9	10	39
Gemma	14	24	17	13	15	83
nmt	14	58	28	31	45	176

流暢さに関する誤訳は、nmt、Gemma、Claude、Gemini の順に多く指摘された。Claude は審査書類、Gemini は韓日、Gemma と nmt は中日で誤訳指摘が多く見られた。

表 8.3.7-5 評価手法②流暢さ 平均スコア

モデル	英日	中日	韓日	日英	審査	全言語
Claude	4.70	4.64	4.85	4.58	4.71	4.70
Gemini	4.86	4.80	4.93	4.71	4.85	4.83
Gemma	4.58	4.55	4.76	4.54	4.63	4.61
nmt	4.66	4.61	4.76	4.71	4.48	4.64

評価手法②の流暢さは、全てのモデルで平均スコアが 4.5 を超え全体的に高い結果となった。

### 8.3.7.3 語の誤訳指摘数、正確性（用語）の評価結果の比較

下表に、各モデルの評価手法①の語に関する誤訳、評価手法②の正確性（用語）の評価結果について、評価手法①は、語に関する誤訳の指摘数を、評価手法②は 4 段階評価の平均値を集計した(表 8.3.7-6～表 8.3.7-7)。

表 8.3.7-6 評価手法①用語に関する誤訳 誤訳指摘数

モデル	英日	中日	韓日	日英	審査	全言語
Claude	11	32	20	9	10	82
Gemini	5	4	6	4	10	29
Gemma	28	48	33	13	24	146
nmt	37	42	31	31	26	167

用語に関する誤訳の指摘は、Gemma、nmt、Claude、Gemini の順に多く指摘された。言語別に見ると中日の Gemma が最も多い誤訳指摘数となった。

表 8.3.7-7 方式②正確性（用語） 平均スコア

モデル	英日	中日	韓日	日英	審査	全言語
Claude	3.83	3.73	3.93	3.66	3.91	3.81
Gemini	3.91	3.88	3.94	3.79	3.95	3.89
Gemma	3.62	3.43	3.79	3.48	3.74	3.61
nmt	3.70	3.75	3.79	3.73	3.76	3.75

評価手法②の正確性（用語）は、Gemini、Claude がスコア 3.8 台とほぼ同等の結果となり、以下、nmt、Gemma の順となった。

#### 8.3.7.4 出願人・発明者の評価結果の比較

出願人、発明者について、評価手法①の文の誤訳指摘数、評価手法②の正確性（文）の平均スコアを集計した<sup>23</sup>(表 8.3.7-8～表 8.3.7-9)。

表 8.3.7-8 評価手法①文の誤訳指摘回数

モデル	出願人、発明者
Claude	16
Gemini	5
Gemma	50
nmt	64

表 8.3.7-9 評価手法②正確性 平均スコア

モデル	出願人、発明者
Claude	4.56

<sup>23</sup> 大規模言語モデルによる評価は、評価基準から、文章を対象にした評価となっている。一方出願人や発明者は固有名詞であるため、評価手法と評価対象が対応していない。そのため、下記表に示す評価結果は参考程度の扱いとするのが適当と考える。

Gemini	4.66
Gemma	3.78
nmt	3.29

評価手法①の正確性の誤訳指摘は、Gemini が指摘数 5 と最も少なく、Claude が 16、Gemma が 50、nmt が 64 となった。評価手法②の正確性も Gemini、Claude、Gemma、nmt の順となった。

### 8.3.8 前後処理を導入したモデル

大規模言語モデルによる機械翻訳を行う際に、特許翻訳に適すると考えられる特定の前後処理を導入した場合の翻訳品質を評価し、ベースモデルからの変化の度合いを確認した。本項でその結果をまとめる。

本評価では、大規模言語モデル Claude をベースモデルとし、このモデルに 3 種の前後処理<sup>24</sup>を導入した場合に機械翻訳文の翻訳品質水準がどのように変化するかを調べた。

#### 8.3.8.1 前後処理①

前後処理①モデルについて、英日、中日、韓日、日英それぞれ「特許請求の範囲」と「明細書」の 2 項目、200 文を評価した。下の各表に、前後処理①モデルの各評価の集計結果を、ベースモデル Claude との対比で示す(表 8.3.8.1-1～表 8.3.8.1-7)。比較条件を合わせるため、ベースモデルの評価結果も「特許請求の範囲」と「明細書」の 2 項目のみの合算値とした。

表 8.3.8.1-1 評価手法①正確性 誤訳指摘数

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	18	10	8	12	48
前後処理①	18	9	16	17	60

表 8.3.8.1-2 評価手法②正確性 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.89	4.97	4.90	4.97	4.93
前後処理①	4.86	4.92	4.87	4.94	4.90

<sup>24</sup> 前後処理については、3 章公開情報調査（前後処理）を参照のこと。

表 8.3.8.1-3 評価手法③ 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	93.39	92.59	91.17	97.56	93.67
前後処理①	93.61	93.40	91.60	97.27	93.97

正確性の観点では、評価手法①の前後処理①は韓日、日英で誤訳指摘数がベースモデルより増えた。評価手法②の前後処理①では全ての言語で若干のスコアの低下が見られた。評価手法③の前後処理①では日英を除きスコアの上昇がみられた。

表 8.3.8.1-4 評価手法① 流暢さ 誤訳指摘数

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	10	6	5	6	27
前後処理①	7	11	12	7	37

表 8.3.8.1-5 評価手法②流暢さ 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.54	4.52	4.44	4.83	4.58
前後処理①	4.51	4.56	4.48	4.84	4.60

流暢さの観点では、評価手法①の前後処理①の誤訳指摘数は増加した一方、評価手法②の前後処理①のスコアは若干の向上が見られた。

表 8.3.8.1-6 評価手法① 用語 誤訳指摘数

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	6	15	9	4	34
前後処理①	12	12	8	6	38

表 8.3.8.1-7 評価手法②正確性 (用語) 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	3.85	3.71	3.63	3.93	3.78
前後処理①	3.78	3.75	3.73	3.89	3.79

流暢さの観点では、評価手法①の前後処理①の誤訳指摘数は英日、日英で誤訳指摘数が増

加した。評価手法②の前後処理①は日英、中日で微減し、中日が微増、韓日で0.1ポイントの増加が見られた。

### 8.3.8.2 前後処理②

前後処理②モデルについて、英日の全4項目（「発明の名称」「要約」「特許請求の範囲」「明細書」）、及び審査書類（日英）の全3種（「拒絶理由通知」「意見書」「補正書」）を対象に各評価を実施した。下表に、その結果をベースモデル Claude と対比して示す(表 8.3.8.2-1～表 8.3.8.2-7)。

表 8.3.8.2-1 評価手法①正確性 誤訳指摘数

モデル	英日	日英(審査)	全言語
Claude (ベース)	25	26	51
前後処理②	15	24	39

表 8.3.8.2-2 評価手法②正確性 平均スコア

モデル	英日	日英(審査)	全言語
Claude (ベース)	4.91	4.98	4.94
前後処理②	4.93	4.94	4.93

表 8.3.8.2-3 評価手法③ 平均スコア

モデル	英日	日英(審査)	全言語
Claude (ベース)	94.86	97.08	95.97
前後処理②	96.29	97.44	96.87

正確性の観点では、評価手法①の前後処理②の誤訳指摘数は減少、評価手法②の前後処理①では英日のスコアに若干の低下が見られた。また評価手法③の前後処理②は英日で1.4ポイント、日英で0.3ポイントのスコアの増加が見られた。

表 8.3.8.2-4 評価手法① 流暢さ 誤訳指摘数

モデル	英日	日英(審査)	全言語
Claude (ベース)	11	15	26

前後処理②	12	9	21
-------	----	---	----

表 8.3.8.2-5 評価手法②流暢さ 平均スコア

モデル	英日	日英(審査)	全言語
Claude (ベース)	4.70	4.71	4.70
前後処理②	4.76	4.75	4.75

流暢さの観点では、評価手法①の前後処理②の誤訳指摘数は日英で指摘数の減少が見られた。評価手法②の前後処理②は英日、日英でスコアの微増が見られた。

表 8.3.8.2-6 評価手法①用語 誤訳指摘数

モデル	英日	日英(審査)	全言語
Claude (ベース)	11	10	21
前後処理②	5	16	21

表 8.3.8.2-7 評価手法②正確性 (用語) 平均スコア

モデル	英日	日英(審査)	全言語
Claude (ベース)	3.83	3.91	3.87
前後処理②	3.87	3.93	3.90

用語の観点では、評価手法①の前後処理②の誤訳指摘数は英日で減少、日英で増加が見られた。評価手法②の前後処理②では各言語ともに若干のスコアの上昇が見られた。

### 8.3.8.3 前後処理⑤

前後処理⑤モデルは、前後処理①と同様、英日、中日、韓日、日英それぞれで「特許請求の範囲」と「明細書」の2項目、200文を評価対象とした。下の各表に、前後処理⑤モデルの各評価の集計結果を、ベースモデル Claude の評価結果との対比で示す(表 8.3.8.3-1～表 8.3.8.3-7)。

表 8.3.8.3-1 評価手法①正確性 誤訳指摘数

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	18	10	8	12	48

前後処理⑤	18	18	10	11	57
-------	----	----	----	----	----

表 8.3.8.3-2 評価手法②正確性 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.89	4.97	4.90	4.97	4.93
前後処理⑤	4.91	4.94	4.88	4.98	4.92

表 8.3.8.3-3 評価手法③ 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	93.39	92.59	91.17	97.56	93.67
前後処理⑤	94.86	91.94	94.29	98.26	94.84

正確性の観点では、評価手法①の前後処理⑤の誤訳指摘数は中日、韓日で増加が見られた。評価手法②の前後処理⑤は英日、日英でスコアの微増、中日、韓日ではスコアの微減が見られた。また、評価手法③の前後処理⑤は中日でスコアの低下が見られた。

表 8.3.8.3-4 評価手法① 流暢さ 誤訳指摘数

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	10	6	5	6	27
前後処理⑤	6	6	13	4	29

表 8.3.8.3-5 評価手法②流暢さ 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.54	4.52	4.44	4.83	4.58
前後処理⑤	4.63	4.53	4.62	4.83	4.65

流暢さの観点では、評価手法①の前後処理⑤の誤訳指摘数は韓日で増加、英日、日英で減少が見られた。評価手法②の前後処理⑤は日英以外の言語でスコアの上昇が見られた。

表 8.3.8.3-6 評価手法① 用語 誤訳指摘数

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	6	15	9	4	34
前後処理⑤	8	11	10	4	33

表 8.3.8.3-7 評価手法②正確性（用語） 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude（ベース）	3.85	3.71	3.63	3.93	3.78
前後処理⑤	3.80	3.78	3.77	3.91	3.81

用語の観点では、評価手法①の前後処理⑤の誤訳指摘数は英日、韓日で若干の増加が、中日で若干の減少見られた。評価手法②の前後処理⑤は中日、韓日でスコアの増加、英日、日英でスコアの微減が見られた。

### 8.3.9 プロンプトに工夫を施したモデル

大規模言語モデルによる機械翻訳を行う際に、特許翻訳に適すると考えられるプロンプトに工夫を施した6種のモデルについて、ベースモデルからの変化の度合いを確認した。

#### 8.3.9.1 プロンプト①

プロンプト①では、翻訳対象の項目（発明の名称、要約、請求の範囲、明細書等）および、想定読者に「特許のサーチャー」を指定した。以下、このプロンプトを用いて Claude で生成させた機械翻訳文の評価結果を各表に示す(表 8.3.9.1-1～表 8.3.9.1-7)。なお、評価対象項目は各言語とも「発明の名称」「特許請求の範囲」「明細書」の3項目である。

表 8.3.9.1-1 評価手法①正確性 誤訳指摘数

モデル	英日	中日	韓日	日英	全言語
Claude（ベース）	22	13	10	15	60
プロンプト①	15	15	10	13	53

表 8.3.9.1-2 評価手法②正確性 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude（ベース）	4.91	4.93	4.89	4.96	4.92
プロンプト①	4.91	4.94	4.90	4.95	4.92

表 8.3.9.1-3 評価手法③ 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	94.86	93.77	92.59	97.60	94.70
プロンプト①	94.30	92.71	93.12	97.90	94.51

正確性の観点では、評価手法①のプロンプト①の誤訳指摘数は英日、日英で指摘の減少が、中日で指摘の増加が見られた。評価手法②のプロンプト①は中日、韓日でスコアの微増、日英でスコアの微減が見られた。また、評価手法③のプロンプト①は英日、中日のスコアの減少が見られた。

表 8.3.9.1-4 評価手法① 流暢さ 誤訳指摘数

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	11	6	6	8	31
プロンプト①	15	8	5	6	34

表 8.3.9.1-5 評価手法②流暢さ 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.66	4.64	4.56	4.87	4.68
プロンプト①	4.60	4.59	4.58	4.87	4.66

流暢さの観点では、評価手法①のプロンプト①の誤訳指摘数は英日、中日で増加が、韓日、日英で減少が見られた。評価手法②のプロンプト①は英日、中日でスコアの低下が、韓日でスコアの増加が見られた。

表 8.3.9.1-6 評価手法① 用語 誤訳指摘数

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	7	23	14	7	51
プロンプト①	7	22	13	9	51

表 8.3.9.1-7 評価手法②正確性（用語） 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude（ベース）	3.84	3.72	3.65	3.92	3.78
プロンプト①	3.85	3.74	3.71	3.94	3.81

用語の観点では、評価手法①のプロンプト①の誤訳指摘数は日英で増加が見られた。評価手法②のプロンプト①は全ての言語でスコアの微増が見られた。

### 8.3.9.2 プロンプト②

プロンプト②では、入力文の技術分野の情報として、その文を採取した文献の筆頭 IPC サブクラスの説明文を示した。以下、Claude にこのプロンプトを用いた機械翻訳文の評価結果を各表に示す(表 8.3.9.2-1～表 8.3.9.2-7)。なお、各言語とも対象項目は「発明の名称」のみである。

表 8.3.9.2-1 評価手法①正確性 誤訳指摘数

モデル	英日	中日	韓日	日英	全言語
Claude（ベース）	4	3	2	3	12
プロンプト②	2	4	1	5	12

表 8.3.9.2-2 評価手法②正確性 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude（ベース）	4.95	4.86	4.88	4.95	4.91
プロンプト②	4.93	4.92	4.93	4.89	4.92

表 8.3.9.2-3 評価手法③ 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude（ベース）	97.79	96.14	95.44	97.69	96.77
プロンプト②	97.71	96.71	95.81	96.85	96.77

正確性の観点では、評価手法①のプロンプト②の誤訳指摘数は中日、日英で若干指摘の増加が見られた。評価手法②のプロンプト②は中日、韓日でスコアの微増が見られた。また、評価手法③のプロンプト②は中日、韓日でスコアの微増が見られた。

表 8.3.9.2-4 評価手法① 流暢さ 誤訳指摘数

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	1	0	1	2	4
プロンプト②	0	0	0	0	0

表 8.3.9.2-5 評価手法②流暢さ 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.91	4.88	4.80	4.95	4.89
プロンプト②	4.94	4.92	4.81	4.97	4.91

流暢さの観点では、評価手法①のプロンプト②は中日以外の言語で指摘の減少が見られた。評価手法②のプロンプト②は全ての言語でスコアの微増が見られた。

表 8.3.9.2-6 評価手法① 用語 誤訳指摘数

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	1	8	5	3	17
プロンプト②	4	4	5	2	15

表 8.3.9.2-7 評価手法②正確性 (用語) 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	3.82	3.73	3.68	3.91	3.79
プロンプト②	3.83	3.81	3.76	3.88	3.82

用語の観点では、評価手法①のプロンプト②の英日で指摘の増加が見られた。評価手法②のプロンプト②は英日、中日、韓日でスコアの微増が、日英で微減が見られた。

### 8.3.9.3 プロンプト③&プロンプト④

プロンプト③は、プロンプト①「翻訳対象の項目（発明の名称、要約、請求の範囲、明細書等）および、想定読者に「特許のサーチャー」を指定」とプロンプト②「技術分野の情報として、IPCの説明文（サブクラス）を指定」の双方を組み合わせて使用したモデルである。一方、プロンプト④は、プロンプト③に対して更に「文構造を分析し、主要要素（主語、動詞、目的語）を特定する」という CoT(Chain of Thought)を追加したモデルである。両モデルとも各言語の「特許請求の範囲」が対象項目であるため、本項でまとめて論じる(表 8.3.9.3-

1～表 8.3.9.3-7)。

表 8.3.9.3-1 評価手法①正確性 誤訳指摘数

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	10	5	4	7	26
プロンプト③	12	6	6	4	28
プロンプト④	6	12	5	9	32

表 8.3.9.3-2 評価手法②正確性 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.85	4.95	4.84	4.96	4.90
プロンプト③	4.87	4.96	4.82	4.94	4.90
プロンプト④	4.93	4.92	4.82	4.94	4.90

表 8.3.9.3-3 評価手法③ 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	90.43	90.65	86.37	97.32	91.19
プロンプト③	89.36	90.19	87.54	97.45	91.14
プロンプト④	90.56	89.28	88.67	97.62	91.53

正確性の観点では、評価手法①のプロンプト③の誤訳指摘数は日英以外で若干指摘の増加が見られた。評価手法②のプロンプト③は英日、中日でスコアの微増が、韓日、日英で微減が見られた。また、評価手法③のプロンプト③は韓日、日英でスコアの増加が見られた。

プロンプト④では評価手法①が英日以外で指摘数の増加が見られた。評価手法②は英日以外でスコアの微減が見られた。また評価手法③では、中日以外でスコアの上昇が見られた。

表 8.3.9.3-4 評価手法① 流暢さ 誤訳指摘数

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	8	4	3	5	20
プロンプト③	13	7	3	2	25
プロンプト④	12	8	4	3	27

表 8.3.9.3-5 評価手法②流暢さ 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.26	4.35	4.10	4.82	4.38
プロンプト③	4.26	4.30	4.16	4.81	4.38
プロンプト④	4.15	4.32	4.33	4.89	4.42

流暢さの観点では、評価手法①のプロンプト③の誤訳指摘数は英日、中日で指摘の増加が見られた。評価手法②のプロンプト③は韓日でスコアの微増が見られた。

プロンプト④では評価手法①が日英以外で指摘数の増加が見られた。評価手法②は韓日、日英でスコアの微増が、英日、中日でスコアの微減が見られた。

表 8.3.9.3-6 評価手法①用語 誤訳指摘数

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	1	12	7	3	23
プロンプト③	2	8	5	3	18
プロンプト④	1	6	7	3	17

表 8.3.9.3-7 評価手法②正確性 (用語) 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	3.86	3.67	3.48	3.92	3.73
プロンプト③	3.87	3.71	3.65	3.90	3.78
プロンプト④	3.85	3.66	3.60	3.87	3.75

用語の観点では、評価手法①のプロンプト③の誤訳指摘数は中日、韓日で指摘の減少が見られた。評価手法②のプロンプト③は日英以外でスコアの微増が見られた。

プロンプト④では評価手法①が中日で指摘数の減少が見られ、韓日、日英で同等の指摘数だった。評価手法②は全ての言語でスコアの微減が見られた。

#### 8.3.9.4 プロンプト⑧&プロンプト⑨

プロンプト⑧とプロンプト⑨も、評価対象項目が各言語の「明細書」で共通するため、本項でまとめて結果を示す(表 8.3.9.4-1～表 8.3.9.4-7)。どちらもプロンプト③に CoT を追加したモデルである。

表 8.3.9.4-1 評価手法①正確性 誤訳指摘数

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	8	5	4	5	22
プロンプト⑧	7	5	4	7	23
プロンプト⑨	6	6	4	7	23

表 8.3.9.4-2 評価手法②正確性 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.92	4.98	4.96	4.98	4.96
プロンプト⑧	4.90	4.98	4.85	4.94	4.92
プロンプト⑨	4.93	4.94	4.87	4.97	4.93

表 8.3.9.4-3 評価手法③ 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	96.35	94.52	95.96	97.79	96.16
プロンプト⑧	96.33	95.17	94.96	97.44	95.98
プロンプト⑨	96.83	95.25	96.55	98.24	96.72

正確性の観点では、評価手法①のプロンプト⑧の誤訳指摘数は日英で若干指摘の増加が見られた。評価手法②のプロンプト⑧は中日以外でスコアの微減が見られた。また、評価手法③のプロンプト⑧は中日以外でスコアの微減が見られた。

プロンプト⑨では評価手法①が中日、日英で若干指摘数の増加が見られた。評価手法②では中日、韓日、日英でスコアの微減が見られた。また評価手法③では、全ての言語でスコアの上昇が見られた。

表 8.3.9.4-4 評価手法① 流暢さ 誤訳指摘数

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	2	2	2	1	7
プロンプト⑧	0	3	4	1	8
プロンプト⑨	2	2	2	1	7

表 8.3.9.4-5 評価手法②流暢さ 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	4.81	4.69	4.78	4.84	4.78
プロンプト⑧	4.82	4.72	4.79	4.82	4.79
プロンプト⑨	4.77	4.75	4.84	4.80	4.79

流暢さの観点では、評価手法①のプロンプト⑧の誤訳指摘数は中日、韓日で若干指摘の増加が見られた。評価手法②のプロンプト⑧は日英以外でスコアの微増が見られた。

プロンプト⑨では評価手法①が指摘数に変化はなかった。評価手法②では英日、日英でスコアの微減が見られた。

表 8.3.8.4-6 評価手法①用語 誤訳指摘数

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	5	3	2	1	11
プロンプト⑧	4	0	1	3	8
プロンプト⑨	6	4	4	1	15

表 8.3.9.4-7 評価手法②正確性 (用語) 平均スコア

モデル	英日	中日	韓日	日英	全言語
Claude (ベース)	3.83	3.75	3.78	3.94	3.83
プロンプト⑧	3.79	3.84	3.79	3.95	3.84
プロンプト⑨	3.81	3.76	3.81	3.92	3.83

用語の観点では、評価手法①のプロンプト⑧の誤訳指摘数は日英で若干指摘の増加が見られた。評価手法②のプロンプト⑧は日英以外でスコアの微増が見られた。

プロンプト⑨では評価手法①が日英以外で若干の増加が見られた。評価手法②では英日、日英でスコアの微減が見られた。

### 8.3.10 特許由来の学習データを追加学習させたモデル

Gemma に特許由来の対訳コーパスを学習させたモデルについて、学習前のモデルと比較した。学習させたモデルは「発明の名称」「要約」「特許請求の範囲」「明細書」「拒絶理由通知書」「意見書」「補正書」それぞれ 1,500 文対ずつ学習させたモデル(f1500\_gemma)を評価

した(表 8.3.10.1-1～表 8.3.10.1-7)。

表 8.3.10.1-1 評価手法①正確性 誤訳指摘数

モデル	英日	中日	韓日	日英	審査	全言語
Gemma (ベース)	85	83	70	69	103	410
f1500_gemma	74	90	63	79	95	401

表 8.3.10.1-2 評価手法②正確性 平均スコア

モデル	英日	中日	韓日	日英	審査	全言語
Gemma (ベース)	4.66	4.62	4.59	4.78	4.73	4.67
f1500_gemma	4.65	4.69	4.66	4.78	4.76	4.71

表 8.3.10.1-3 評価手法③ 平均スコア

モデル	英日	中日	韓日	日英	審査	全言語
Gemma (ベース)	88.67	86.70	87.23	93.66	92.51	89.75
f1500_gemma	88.50	87.31	88.40	93.80	91.66	89.93

正確性の観点では、評価手法①の f1500\_gemma の誤訳指摘数は中日、日英で指摘の増加が見られた。評価手法②の f1500\_gemma は英日以外でスコアの微増が見られた。また評価手法③は英日、日英（審査書類）でスコアの微減が見られた。

表 8.3.10.1-4 評価手法① 流暢さ 誤訳指摘数

モデル	英日	中日	韓日	日英	審査	全言語
Gemma (ベース)	14	24	17	13	15	83
f1500_gemma	16	26	17	24	24	107

表 8.3.10.1-5 評価手法②流暢さ 平均スコア

モデル	英日	中日	韓日	日英	審査	全言語
Gemma (ベース)	4.58	4.55	4.54	4.76	4.63	4.61
f1500_gemma	4.60	4.55	4.59	4.75	4.59	4.61

流暢さの観点では、評価手法①の f1500\_gemma の誤訳指摘数は韓日以外で指摘の増加が見られた。評価手法②の f1500\_gemma は英日、中日でスコアの微増が、日英、日英（審査書類）でスコアの微減が見られた。

表 8.3.10.1-6 評価手法① 用語 誤訳指摘数

モデル	英日	中日	韓日	日英	審査	全言語
Gemma (ベース)	28	48	33	13	24	146
f1500_gemma	37	43	27	18	17	142

表 8.3.10.1-7 評価手法②正確性 (用語) 平均スコア

モデル	英日	中日	韓日	日英	審査	全言語
Gemma (ベース)	3.62	3.43	3.48	3.79	3.74	3.61
f1500_gemma	3.57	3.52	3.52	3.79	3.72	3.62

用語の観点では、評価手法①の f1500\_gemma の誤訳指摘数は英日、日英で指摘の増加が見られた。評価手法②の f1500\_gemma は英日、日英（審査書類）でスコアの微減が見られた。

### 8.3.11 大規模言語モデルによる評価の事例

本節では、大規模言語モデルによる評価について英日の結果を使用し、各評価手法でどのような指摘がされているかの事例を示す。

・評価手法① エラーカテゴリ「accuracy」の事例

[Claude 請求項#69 前後処理なし プロンプト工夫なし]

原文	A robot arm as claimed in claim 4, wherein the controller is configured to, following applying the braking current claimed in claim 4, in response to continuing to detect a fault in the robot arm, apply a braking current between a second <b>pair of phases</b> of the motor only, from the first phase of the second pair to the second phase of the second pair.
参照訳	前記コントローラが、請求項4に記載の前記制動電流を印加することに続いて、前記ロボットアームの故障を検出し続けることに応答して、前記モータの第2の <b>対の相</b> の間のみ、前記第2の対の前記第1の相から前記第2の対の前記第2の相まで、制動電流を印加するように構成されている、請求項4に記載のロボットアーム。
機械翻訳	請求項 4 に記載のロボットアームにおいて、前記コントローラは、請求項 4 に記載の制動電流を印加した後、ロボットアームの故障を継続して検出することに応答して、 <b>第 2 の相対</b> の第 1 相から <b>第 2 の相対</b> の第 2 相へ、モータの <b>第 2 の相対</b> 間のみ制動電流を印加するように構成される。
評価①結果	エラーカテゴリ: accuracy サブカテゴリ: mistranslation エラーレベル: critical
評価①理由	<b>The technical term "a pair of phases" is incorrectly translated as "相対" (relative).</b> This is a critical error as it makes the technical description of the motor control nonsensical. The correct translation should be something like "相の対" or "相ペア". <b>This mistranslation is</b>

	<b>repeated three times (“a second pair of phases” -&gt; “第2の相対”, “the second pair” -&gt; “第2の相対”), rendering the core technical detail of the claim incomprehensible.</b>
備考	評価②結果 正確性(文):4/5 正確性(用語):2/4 流暢さ:3/5 評価③結果 スコア:50

この事例では、「pair of phases」（「対の相」又は「相の対」）を「相対」として翻訳していたものに対して、「accuracy」の問題があるとされた。日本語では、「の」を削除しても意味が通じることもあるが、「相対」といえば「relative」の意味であることも考慮された上で、致命的な誤訳であると判断されていることが、理由の出力から伺える。

・評価手法① エラーカテゴリ「fluency」の事例

[Gemma 請求項#92 前後処理なし プロンプト工夫なし]

原文	The method according to claim 2, further comprising exposing the peptide ligand displayed on the genetic display system to a second probe after step (c), wherein <b>the second probe</b> binds to the genetic display system, and comprises or <b>is linkable to</b> a second signalling group.
参照訳	前記遺伝子ディスプレイシステム上に提示される前記ペプチドリガンドをステップ(c)の後で第2のプローブに曝露するステップをさらに含み、 <b>前記第2のプローブが前記遺伝子ディスプレイシステムに結合し、第2のシグナル発生基を含むか、これに連結可能である、</b> 請求項2に記載の方法。
機械翻訳	請求項2に記載の方法であって、ステップ(c)の後、遺伝子ディスプレイシステムに表示されているペプチドリガンドを、第二のプローブに晒すことを特徴とする方法であって、 <b>前記第二のプローブは、遺伝子ディスプレイシステムに結合し、第二のシグナルグループを含むか、または第二のシグナルグループと結合可能な方法。</b>
評価①結果	エラーカテゴリ: fluency サブカテゴリ: grammar エラーレベル: critical
評価①理由	The end of the sentence is ungrammatical. <b>The phrase “...結合可能な方法” (a method that is linkable...) incorrectly modifies “方法” (method).</b> The original text states that the *probe* “comprises or is linkable to a second signalling group.” The translation incorrectly attributes this property to the method itself, creating a nonsensical and ungrammatical sentence structure. In the context of a patent claim, this is a critical error as it misrepresents a key feature.
備考	評価②結果 正確性(文):3/5 正確性(用語):3/4 流暢さ:2/5 評価③結果 スコア:45

この事例では、第2のプローブが第2のシグナルグループに結合可能である、という意味の英語に対し、文末で「結合可能な方法」と翻訳されていることに対して、「fluency」の問題があるとされている。「結合可能な方法」との記載では、方法自体が結合可能（「a method that is linkable」）と解されてしまうため、致命的な問題と判断されていることが、理由の出力から伺える。

・評価手法① エラーカテゴリ「style」の事例

[nmt 要約#65 前後処理なし プロンプト工夫なし]

原文	A method <b>to operate</b> a UE <b>for handling</b> security policy <b>for user plane protection</b> of communications in a communications system is provided.
参照訳	通信システムにおける通信の <b>ユーザプレーン保護のための</b> セキュリティポリシーを <b>ハンドリングするための</b> UEを <b>動作させるための</b> 方法が提供される。

機械翻訳	通信システムにおける通信のユーザープレーン保護のためのセキュリティポリシーを処理するために UE を動作させるための方法が提供される。
評価①結果	エラーカテゴリ: style サブカテゴリ: awkward エラーレベル: major
評価①理由	The translation is a literal, word-for-word rendering of the English source. The chain of prepositional phrases in English ('for handling... for... of... in...') is translated into <b>a long and convoluted chain of modifiers in Japanese ('...のための...ために...ための...')</b> . This results in a sentence that is grammatically correct but stylistically very unnatural and difficult to parse, disrupting the reading flow.
備考	評価②結果 正確性(文): 5/5 正確性(用語): 4/4 流暢さ: 5/5 評価③結果 スコア: 100

この事例では、「のための」、「ために」、「ための」との翻訳について、「style」の問題があるとされた。しかしながら、参照訳でも同様の訳をされていることからわかるように、特許の文書では、時折このような記載は存在し、やや厳しい判断であると考えられる。なお、参照訳を与える手法である、評価手法②、評価手法③では問題があるとはされなかった。

・評価手法① エラーカテゴリ「terminology」の事例

[Gemini 請求項#24 前後処理なし プロンプト工夫なし]

原文	The method of claim 9, further including identifying a nominal pocket geometry and <b>analog</b> pocket tolerance value based on the at least one 3D printer parameter.
参照訳	前記少なくとも1つの3Dプリンタパラメータに基づいて公称ポケットジオメトリ及び <b>類似物</b> ポケット公差値を識別することを更に含む、請求項9に記載の方法。
機械翻訳	少なくとも1つの3Dプリンタパラメータに基づいて、公称ポケット形状および <b>アナログ</b> ポケット公差値を特定することをさらに含む、請求項9に記載の方法。
評価①結果	エラーカテゴリ: terminology サブカテゴリ: inappropriate for context エラーレベル: major
評価①理由	The word 'analog' is translated as 'アナログ'. While this is a direct transliteration, the term 'analog pocket tolerance' is highly unusual in an engineering/3D printing context. It's very likely that 'analog' is used here in the sense of 'analogue' (i.e., corresponding, similar). The direct transliteration 'アナログ' is jarring and technically questionable, as an 'analog tolerance' is not a standard engineering term. <b>A better translation would be '類似の' (similar/analogous).</b> This choice of term could lead to misinterpretation by a technical expert, thus it is a major error.
備考	評価②結果 正確性(文): 5/5 正確性(用語): 4/4 流暢さ: 5/5 評価③結果 スコア: 96

この事例では、「analog」（この文脈では「類似物」）を「アナログ」と翻訳されていたものに対して、「terminology」の問題があるとされていた。なお、評価手法②では、参照訳を与えているにも関わらず指摘がされず、評価手法③では、参考訳の方がよい（「The reference's choice is a better example of professional translation」）旨の出力はあったが、大きな問題とはされなかった。

・評価手法② 「正確性（文）」の事例

[Gemma 明細書#82 前後処理なし プロンプト工夫なし]

原文	LIGANDS OF THE FSH HORMONE RECEPTOR <b>IN THE DIAGNOSIS AND TREATMENT OF TUMORS</b>
----	---

参照訳	<u>腫瘍の診断及び治療におけるFSHホルモン受容体のリガンド</u>
機械翻訳	FSH ホルモン受容体のリガンド、 <u>および腫瘍の診断と治療</u>
評価②結果	正確性(文):2/5 正確性(用語):4/4 流暢さ:3/5
評価②正確性(文)理由	2/5 stars: The translation correctly identifies the key components but fails to capture the crucial relationship between them. The source describes "Ligands... IN the diagnosis and treatment...", indicating the ligands' application. <u>The translation changes this to "Ligands... AND the diagnosis and treatment..." ("...リガンド、および...診断と治療"), presenting them as two separate, parallel items.</u> This fundamentally alters the meaning of the title, as the core concept of the ligands' purpose is lost.
備考	評価①結果 エラーカテゴリ:accuracy サブカテゴリ:mistranslation エラーレベル:major 評価③結果 スコア:35

この事例では、「LIGANDS OF THE FSH HORMONE RECEPTOR」に係るべき「IN THE DIAGNOSIS AND TREATMENT OF TUMORS」を、「および腫瘍の診断と治療」と、「FSH ホルモン受容体のリガンド」に係らない形で翻訳したものに対して、「正確性(文)」の問題があるとされた。用語の訳には誤りがない事例であるが、「正確性(文)」として誤訳の指摘がされている。

・評価手法②「正確性(用語)」の事例

[Gemini 請求項#89 前後処理なし プロンプト工夫なし]

原文	The method as recited in claim 1, wherein the <b>hypodermic tubing</b> (20) is provided with an internal volume fixed to the surface of the substrate (10) forming into a gas distribution system which supplies externally pressurized gas.
参照訳	前記埋設チューブ(20)は、内容積を有し、前記基板(10)の表面に固定されて、外部から加圧ガスを供給するガス分配システムを形成する、ことを特徴とする、請求項1に記載の方法。
機械翻訳	請求項1に記載の方法であって、前記皮下注射用チューブ(20)は、内部容積を備え、基板(10)の表面に固定されて、外部から加圧されたガスを供給するガス分配システムを形成する、方法。
評価②結果	正確性(文):1/5 正確性(用語):2/4 流暢さ:4/5
評価②正確性(用語)理由	2 stars: <u>The term "hypodermic tubing" is translated literally as 「皮下注射用チューブ」(tubing for subcutaneous injection).</u> Given the context of a gas distribution system on a substrate and the human reference's use of 「埋設チューブ」(embedded tubing), the literal medical translation is a mistranslation of a key technical term.
備考	評価①結果 エラーカテゴリ:no-error サブカテゴリ:no-error エラーレベル:no-error 評価③結果 スコア:75

この事例では、「hypodermic tubing」を「皮下注射用チューブ」と訳していることに対して、「正確性(用語)」の誤りであると指摘がされている。一般に、「hypodermic」は「皮下の」という意味であるものの、理由の出力をみると、「ガス分配システム」の文脈では「埋設チューブ」が正しい訳であるとの判断がされていることが伺える。なお、参照訳を与えない評価手法①では、エラーが発見されておらず、参照訳の提示が有効に働いた可能性がある。

・評価手法②「流暢さ」の事例

[Claude 請求項#74 前後処理なし プロンプト工夫なし]

原文	<b>The method of claim 26, wherein</b> the doxorubicin is administered before, after, or simultaneously with the protein.
参照訳	ドキソルビシンが、前記タンパク質の前、前記タンパク質の後、または前記タンパク質と同時に投与される。 <b>請求項 26 記載の方法。</b>
機械翻訳	<b>請求項 26 に記載の方法であって、</b> ドキソルビシンが、前記タンパク質の前に、後に、または同時に投与される、 <b>方法。</b>
評価②結果	正確性(文):5/5 正確性(用語):4/4 流暢さ:3/5
評価②流暢さ理由	3 stars: <b>The sentence structure 「～であって、～、方法」 is redundant because it repeats 「方法」(method) at the end.</b> This is unnatural in Japanese patent claims. While the meaning is generally understandable, this structural flaw makes the sentence awkward and less clear than the human reference.
備考	評価①結果 エラーカテゴリ:fluency サブカテゴリ:grammar エラーレベル:major 評価③結果 スコア:82

この事例では、「請求項 26 に記載の方法であって、・・・方法。」という翻訳について、流暢さの問題があるとされている。しかしながら、当該記載は、参照訳の記載とは違うものの、特許の文書ではよくみられる記載であり、やや厳しい判断であると考えられる。

・評価手法③の事例

[Claude 要約#55 前後処理なし プロンプト工夫なし]

原文	There is provided a camera installation structure for inside of <b>operation room</b> that can obtain a camera image with reduced shake to enable remote operation and hardly obstructs rear visual field in a case where an operator boards a working machine.
参照訳	遠隔操作可能とするため揺れの少ないカメラ画像を得られ、オペレータが作業機械に搭乗した場合の後方視界の妨げになりにくい、 <b>操作室内</b> のカメラの設置構造を提供する。
機械翻訳	作業者が作業機械に搭乗した場合において、遠隔操作を可能にするために揺れの少ないカメラ画像を得ることができ、かつ後方視野をほとんど妨げない <b>手術室内</b> 用カメラ設置構造が提供される。
評価③結果	スコア:45
評価③理由	This translation has a critical error in terminology that completely changes the context of the invention, despite being grammatically correct and accurate in other parts. <b>**Positive Points:**</b> * The overall sentence structure is clear and follows patent language conventions. * Most of the technical concepts are translated accurately: * "camera image with reduced shake" -> 「揺れの少ないカメラ画像」(Good) * "to enable remote operation" -> 「遠隔操作を可能にするために」(Perfect) * "hardly obstructs rear visual field" -> 「後方視野をほとんど妨げない」(Good, very literal) * "an operator boards a working machine" -> 「作業者が作業機械に搭乗した場合」(Good) <b>**Negative Points:**</b> <b>**Critical Error:**</b> <b>"operation room" is translated as 「手術室」(shujutsu-shitsu), which means a surgical operating room in a hospital.</b> The context of a "working machine" (作業機械) clearly indicates that "operation room" refers to an operator's cab or control room, which should be 「操作室」(sousa-shitsu) as used in the human reference. * This single error completely misrepresents the field of the invention, creating a nonsensical scenario of an operator boarding a "working machine" inside a "surgery room". In a patent document, where the field of application is crucial, this is a fatal flaw. <b>**Scoring:**</b> While most of the functional descriptions are correct, the fundamental misunderstanding of the invention's context is a major failure. The translation is grammatically sound but semantically nonsensical due to the key term error. This severely compromises the meaning of the entire sentence. Therefore, despite the correct translation of other phrases, the score must be significantly penalized. [[[45]]]

備考	評価①結果 エラーカテゴリ:no-error サブカテゴリ:no-error エラーレベル:no-error 評価②結果 正確性(文):2/5 正確性(用語):2/4 流暢さ:4/5
----	--

この事例では、「operation room」を「手術室」と翻訳したことに対して、大きな減点がされている。「作業機械」という文脈によれば「作業室」との訳が適切と判断されていることが、理由の出力から伺える。なお、参照訳を与えない評価手法①では、エラーが発見されておらず、参照訳の提示が有効に働いた可能性がある。

[Gemini 明細書#44 前後処理なし プロンプト工夫なし]

原文	Meanwhile, <b>when</b> the wafer W <b>is immersed</b> in the processing bath 31 (step S101, Yes), for example, when the wafer W is etched in the processing bath 31, the <b>controller</b> performs a mixture liquid generation processing of generating the mixture liquid M <b>by the mixing device 10</b> (step S108).
参照訳	一方、処理槽31でウェハWが浸漬されている場合(ステップS101, Yes)、たとえば、処理槽31でウェハWがエッチング処理されている場合、 <b>制御部</b> は、 <b>混合装置10</b> で混合液Mを生成する混合液生成処理を実施する(ステップS108)。
機械翻訳	一方、ウェハWが処理槽31に <b>浸漬されると</b> (ステップS101でYes)、例えば、処理槽31でウェハWがエッチングされると、 <b>コントローラ</b> は、 <b>混合装置10</b> によって混合液Mを生成する混合液生成処理を実行する(ステップS108)。
評価③結果	スコア:70
評価③理由	This translation has a few issues that prevent it from being a high-quality patent translation, although the core meaning is mostly understandable. 1. <b>**Tense/Aspect Mismatch**</b> : The most significant error is the use of 「～されると」 (e.g., 「浸漬されると」, 「エッチングされると」). This phrasing implies a trigger or a point in time when an action is completed. The English source “when the wafer W is immersed” and “when the wafer W is etched” describes a continuous state or condition. The human reference correctly captures this with <b>「～されている場合」</b> (“ <b>in the case that it is immersed/being etched</b> ”), which accurately reflects the <b>ongoing state</b> during which the controller performs its action. The machine translation changes the logic from “while the wafer is in this state...” to “upon the wafer entering this state...”. 2. <b>**Terminology Choice**</b> : The translation uses <b>「コントローラ」</b> for <b>“controller”</b> . While a direct transliteration, <b>the standard and preferred term in Japanese patents is 「制御部」</b> , as used in the human reference. 3. <b>**Particle Choice**</b> : <b>The use of 「によって」 for “by the mixing device 10” is grammatically correct, but 「で」 as used in the reference is more concise and natural in this context.</b> Due to the inaccurate representation of the process timing/condition and the use of non-standard terminology, the translation is significantly flawed for a professional patent document. [[[70]]]
備考	評価①結果 エラーカテゴリ:no-error サブカテゴリ:no-error エラーレベル:no-error 評価②結果 正確性(文):5/5 正確性(用語):4/4 流暢さ:5/5

この事例では、「when ... is immersed」を「浸漬されると」と翻訳したこと（参照訳では「浸漬されている場合」）、「controller」を「コントローラ」と翻訳したこと（参照訳では「制御部」）、「by the mixing device 10」を「混合装置10で」と翻訳したこと（参照訳では「混合装置10によって」）に対して、比較的大きな減点がされており、やや厳しい判断がされていると考えられる。なお、参照訳を与えない評価手法①では、エラーが発見されておらず、参照訳に過度に合わせようとした可能性がある。

[Gemma 請求項#29 前後処理なし プロンプト工夫なし]

原文	The control device according to claim 1, wherein the first target rotation speed is a predetermined minimum rotation speed.
参照訳	前記第1目標回転数が、所定の最低回転数である、請求項1に記載の制御装置。
機械翻訳	請求項1に記載の制御装置において、第一の目標回転速度は、予め定められた最小回転速度である。
評価③結果	スコア:90
評価③理由	<p>This is a high-quality translation that accurately conveys the full meaning of the source text. All technical terms (“first target rotation speed”, “predetermined minimum rotation speed”, “control device”, “claim 1”) are translated correctly. The main point of distinction is the sentence structure, which is a matter of style and convention in patent writing. * **Human Reference:** `前記第1目標回転数が、所定の最低回転数である、請求項1に記載の制御装置。` This uses a very common and professional structure for Japanese dependent claims, where the limiting feature (`...である、`) directly modifies the noun (`...制御装置`) at the end. It is a single, concise noun phrase. * **Translation to Score:** `請求項1に記載の制御装置において、第一の目標回転速度は、予め定められた最小回転速度である。` This structure breaks the claim into two parts: “In the control device according to claim 1,” and then a separate declarative sentence “the first target rotation speed is...”. <b><u>While grammatically correct and perfectly understandable, this `...において、...は...である` structure is less common and slightly less elegant for a dependent claim than the structure used in the reference.</u></b> It reads more like a sentence from the detailed description section of the patent rather than the claim itself. The word choices (`第一の` vs. `前記第1`, `予め定められた` vs. `所定の`, `回転速度` vs. `回転数`) are all acceptable and largely synonymous in this context. The translation is excellent in terms of meaning but slightly less idiomatic in terms of patent claim structure compared to the human reference. It is a very good translation, but not perfect from a professional patent translator’s stylistic perspective. [[[90]]]</p>
備考	<p>評価①結果 エラーカテゴリ:no-error サブカテゴリ:no-error エラーレベル:no-error          評価②結果 正確性(文):5/5 正確性(用語):3/4 流暢さ:4/5</p>

この事例では、「において、...は...である」という構成について、減点がされている。請求項の記載として、「である。」で終了することは一般的では無いと考えられるところ、理由の出力でも、従属項の記載として一般的であるかが考慮されていることが伺える。なお、参照訳を与えない評価手法①では、エラーが発見されておらず、参照訳の提示が有効に働いた可能性がある。

## 8.4 評価結果に基づいた調査・分析

自動評価・人手評価の過程及び結果をもとに、以下の観点について調査・分析を行い、大規模言語モデルによる機械翻訳の課題とその解決策を検討した。また、大規模言語モデルによる精度評価の有用性について、自動評価・人手評価と比較した。さらに、文中タグや定型文を多く含むといった特徴を有する実際の特許文書を想定した分析を行った。

(調査・分析観点)

- ・ 翻訳精度と機械翻訳の誤りの傾向
- ・ 追加学習、前後処理、プロンプトの工夫による改善点や改悪点
- ・ 同一語の訳ゆれの程度

### 8.4.1 翻訳精度と機械翻訳の誤りの傾向

翻訳精度に関しては、語彙やフレーズの参照訳との一致を主に評価する BLEU と、BLEU よりも語順を重視して評価を行う RIBES の値から考察することができる。

自動評価結果は第 8.1.7 項に示したとおりであり、全言語全項目の BLEU は、Gemini が 53.58、Claude が 51.52、Gemma が 45.63 となり、いずれも nmt の BLEU の 56.45 よりも低い値となった。ただし、最も低い Gemma であっても一般に機械翻訳で高品質と言える BLEU 40 を大きく上回った。

また、機械翻訳の誤りの傾向については、BLEU の score からより深く分析することができる。具体的には、score1 が低いのは語彙選択ミス・訳抜けが、score2 が低い場合は不自然な語のつながりが、score3 が低い場合は文構造の乱れが、score4 が低い場合は語彙・流暢性の問題が生じているといえる。

結果としては、第 8.1.7 項に示したとおりであり、LLM モデル 3 種 (Claude、Gemini、Gemma) についての傾向をみると、Gemini は score1 が高く、score4 も比較的高い、また、Claude は score1 が高く、score4 は伸びにくい、そして、Gemma は score1~4 が全体的に低かった。また、nmt は、score1~4 が全体的に高く、BLEU 自体も他の LLM モデル 3 種よりも高いという結果となった。

また、RIBES については、全言語全項目の平均値について、Gemini の RIBES が 86.22、Claude の RIBES が 84.70、Gemma の RIBES が 81.84 となり、いずれも nmt の RIBES の 88.54 よりも低い値となった。ただし、1 位の nmt と、4 位の Gemma との差は 6.7 であ

り、BLEU の差 10.82 よりも小さい。

このことは語彙やフレーズの一致を主に評価する BLEU と、BLEU よりも語順を重視して評価を行う RIBES の特性を踏まえると、各モデル間の差は語順や構文能力の違いによるものというよりも、語彙やフレーズの参照訳への忠実性の違いに起因すると考えられる。

ただし、自動評価値から考察可能な「機械翻訳の誤りの傾向」は、あくまで参照訳からの語彙・フレーズ一致面での差の有無であることに注意が必要である。

人手評価では、LLM モデル 3 種 (Claude、Gemini、Gemma) に対し、「内容伝達レベルの評価」、「流暢さの評価」、「重要技術用語の訳質評価」、「誤訳のカテゴリ別カウント」及び「連続文での技術用語の訳ゆれの評価」を実施した。評価に際しては、特許文献由来の大量の対訳コーパスを学習させたニューラル機械翻訳モデル (nmt) を対照に用いた。

結果は第 8.2.8 項に示したとおりであり、Claude と Gemini は、翻訳文の実用性の総合評価といえる内容伝達レベルの評価をはじめ、各評価観点において nmt を上回る高評価となった。nmt の内容伝達レベル自体、全言語平均で 4.71 (5 段階評価) と機械翻訳として十分に良好であるが、Claude は 4.84、Gemini は 4.86 と、これをさらに上回る高評価であった。Gemma は 4.56 と他の 2 モデルとはやや差があり、nmt よりも低値となったが、レベル 4 は大きく上回っており、特許文の機械翻訳品質としては良好な水準であった。

翻訳の誤り傾向に関しては、誤訳のカテゴリ別カウントの集計結果のとおり、各 LLM モデルとも「①技術用語の誤訳」が多かった。ただし、これは nmt も同様であり、特許文を翻訳対象とした際の通常の傾向といえる。また、特に Claude と Gemma は英日・中日において「請求項であるのに体言止めとなっていない」不備が多発したが、中日では nmt も同じ状況であり、むしろ、Gemini のみ本事象の発生数が他より大幅に少ない点がこのモデルの長所の一つとなっている。

他の誤訳カテゴリに関しても、特定のモデルのみで多発しているものは見られず、nmt より翻訳精度が良好な Claude と Gemini は各種誤訳のカウントも全体的に少なくなり、nmt より低評価であった Gemma ではカウントが多くなっている。つまり、各 LLM モデルとも誤訳の傾向自体は nmt とおおむね同様であり、LLM モデル特有の懸念すべき誤訳の傾向は検出されなかった。

LLM 3 モデルと、nmt との合計 4 モデルについての評価結果は、人手評価 (翻訳文の実用性の総合評価といえる内容伝達レベル) と自動評価とで異なるものとなったが、自動評価の同義語・自然な言いかえを評価しにくい点、語順の違いが許容されにくい点等が評価の差

として現れたものといえる。

#### 8.4.2 追加学習、前後処理、プロンプトの工夫による改善点や改悪点

人手評価では、Claude ベースモデルに前後処理を導入した 3 モデルと、同じく Claude をベースにしてプロンプトに工夫を加えた 6 モデル、そして Gemma をベースに特許由来の対訳コーパスを 1,500 件追加学習させた 1 モデルを評価対象とした。これらのモデルの評価スコアをベースモデルと対比することで、前後処理やプロンプトの工夫、特許由来コーパスの追加学習による翻訳品質改善効果を測った。

前後処理に関しては、3 種のモデル（前後処理①、②、⑤）をベースモデルと比較したが、各モデルとも評価値は小さな変動にとどまり、また言語・項目によって向上と低下が混在する結果となった。Claude がベースモデルの時点で非常に翻訳精度が高く、前後処理で期待される水準にすでに達していたため、誤差レベルの小変動に終始したものと推測される。そうした中、前後処理①は、ベースモデルで改善の余地が比較的大きかった中日の重要技術用語の A 判定率において顕著に数値が上昇しており、一定の改善効果が得られたと判断される。前後処理②モデルについても、評価対象言語や項目が少なく、かつ前後処理①や⑤と異なるため横並びの評価は難しいものの、評価対象（英日特許文献、日英審査書類）に関してはプラス方向の変動が大半で、かつ改善の度合いも比較的大きかった。このため、少なくとも英日特許と日英審査書類に関しては一定の改善効果が得られる可能性が高い。

自動評価に関しては、前後処理①は BLEU について韓日の数値向上が特に著しかった（ベースの Claude について BLEU が 51.06 から、57.45 に向上）ものの、人手評価値の変化は内容伝達レベルが 4.91 から 4.92 に変化した等、ほぼなかったといえる。前後処理②については、ベースモデルに比べて劣化こそしなかったものの、BLEU、RIBES とも、ほぼ効果がないといえるレベルであった。前後処理⑤は前後処理①同様、BLEU について韓日の数値向上が特に著しかった（ベースの Claude について BLEU が 51.06 から、60.27 に向上）ものの人手評価はほぼ横ばいであった。興味深いことに、RIBES は前後処理①ではベースモデル 82.41 に対して、82.57 と上昇幅が大きくなかったものの、前後処理⑤では 86.99 に向上している。

プロンプトの工夫に関しては、プロンプト①、②、③、④、⑧、⑨を Claude ベースモデルに使用した 6 モデルを評価した。これらも「ベースモデルの時点で翻訳精度が高く、改善の余地が少ない」点では前後処理と同じであり、評価結果も同様の傾向となった。そのような中、プロンプトに「技術分野の情報として、IPC の説明文（サブクラス）を指定したプロンプト②モデルと、プロンプト②に他の工夫を組み合わせたプロンプト③～⑨モデルのうちの 3 モデル（②、③、⑧）において、中日の重要技術用語の A 判定が顕著に向上した。

それぞれ評価対象とした項目が異なるためモデル間の優劣はつけ難いが、これら 3 種のプロンプトに関しては、重要技術用語の適訳性において一定の効果を有すると判断される。

その点、自動評価についてはいずれのプロンプトについても効果は実質的にみられなかった。

追加学習に関しては、Gemma ベースモデルに特許由来の対訳コーパス 1,500 文を追加学習させたモデル (f1500\_gemma) の評価を行った。Gemma は 3 種の LLM モデルの中では最も評価結果が低く、特許翻訳に関しては改善の余地が大きいモデルといえたが、結果的には各言語、各項目とも評価結果は誤差レベルの小変動にとどまり、顕著な翻訳改善効果は見られなかった。自動評価結果についても同様の傾向であった。この傾向は本事業の学習データが 1500 または 2500 文と比較的少量であったことが要因の 1 つとして考えられる。

なお、Claude、Gemma とともにベースモデルで多発していた英日・中日の「請求項が体言止めで訳されない」不備に関しては、「特許請求の範囲」を対象とした全ての前後処理／プロンプト／追加学習モデルのうち、プロンプト①、③、④で英日の発生数が半減し、プロンプト④では中日の発生数も半減した。これら 3 種のモデルはいずれも「翻訳対象の項目（発明の名称、要約、請求の範囲、明細書等）および、想定読者に「特許のサーチャー」を指定」するプロンプトを採用しており、このプロンプトに請求項の文体改善効果があると考えられる。

#### 8.4.3 同一語の訳ゆれ程度

本評価では、同一語の訳ゆれに関して、誤訳のカテゴリ別カウントの「⑧ 訳ゆれ」と、「連続文での技術用語の訳ゆれの評価」の 2 種の評価を行った。

誤訳カテゴリ別カウントの「⑧ 訳ゆれ」は、1 文中に同一の（訳語統一されるべき）技術用語が複数回出現する原文のみを対象に、その技術用語が文中で同一の訳語に訳されていない場合に誤訳としてカウントした。一方、「連続文での技術用語の訳ゆれの評価」については、テストセット中に同一文献から「同一の技術用語を含む連続する 2 文」を所定数含めておき、その技術用語が 2 文のペアにおいて同一の訳語に訳されているか否かを判定した。つまり、前者は「同一文中の訳ゆれ」、後者は「異なる文間の訳ゆれ」のチェックである。

前者、すなわち「同一文中の訳ゆれ」については、nmt のカウント数が全言語合計で 9 文（⇒8.2.8.4）であるのに対し、Claude が 4 文、Gemini が 1 文、Gemma が 4 文であった。一方、後者の「異なる文間の訳ゆれ」については、全言語合計（対象 148 ペア）で nmt の

25 ペアに対し、Claude が 16 ペア、Gemini が 18 ペア、Gemma が 24 ペアであった。

これらの結果からは、LLM モデル、とりわけ Claude と Gemini は、nmt に比べて訳ゆれがやや発生しにくい性質と見なせる。とはいえ、両モデルとも訳ゆれ自体は一定量発生しており、LLM 方式においても文中・文間の訳ゆれは解消されていない。LLM モデルも nmt と同様、用語辞書を用いず学習データ中から最良と判断した訳語を採用する方式であり、また本評価の対象とした LLM モデルはいずれも文単位で独立した翻訳を行う方式であって、先行する文で採用した訳語の情報が保持されない点も nmt と同じである。このような方式では、技術用語の訳ゆれは宿命的なものとなる。結果、今回評価対象とした各 LLM モデルにおいても技術用語の訳ゆれは根本的には解消されていない。

#### 8.4.4 大規模言語モデルによる評価の有用性

##### 8.4.4.1 各手法と人手評価との比較

大規模言語モデルによる評価の有用性を検討するため、人手評価結果との関係を分析した。具体的には、同一文の大規模言語モデルによる評価の手法①、②、③の評価結果と人手評価結果がどの程度一致しているかを比較した。比較した結果を言語方向毎に集計し、その結果を分析した。

##### 評価手法①と人手評価との比較

大規模言語モデルによる評価の手法①の評価結果は、文の誤訳、流暢さ、用語の誤訳観点で誤訳有無の指摘をする方式である。指摘は 1 文に対して複数指摘されることもある。一方人手評価は、1 文に対して 1 つの値で評価されるため直接の比較ができない。そのため、評価手法①の評価結果を下記の手順で 1 つの評価値に変換して、変換した評価値と人手評価結果とを比較した(表 8.4.4.1-1～表 8.4.4.1-12)。

評価手法①は評価観点（文の誤訳、流暢さ、用語の誤訳）毎に誤訳レベルを提示する。誤訳レベルは Critical、Major、Minor の三種類がある。この誤訳レベルの個数に応じ、評価値の満点から減点する方式で正確性（文）、流暢さ、正確性（用語）の評価値を決めた。減点は、誤訳レベル Critical ⇒ 評価値 1、Major ⇒ マイナス 2、Minor ⇒ マイナス 1 とし、正確性、流暢さは 5 点満点からの減点、用語は 4 点満点からの減点とした。減点結果が 1 よりも小さい値となったときは、評価値を 1 とした<sup>25</sup>。(例) 文の誤訳について Major が 1 つ、Minor が 1 つの場合、5(満点) - 2(Major) - 1(Minor) ⇒ 評価値 2 となる。

##### 評価手法①の評価値の変換方法

<sup>25</sup> 各誤訳レベルの減点値は、人手評価との一致度、相関が高くなる値を使用した。

表 8.4.4.1-1 人手評価（内容伝達）と正確性（文）との関係（英日）

人手評価(内容伝達)	評価手法①(正確性(文))					総計
	5	4	3	2	1	
5	76.06%	0.64%	2.44%	0.06%	0.33%	79.53%
4	12.31%	0.61%	2.11%	0.06%	0.56%	15.64%
3	2.50%	0.19%	0.83%	0.03%	0.22%	3.78%
2	0.42%	0.06%	0.39%	0.06%	0.14%	1.06%
<b>総計</b>	<b>91.28%</b>	<b>1.50%</b>	<b>5.78%</b>	<b>0.19%</b>	<b>1.25%</b>	<b>100.00%</b>

表 8.4.4.1-2 人手評価（内容伝達）と正確性（文）との関係（中日）

人手評価(内容伝達)	評価手法①(正確性(文))					総計
	5	4	3	2	1	
5	72.72%	0.91%	2.34%	0.06%	0.41%	76.44%
4	12.91%	0.72%	1.97%	0.06%	0.19%	15.84%
3	4.03%	0.16%	1.38%	0.00%	0.66%	6.22%
2	1.09%	0.00%	0.09%	0.00%	0.06%	1.25%
1	0.16%	0.00%	0.03%	0.00%	0.06%	0.25%
<b>総計</b>	<b>90.91%</b>	<b>1.78%</b>	<b>5.81%</b>	<b>0.13%</b>	<b>1.38%</b>	<b>100.00%</b>

表 8.4.4.1-3 人手評価（内容伝達）と正確性（文）との関係（韓日）

人手評価(内容伝達)	評価手法①(正確性(文))					総計
	5	4	3	2	1	
5	84.94%	0.66%	2.41%	0.06%	0.34%	88.41%
4	6.53%	0.31%	1.81%	0.09%	1.09%	9.84%
3	0.75%	0.06%	0.41%	0.00%	0.53%	1.75%
<b>総計</b>	<b>92.22%</b>	<b>1.03%</b>	<b>4.63%</b>	<b>0.16%</b>	<b>1.97%</b>	<b>100.00%</b>

表 8.4.4.1-4 人手評価（内容伝達）と正確性（文）との関係（日英）

人手評価(内容伝達)	評価手法①(正確性(文))					総計
	5	4	3	2	1	
5	76.75%	1.97%	2.78%	0.00%	0.28%	81.78%
4	11.44%	0.81%	1.28%	0.09%	0.44%	14.06%
3	1.84%	0.16%	1.06%	0.00%	0.44%	3.50%
2	0.22%	0.03%	0.38%	0.03%	0.00%	0.66%
<b>総計</b>	<b>90.25%</b>	<b>2.97%</b>	<b>5.50%</b>	<b>0.13%</b>	<b>1.16%</b>	<b>100.00%</b>

表 8.4.4.1-5 人手評価（流暢さ）と流暢さとの関係（英日）

人手評価(流暢さ)	評価手法①(流暢さ)					総計
	5	4	3	2	1	
5	73.00%	0.47%	0.81%	0.00%	0.00%	74.28%
4	23.47%	0.17%	1.94%	0.03%	0.03%	25.64%
3	0.08%	0.00%	0.00%	0.00%	0.00%	0.08%
<b>総計</b>	<b>96.56%</b>	<b>0.64%</b>	<b>2.75%</b>	<b>0.03%</b>	<b>0.03%</b>	<b>100.00%</b>

表 8.4.4.1-6 人手評価（流暢さ）と流暢さとの関係（中日）

人手評価(流暢さ)	評価手法①(流暢さ)					総計
	5	4	3	2	1	
5	73.06%	0.53%	1.25%	0.00%	0.03%	74.88%
4	17.94%	0.72%	0.94%	0.16%	0.09%	19.84%
3	3.16%	0.22%	0.34%	0.06%	0.06%	3.84%
2	0.91%	0.06%	0.03%	0.03%	0.00%	1.03%
1	0.34%	0.00%	0.03%	0.03%	0.00%	0.41%
<b>総計</b>	<b>95.41%</b>	<b>1.53%</b>	<b>2.59%</b>	<b>0.28%</b>	<b>0.19%</b>	<b>100.00%</b>

表 8.4.4.1-7 人手評価（流暢さ）と流暢さとの関係（韓日）

人手評価(流暢さ)	評価手法①(流暢さ)					総計
	5	4	3	2	1	
5	92.44%	1.09%	1.88%	0.06%	0.03%	95.50%
4	3.81%	0.22%	0.34%	0.00%	0.06%	4.44%
3	0.00%	0.00%	0.03%	0.00%	0.03%	0.06%
<b>総計</b>	<b>96.25%</b>	<b>1.31%</b>	<b>2.25%</b>	<b>0.06%</b>	<b>0.13%</b>	<b>100.00%</b>

表 8.4.4.1-8 人手評価（流暢さ）と流暢さとの関係（日英）

人手評価(流暢さ)	評価手法①(流暢さ)					総計
	5	4	3	2	1	
5	90.94%	1.38%	0.75%	0.06%	0.00%	93.13%
4	5.31%	0.31%	0.59%	0.00%	0.06%	6.28%
3	0.31%	0.00%	0.22%	0.00%	0.03%	0.56%
2	0.03%	0.00%	0.00%	0.00%	0.00%	0.03%
<b>総計</b>	<b>96.59%</b>	<b>1.69%</b>	<b>1.56%</b>	<b>0.06%</b>	<b>0.09%</b>	<b>100.00%</b>

表 8.4.4.1-9 人手評価（重要技術用語）と正確性(用語)との関係（英日）

人手評価(重要技術用語)	評価手法①正確性（用語）				総計
	4	3	2	1	
A	90.21%	2.14%	1.39%	0.20%	93.94%
B	5.48%	0.35%	0.17%	0.06%	6.06%
<b>総計</b>	<b>95.70%</b>	<b>2.48%</b>	<b>1.56%</b>	<b>0.26%</b>	<b>100.00%</b>

表 8.4.4.1-10 人手評価（重要技術用語）と正確性(用語)との関係（中日）

人手評価(重要技術用語)	評価手法①正確性（用語）				総計
	4	3	2	1	
A	78.81%	1.53%	2.10%	0.22%	82.66%
B	9.18%	0.70%	0.51%	0.13%	10.52%
C	4.81%	0.35%	0.70%	0.19%	6.05%
D	0.41%	0.03%	0.32%	0.00%	0.76%
<b>総計</b>	<b>93.21%</b>	<b>2.61%</b>	<b>3.63%</b>	<b>0.54%</b>	<b>100.00%</b>

表 8.4.4.1-11 人手評価（重要技術用語）と正確性（用語）との関係（韓日）

人手評価(重要技術用語)	評価手法①正確性（用語）				総計
	4	3	2	1	
A	88.35%	1.31%	2.02%	0.38%	92.06%
B	4.87%	0.19%	0.38%	0.03%	5.47%
C	1.54%	0.03%	0.42%	0.42%	2.40%
D	0.06%	0.00%	0.00%	0.00%	0.06%
<b>総計</b>	<b>94.81%</b>	<b>1.54%</b>	<b>2.82%</b>	<b>0.83%</b>	<b>100.00%</b>

表 8.4.4.1-12 人手評価（重要技術用語）と正確性（用語）との関係（日英）

人手評価(重要技術用語)	評価手法①正確性（用語）				総計
	4	3	2	1	
A	92.39%	1.67%	0.80%	0.10%	94.96%
B	3.66%	0.10%	0.42%	0.06%	4.24%
C	0.58%	0.06%	0.06%	0.00%	0.71%
D	0.06%	0.00%	0.03%	0.00%	0.10%
<b>総計</b>	<b>96.69%</b>	<b>1.83%</b>	<b>1.32%</b>	<b>0.16%</b>	<b>100.00%</b>

## 評価手法②と人手評価との比較

大規模言語モデルによる評価の手法②の評価結果は、各評価文に対して文の正確性、文の流暢さ、用語の正確性の3つの観点で評価を行う。評価値は5～1の5段階（文の正確性、文の流暢さ）、4～1の4段階（用語の正確性）で評価結果を提示する（表 8.4.4.1-13～表 8.4.4.1-24）。

表 8.4.4.1-13 人手評価（内容伝達）と正確性（文）との関係（英日）

人手評価(内容伝達)	評価手法②(正確性(文))					総計
	5	4	3	2	1	
5	75.39%	2.47%	1.47%	0.19%	0.00%	79.53%
4	11.94%	1.75%	1.36%	0.53%	0.06%	15.64%
3	2.53%	0.50%	0.56%	0.14%	0.06%	3.78%
2	0.36%	0.28%	0.31%	0.08%	0.03%	1.06%
<b>総計</b>	<b>90.22%</b>	<b>5.00%</b>	<b>3.69%</b>	<b>0.94%</b>	<b>0.14%</b>	<b>100.00%</b>

表 8.4.4.1-14 人手評価（内容伝達）と正確性（文）との関係（中日）

人手評価(内容伝達)	評価手法②(正確性(文))					総計
	5	4	3	2	1	
5	73.28%	2.22%	0.84%	0.09%	0.00%	76.44%
4	12.44%	2.47%	0.88%	0.06%	0.00%	15.84%
3	3.97%	0.88%	0.78%	0.53%	0.06%	6.22%
2	0.81%	0.13%	0.22%	0.06%	0.03%	1.25%
1	0.03%	0.03%	0.13%	0.03%	0.03%	0.25%
<b>総計</b>	<b>90.53%</b>	<b>5.72%</b>	<b>2.84%</b>	<b>0.78%</b>	<b>0.13%</b>	<b>100.00%</b>

表 8.4.4.1-15 人手評価（内容伝達）と正確性（文）との関係（韓日）

人手評価(内容伝達)	評価手法②(正確性(文))					総計
	5	4	3	2	1	
5	84.41%	1.94%	1.50%	0.53%	0.03%	88.41%
4	5.47%	1.53%	1.88%	0.78%	0.19%	9.84%
3	0.09%	0.19%	0.53%	0.66%	0.28%	1.75%
<b>総計</b>	<b>89.97%</b>	<b>3.66%</b>	<b>3.91%</b>	<b>1.97%</b>	<b>0.50%</b>	<b>100.00%</b>

表 8.4.4.1-16 人手評価（内容伝達）と正確性（文）との関係（日英）

評価手法②(正確性(文))					
人手評価(内容伝達)	5	4	3	2	総計
5	79.31%	1.56%	0.81%	0.09%	81.78%
4	11.38%	1.75%	0.66%	0.28%	14.06%
3	2.13%	0.41%	0.56%	0.41%	3.50%
2	0.44%	0.06%	0.16%	0.00%	0.66%
<b>総計</b>	<b>93.25%</b>	<b>3.78%</b>	<b>2.19%</b>	<b>0.78%</b>	<b>100.00%</b>

表 8.4.4.1-17 人手評価（流暢さ）と流暢さとの関係（英日）

評価手法②(流暢さ)					
人手評価(流暢さ)	5	4	3	2	総計
5	60.39%	12.11%	1.64%	0.14%	74.28%
4	9.39%	14.08%	2.03%	0.14%	25.64%
3	0.03%	0.03%	0.03%	0.00%	0.08%
<b>総計</b>	<b>69.81%</b>	<b>26.22%</b>	<b>3.69%</b>	<b>0.28%</b>	<b>100.00%</b>

表 8.4.4.1-18 人手評価（流暢さ）と流暢さとの関係（中日）

評価手法②(流暢さ)						
人手評価(流暢さ)	5	4	3	2	1	総計
5	56.28%	16.13%	2.38%	0.06%	0.03%	74.88%
4	9.63%	7.72%	2.38%	0.13%	0.00%	19.84%
3	1.41%	1.69%	0.59%	0.16%	0.00%	3.84%
2	0.47%	0.34%	0.16%	0.03%	0.03%	1.03%
1	0.03%	0.09%	0.25%	0.00%	0.03%	0.41%
<b>総計</b>	<b>67.81%</b>	<b>25.97%</b>	<b>5.75%</b>	<b>0.38%</b>	<b>0.09%</b>	<b>100.00%</b>

表 8.4.4.1-19 人手評価（流暢さ）と流暢さとの関係（韓日）

評価手法②(流暢さ)						
人手評価(流暢さ)	5	4	3	2	1	総計
5	63.44%	28.00%	3.41%	0.63%	0.03%	95.50%
4	2.34%	1.75%	0.31%	0.00%	0.03%	4.44%
3	0.00%	0.00%	0.06%	0.00%	0.00%	0.06%
<b>総計</b>	<b>65.78%</b>	<b>29.75%</b>	<b>3.78%</b>	<b>0.63%</b>	<b>0.06%</b>	<b>100.00%</b>

表 8.4.4.1-20 人手評価（流暢さ）と流暢さとの関係（日英）

人手評価(流暢さ)	評価手法②(流暢さ)				総計
	5	4	3	2	
5	81.34%	10.84%	0.91%	0.03%	93.13%
4	2.94%	2.72%	0.63%	0.00%	6.28%
3	0.09%	0.19%	0.25%	0.03%	0.56%
2	0.00%	0.03%	0.00%	0.00%	0.03%
<b>総計</b>	<b>84.38%</b>	<b>13.78%</b>	<b>1.78%</b>	<b>0.06%</b>	<b>100.00%</b>

表 8.4.4.1-21 人手評価（重要技術用語）と正確性（用語）との関係（英日）

人手評価(重要技術用語)	評価手法②正確性（用語）				総計
	4	3	2	1	
A	76.76%	13.65%	2.46%	0.03%	92.89%
B	3.48%	1.94%	0.57%	0.00%	5.99%
C	0.20%	0.51%	0.40%	0.00%	1.11%
<b>総計</b>	<b>80.45%</b>	<b>16.10%</b>	<b>3.43%</b>	<b>0.03%</b>	<b>100.00%</b>

表 8.4.4.1-22 人手評価（重要技術用語）と正確性（用語）との関係（中日）

人手評価(重要技術用語)	評価手法②正確性（用語）				総計
	4	3	2	1	
A	68.61%	10.68%	3.31%	0.06%	82.66%
B	4.53%	5.10%	0.89%	0.00%	10.52%
C	2.17%	2.33%	1.56%	0.00%	6.05%
D	0.13%	0.38%	0.22%	0.03%	0.76%
<b>総計</b>	<b>75.43%</b>	<b>18.48%</b>	<b>5.99%</b>	<b>0.10%</b>	<b>100.00%</b>

表 8.4.4.1-23 人手評価（重要技術用語）と正確性（用語）との関係（韓日）

人手評価(重要技術用語)	評価手法②正確性（用語）				総計
	4	3	2	1	
A	71.41%	15.88%	4.64%	0.13%	92.06%
B	2.27%	2.53%	0.67%	0.00%	5.47%
C	0.13%	0.38%	1.89%	0.00%	2.40%
D	0.03%	0.00%	0.03%	0.00%	0.06%
<b>総計</b>	<b>73.85%</b>	<b>18.79%</b>	<b>7.23%</b>	<b>0.13%</b>	<b>100.00%</b>

表 8.4.4.1-24 人手評価（重要技術用語）と正確性（用語）との関係（日英）

人手評価(重要技術用語)	評価手法②(正確性(用語))				総計
	4	3	2	1	
A	86.42%	6.33%	2.15%	0.06%	94.96%
B	2.73%	1.16%	0.35%	0.00%	4.24%
C	0.19%	0.35%	0.16%	0.00%	0.71%
D	0.10%	0.00%	0.00%	0.00%	0.10%
<b>総計</b>	<b>89.43%</b>	<b>7.84%</b>	<b>2.67%</b>	<b>0.06%</b>	<b>100.00%</b>

### 評価手法③と人手評価との比較

大規模言語モデルによる評価手法③は、各評価文に対して100～0の評価値を提示する。評価手法③では複数の評価観点がないことから、人手評価結果の内容伝達の結果との関係を調査した。なお、評価手法③の評価値の範囲は人手評価の内容伝達の評価値5～1と異なるため、評価手法③の評価値を5～1に変換し比較を行った<sup>26</sup>(表 8.4.4.1-25～表 8.4.4.1-28)。

表 8.4.4.1-25 人手評価（内容伝達）と評価手法③の評価との関係（英日）

人手評価(内容伝達)	評価手法③（5段階に変換）					総計
	5	4	3	2	1	
5	73.94%	4.64%	0.81%	0.14%	0.00%	79.53%
4	12.89%	1.44%	0.92%	0.39%	0.00%	15.64%
3	2.64%	0.72%	0.22%	0.17%	0.03%	3.78%
2	0.44%	0.25%	0.28%	0.08%	0.00%	1.06%
<b>総計</b>	<b>89.92%</b>	<b>7.06%</b>	<b>2.22%</b>	<b>0.78%</b>	<b>0.03%</b>	<b>100.00%</b>

表 8.4.4.1-26 人手評価（内容伝達）と評価手法③の評価との関係（中日）

人手評価(内容伝達)	評価手法③（5段階に変換）					総計
	5	4	3	2	1	
5	70.72%	4.75%	0.88%	0.09%	0.00%	76.44%
4	11.00%	3.88%	0.88%	0.09%	0.00%	15.84%
3	3.38%	1.56%	0.69%	0.53%	0.06%	6.22%
2	0.75%	0.22%	0.16%	0.13%	0.00%	1.25%

<sup>26</sup> 方式③の評価値を0～19, 20～39, 40～59, 60～79, 80～100の5段階に変換した。

1	0.00%	0.00%	0.19%	0.00%	0.06%	0.25%
<b>総計</b>	<b>85.84%</b>	<b>10.41%</b>	<b>2.78%</b>	<b>0.84%</b>	<b>0.13%</b>	<b>100.00%</b>

表 8.4.4.1-27 人手評価（内容伝達）と評価手法③の評価との関係（韓日）

人手評価(内容伝達)	評価手法③（5段階に変換）					総計
	5	4	3	2	1	
5	82.22%	4.63%	1.38%	0.19%	0.00%	88.41%
4	5.19%	2.22%	1.59%	0.81%	0.03%	9.84%
3	0.09%	0.25%	0.50%	0.66%	0.25%	1.75%
<b>総計</b>	<b>87.50%</b>	<b>7.09%</b>	<b>3.47%</b>	<b>1.66%</b>	<b>0.28%</b>	<b>100.00%</b>

表 8.4.4.1-28 人手評価（内容伝達）と評価手法③の評価との関係（日英）

人手評価(内容伝達)	評価手法③（5段階に変換）					総計
	5	4	3	2	1	
5	80.38%	0.94%	0.38%	0.09%	0.00%	81.78%
4	12.34%	1.13%	0.44%	0.16%	0.00%	14.06%
3	2.25%	0.59%	0.47%	0.16%	0.03%	3.50%
2	0.41%	0.22%	0.03%	0.00%	0.00%	0.66%
<b>総計</b>	<b>95.38%</b>	<b>2.88%</b>	<b>1.31%</b>	<b>0.41%</b>	<b>0.03%</b>	<b>100.00%</b>

以上の結果をもとに各手法がどの程度人手評価と近い結果を示すかを、評価値の一致度と、相関係数を計算した(表 8.4.4.1-29～表 8.4.4.1-32)。一致度の計算には、重み付けカッパ係数 (weighted  $\kappa$ )<sup>27</sup>を使用した。相関係数にはいくつか種類があるが、ここでは翻訳評価手法自体の評価で使用される手法の1つであるスピアマンの順位相関係数<sup>28</sup>を使用した。一致度を見ると、韓日の評価手法②、③と内容伝達との間である程度一致した。

表 8.4.4.1-29 各手法と人手評価の一致度

	英日	中日	韓日	日英
--	----	----	----	----

<sup>27</sup> weighted  $\kappa$  (重み付けカッパ係数) とは、カテゴリに“順序”があるときに、不一致の程度に応じてペナルティ (重み) を変えて一致度合いを計算する手法。詳細については、[https://en.wikipedia.org/wiki/Cohen%27s\\_kappa](https://en.wikipedia.org/wiki/Cohen%27s_kappa) や <https://www.med.osaka-u.ac.jp/pub/kid/clinicaljournalclub12.html> 参照のこと。

<sup>28</sup> 相関係数および値の解釈は、<https://xgrapher.com/articles/spearman> を参考とした。

	内容伝達	流暢さ	内容伝達	流暢さ	内容伝達	流暢さ	内容伝達	流暢さ
評価手法①	0.32	0.14	0.28	0.14	0.37	0.14	0.33	0.27
評価手法②	0.33	0.38	0.35	0.31	<b>0.51</b>	0.03	0.33	0.31
評価手法③	0.26	-	0.38	-	<b>0.52</b>	-	0.27	-

表 8.4.4.1-30 一致度の目安

一致度	解釈
0.8 ~ ≤1	ほぼ完全に一致
0.6 ~ ≤0.8	かなり一致
0.4 ~ ≤0.6	ある程度一致
0.2 ~ ≤0.4	少しは一致
0 ~ ≤0.2	わずかに一致
~ ≤0	不一致

相関関係の結果から英日で評価手法②と流暢さとの間で中程度の相関が、中日は評価手法③と内容伝達に中程度の相関が、韓日は評価手法②と内容伝達に中程度の相関がみられた。

表 8.4.4.1-31 各手法と人手評価との相関係数

	英日		中日		韓日		日英	
	内容伝達	流暢さ	内容伝達	流暢さ	内容伝達	流暢さ	内容伝達	流暢さ
評価手法①	0.31	0.16	0.27	0.18	0.40	0.13	0.27	0.21
評価手法②	0.31	<b>0.42</b>	0.34	0.29	<b>0.52</b>	0.06	0.32	0.31
評価手法③	0.22	-	<b>0.41</b>	-	0.36	-	0.35	-

表 8.4.4.1-32 相関の強さの目安

相関係数	解釈
0.7 ~ 1.0	強い相関がある
0.4 ~ 0.7	中程度の相関がある
0.2 ~ 0.4	弱い相関がある
0.0 ~ 0.2	ほとんど相関なし

#### 8.4.4.2 自動評価との比較

機械翻訳の評価は人手による評価と機械が評価を行う自動評価に分類される。大規模言語モデルによる評価は自動評価手法の1つとなることから、自動評価として有用となり得るかの調査を行った。調査は自動評価 BLEU、評価手法①の正確性、評価手法②の正確性、評価手法③のスコアそれぞれについて人手評価の内容伝達レベルとの相関係数を求め、どの評価が人手評価との間で最も相関しているかを比較した。比較には、各翻訳のモデル単位の平均値(表 8.4.4.2-1)から相関係数を求めた(表 8.4.4.2-2)。

表 8.4.4.2-1 各翻訳モデルと内容伝達、正確性 (文)、BLEU の平均

	内容伝達	評価手法①	評価手法②	評価手法③	BLEU
<b>大規模言語モデルによる翻訳</b>					
Claude	4.85	4.95	4.93	95.29	44.35
Gemini	4.86	4.98	4.96	97.36	47.41
Gemma	4.56	4.76	4.67	89.61	40.13
<b>前後処理を適用した翻訳</b>					
<b>前後処理①</b>					
Claude	4.79	4.93	4.90	93.97	54.29
<b>前後処理②</b>					
Claude	4.85	4.95	4.93	96.92	43.33
<b>前後処理⑤</b>					
Claude	4.78	4.94	4.92	94.84	52.04
<b>プロンプトを工夫した翻訳</b>					
<b>prompt1</b>					
Claude	4.84	4.95	4.92	94.51	47.71
<b>prompt2</b>					
Claude	4.94	4.96	4.92	96.77	38.96
<b>prompt3</b>					
Claude	4.87	4.93	4.90	91.14	50.51
<b>prompt4</b>					
Claude	4.86	4.91	4.90	91.53	48.55
<b>prompt8</b>					
Claude	4.73	4.95	4.92	95.98	50.62
<b>prompt9</b>					
Claude	4.73	4.94	4.93	96.72	51.11

追加学習済大規模言語モデル					
f1500_gemma	4.55	4.79	4.71	89.84	39.67
ニューラル機械翻訳					
nmt	4.73	4.79	4.79	92.59	56.61

表 8.4.4.2-2 各評価方法と内容伝達との相関係数

評価方法	内容伝達との相関係数
評価手法①	0.56
評価手法②	0.41
評価手法③	0.41
BLEU	0.07

各評価手法と内容伝達との相関係数を比べると、BLEU は評価手法①、評価手法②、評価手法③より低い相関となった。

#### 8.4.4.3 大規模言語モデルによる評価結果の分析

大規模言語モデルによる評価結果を人手評価との比較、自動評価との比較を行った。結果、言語別の相関では高い相関は得られなかった(表 8.4.4.1-31)。人手評価と大規模言語モデルとの相関を比べると、人手評価と BLEU の相関よりは高い相関が得られた。これらの結果より、これまでの自動評価結果に比べ大規模言語モデルによる評価は有用性があると考えられる。但し、大規模言語モデルによる評価は評価で使用するモデルの性能に依存する点に注意が必要である。本調査では、モデル Gemini を使用し評価を行ったが、他のモデルを利用する場合は、評価の傾向を改めて確認する必要がある。

#### 8.4.5 大規模言語モデルによる文中タグの翻訳

特許文書では文中の上付き下付き文字、強調、下線などを表現するために XML のタグが用いられる。これら文中のタグについて大規模言語モデルによる翻訳が適切に対応できているかを調査した。調査は原文中にタグ表現 ("**<タグ名>任意の文字列</タグ名>**"で囲まれた文字列)を含む評価文について、機械翻訳文の中に、正解訳と同じタグ表現を含まないデータを対象に分析した。

調査は人手評価文の大規模言語モデルによる翻訳の 3 モデルおよびニューラル機械翻訳

について、正解訳のタグ表現と機械翻訳のタグ表現が一致しているかを確認し、一致しなかった文について一致しない要因別の文数を集計した(表 8.4.5-1)。

表 8.4.5-1 正解訳のタグ表現と一致しなかった要因

翻訳の種類/モデル	タグ欠落	小さな文字	その他	総計	母数 <sup>29</sup>
<b>大規模言語モデルによる翻訳</b>					
Claude	6	7	2	15	60
Gemini	3	7	2	12	60
Gemma	5	26	5	36	60
<b>ニューラル機械翻訳</b>					
nmt <sup>30</sup>	11		9	20	60
<b>総計</b>	<b>25</b>	<b>40</b>	<b>18</b>	<b>83</b>	

タグ表現が一致しなかった文数は、Gemma が総計 36 件と最も多い結果となった。このうち「タグ欠落」は、例えば正解訳のタグ表現「CO<sub>2</sub>」が「CO2」と機械翻訳されたものを指す。「小さな文字」は、タグ表現「1 × 10<sup>14</sup>」がタグ表現でなく小さな文字を使用して「1×10<sup>14</sup>」のように機械翻訳されたものを指す。「その他」はタグ表現の一部が欠落したものや、タグ内の訳が正解訳と異なるもの等が該当する(表 8.4.5-1)。Gemma におけるタグ表現の不一致の要因を見ると、小さな文字で翻訳されている件数が大多数であった。

表 8.4.5-1 タグが一致しない例

正解訳	機械翻訳	備考
一実施形態においては、t <sub>2</sub> は10 μm以上である。	一実施形態では、t<sub>2</sub>は、少なくとも10 μmである。	閉じタグ</sub>の欠落
Provided is an anti-stress composition including <i>Lactobacillus helveticus</i> as an active ingredient.	An anti-stress composition comprising Lactobacillus helveticus </i> as an active ingredient is provided.	開始タグ<i>の欠落
<i>ラクトバチルス・ヘルベ	活性成分として *ラクトバチ	<i>タグをアスタリスク記号

<sup>29</sup> 母数は3項目(要約、請求項、明細書)×4言語方向×5文(タグ表現を含む件数)=60文

<sup>30</sup> nmt 翻訳時に使用したシステムを確認したところ、タグ処理に対応したバージョンでないことが分かり、再度翻訳したところすべての文でタグが正しく翻訳できることを確認した。

ティカス</i>を有効成分とする抗ストレス用組成物を提供する。	ルス・ヘルベティカス*を含む抗ストレス組成物が提供される。	を使い翻訳。
R<sub>8</sub> ~ R<sub>12</sub>が独立して、H、C(<sub>1~4</sub>)アルキル、ハロゲン原子、ヒドロキシル、又は C(<sub>1~3</sub>)アルコキシである、請求項6に記載の化合物。	請求項6に記載の化合物において、R<sub>8</sub>-R<sub>12</sub>が独立してH、C<sub>1-4</sub>アルキル、ハロゲン、ヒドロキシル又は C<sub>1-3</sub>アルコキシである化合物。	タグ表現自体は問題ないが、数値範囲を示す記号が“~”と“-”で相違した。

#### 8.4.6 大規模言語モデルによる定型文の翻訳

本調査事業では、審査書類（拒絶理由通知書、意見書）からの文を評価対象としている。これら審査書類には法的・経緯式的な言い回しや、引用条文・根拠明示が記載された定型文を含むことを特徴としている。このため、審査書類における定型文が、どの程度の精度で翻訳されているかに着目した調査を実施した。

##### 定型文の例

（進歩性）この出願の下記の請求項に係る発明は、その出願前に日本国内又は外国において、頒布された下記の刊行物に記載された発明又は電気通信回線を通じて公衆に利用可能となった発明に基いて、その出願前にその発明の属する技術の分野における通常の知識を有する者が容易に発明をすることができたものであるから、特許法第29条第2項の規定により特許を受けることができない。

拒絶理由通知書、意見書それぞれに含まれる定型文各10文を対象とし、人手評価のうち「内容伝達レベル」を調査した<sup>31</sup>(表 8.4.6-1)。

表 8.4.6-1 定型文評価結果（内容伝達）

翻訳の種類/モデル	拒絶理由			意見書	
	5	4	3	5	4
大規模言語モデルによる翻訳					

<sup>31</sup> 補正書の評価文には定型文が存在しなかったため、調査対象外とした。

Claude	7	3		9	1
Gemini	9	1		10	
Gemma	5	4	1	8	2
前後処理を適用した翻訳					
前後処理②					
Claude	9	1		9	1
追加学習済大規模言語モデル					
f1500_gemma	4	3	3	7	3
ニューラル機械翻訳					
nmt	9	1		9	1

調査対象件数が少数ではあるが、評価結果から大規模言語モデルによる翻訳では Gemini、Claude、Gemma の順に高い評価となった。Claude については前後処理②を適用した翻訳でベースモデルに比べ拒絶理由の内容伝達レベル 5 が 7 文から 9 文に増加し精度向上が見られた。一方 Gemma に追加学習を施したモデル f1500\_gemma は拒絶理由の内容伝達レベル 5 の文数が 5 から 4 に減少、意見書の内容伝達レベル 5 の文数も 8 から 7 に減少しており、精度の低下が見られた。ニューラル機械翻訳は前後処理②と同等の結果となった。内容伝達の値が低かった文の例を以下に示す。

原文	内容伝達	この拒絶理由通知は、本願と出願人が同一である未確定の先願(出願審査未請求のものも含む)に基づき特許法第 39 条第 1 項の規定に基づいて拒絶理由を通知するものである。
正解文		This is a notification of reasons for refusal to notify the reasons for refusal under the provision of Patent Act Article 39(1), based on the pending prior application (including the one for which a request for examination has not been filed) whose applicant is the same as the present application.
Gemma	3	This rejection notice is based on a pending earlier application (including one for which a request for examination has not been filed) that is identical to this application and notifies the grounds for rejection pursuant to Article 39(1) of the Patent Act.
備考		Gemma は、原文「出願人」部分の訳抜け。

原文	内容	(明確性) この出願は、特許請求の範囲の記載が下記の点で、特許法第 36 条第 6 項第 2 号に規定する要件を満たしていない。
----	----	--

	伝達	
正解文		(Clarity) This application does not satisfy the requirement as provided in Article 36(6)(ii) of the Patent Act since the claims are deemed to be defective in the following.
f1500_gemma	3	Clarity) This application does not meet the requirements of Article 36(6)(2) of the Patent Act in that <u>the claims are not described in the following respect.</u>
備考		the claims are not described in the following respect は「特許請求の範囲について、下記の点に関して（特許法に規定する要件を）満たしていない」との原文の意味から離れて訳されている。

## 9. その他の評価

### 9.1 翻訳に要する時間

#### 9.1.1 1文あたりの翻訳時間

翻訳の実施において翻訳に要した時間やモデルのライセンス費用、その他、翻訳精度以外の観点について整理し、各モデル及びニューラル機械翻訳との比較評価を行った。

翻訳に要した時間は、各モデルによる翻訳（7章）の実施に要した各文の翻訳時間を集計した。モデル Gemma は、下記オンプレミス環境に示すスペックのマシンでの翻訳時間を集計した。前後処理を適用した翻訳では、各方式の前後処理自体<sup>32</sup>の時間と翻訳時間を合算して集計した(表 9.1-1)。Claude, Gemini の処理時間は LLM モデル提供元の混雑状況等により翻訳時間が異なる点に注意が必要である。大規模言語モデルによる翻訳では、Gemini、Claude、Gemma の順に処理時間を要する結果となった。Gemini と Claude の翻訳時間を比較すると、Gemini は Claude の 5.5 倍以上翻訳時間を要した。前後処理を適用した翻訳では、前後処理①、②が⑤に比べ多くの時間を要する結果となった。プロンプトを工夫した翻訳では、prompt4, 8, 9 が処理時間を多く要する結果となった。特に prompt8 は、1 文の平均が 50 秒を超える結果となった。追加学習済大規模言語モデルは、学習前のモデルに比べ 2.5 倍程度多く時間を要する結果となった。ニューラル機械翻訳は、1 文の平均が 0.02 秒となり、大規模言語モデルによる翻訳に比べ大幅に少ない翻訳時間となった。

#### オンプレミス環境

GPU
NVIDIA Ada L40S
GPU memory 48GB
CPU
Intel Xeon Gold 6416H
Memory
1,024GB

<sup>32</sup> 前後処理時間は、前後処理①は、重要用語抽出とコーパス検索時間、を前後処理②は、用例作成時間、前後処理⑤は、類似文検索時間が該当する。

表 9.1-1 翻訳に要する時間  
(1文あたりの平均時間(秒))

	発明の名 称	要約	請求 項	明細 書	拒絶理 由	意見 書	補正書	出願人名/ 発明者名	総計	
<b>大規模言語モデルによる 翻訳</b>										
	Claude	2.27	2.96	3.24	2.78	2.86	2.89	2.85	2.26	2.79
			17.1							
	Gemini	10.02	5	21.92	15.50	14.79	14.14	15.72	7.42	15.52
	Gemma	0.77	1.98	2.56	1.70	1.71	1.79	1.97	0.66	1.71
<b>前後処理を適用した翻訳</b>										
<b>前後処理①</b>										
	Claude		26.26	23.67						24.96
<b>前後処理②</b>										
			18.8							
	Claude	17.02	9	23.15	20.34	19.50	19.78	19.63	12.76	18.88
<b>前後処理⑤</b>										
	Claude		6.93	6.24						6.58
<b>プロンプトを工夫した翻 訳</b>										
<b>prompt1</b>										
	Claude	2.22	4.24	5.20	3.87	3.73	3.80	4.14		3.92
<b>prompt2</b>										
	Claude	2.30								2.30
<b>prompt3</b>										
	Claude			5.08						5.08
<b>prompt4</b>										
	Claude			24.83						24.83
<b>prompt8</b>										
	Claude			51.24						51.24
<b>prompt9</b>										
	Claude			22.75						22.75
<b>追加学習済大規模言語モ デル</b>										

f1500_gemma	1.71	5.18	6.89	4.36	4.64	4.67	5.29	1.13	4.41
f2500_gemma			5.18	4.48					4.83
<b>ニューラル機械翻訳</b>									
nmt	0.01	0.03	0.03	0.02	0.03	0.04	0.05	0.01	0.02

### 9.1.2 1年間あたりの翻訳時間

本項では英中韓それぞれ1年分の公報翻訳に係る時間を机上計算する。  
具体的な算出方法は、以下の算出式を用いた。

・翻訳時間の算出式

1年あたりの翻訳対象公報数 × 1公報に含まれる平均文数（タイトル、要約、発明の詳細な説明、特許請求の範囲の和） × 1文あたりの翻訳時間（9.1.1での「総計」を利用） ÷ 並列数（8とした）

・1年あたりの翻訳対象公報数：2025年1月1日～12月31日の公報件数（当社保有のDBから件数を取得）

言語	調査対象国	対象公報	公報の件数
英語	米国	公開特許公報、登録特許公報	772,239 件
中国語	中国	公開特許公報、公告・登録特許公報、公告・登録実用新案公報	4,440,982 件
韓国語	韓国	公開特許公報、公告・登録特許公報、公開実用新案公報、公告・登録実用新案公報	340,777 件

・1公報に含まれる平均文数：

2024年7月～9月の各国公開特許公報から算出。

調査対象国	1公報あたりの文
米国	516.2 文
中国	508.1 文
韓国	504.3 文

8 並列処理をした際の、1年間分の各国公報の翻訳時間概算値（日数）は以下となった（表 9.1-2）。

表 9.1-2 1年分の各国公報の翻訳時間概算値

	米国文献翻訳日数	中国文献翻訳日数	韓国文献翻訳日数
<b>大規模言語モデルによる翻訳</b>			
Claude	1609	9108	694
Gemini	8951	50666	3859
Gemma	986	5582	425
<b>前後処理を適用した翻訳</b>			
<b>前後処理①</b>			
Claude	14395	81483	6206
<b>前後処理②</b>			
Claude	10888	61635	4694
<b>前後処理⑤</b>			
Claude	3795	21481	1636
<b>プロンプトを工夫した翻訳</b>			
<b>prompt1</b>			
Claude	2261	12797	975
<b>prompt2</b>			
Claude	1326	7508	572
<b>prompt3</b>			
Claude	2930	16584	1263
<b>prompt4</b>			
Claude	14320	81059	6174
<b>prompt8</b>			
Claude	29551	167276	12740
<b>prompt9</b>			
Claude	13120	74269	5656
<b>追加学習済大規模言語モデル</b>			
f1500_gemma	2543	14397	1096
f2500_gemma	2786	15768	1201
<b>ニューラル機械翻訳</b>			
nmt	12	65	5

以上のとおり、各 LLM モデルとも、特許公報の翻訳を 1 文単位で行うと莫大な時間がかかり現実的でない。例えば、Claude で外国特許公報一年分を 8 並列処理で翻訳したとしても、1609 日（米国）、9108 日（中国）、694 日（韓国）を要する計算となる。

以下参考であるが、LLM による特許公報の翻訳を文書単位で行う場合についても補足する。この場合、例えば、米国公報（年間 772,239 件）の翻訳について、1 文献 3 分と仮定（平均的な文数の公報翻訳を数件試行した結果から仮定）した場合、 $3 \text{分} \times 772,239 \text{件} \div 1440 \text{分} (1 \text{日}) \div 8 (\text{並列数}) = 201.1 \text{日}$ となる。ただし、長文である特許文献の読み込みに対応しているモデルが限定される点や、翻訳を指示したとしても要約処理をしてしまうケースがあることなど、ドキュメント単位の翻訳には品質面での問題が残る<sup>33</sup>。

## 9.2 翻訳に要する費用

### 9.2.1 100 万文字あたりの費用

大規模言語モデルによる翻訳の明細書の英日、中日、韓日、日英を対象に、翻訳に要した費用を計算した(表 9.2-1)。具体的には、対象とするモデルは、入出力トークン数に応じた課金がされる Claude 及び Gemini とし、LLM への翻訳指示データ(原文を含むプロンプト)のトークン数(入力トークン)と翻訳結果のトークン数(出力トークン)を集計し、各モデルのトークン数あたりの単価を元に費用を計算した。各モデルの単価は以下となる。また、ニューラル機械翻訳の費用は、当社が一般ユーザー向けにサービス展開しているニューラル機械翻訳の料金(100万文字/月)を用いた<sup>34</sup>。Claude と Gemini の費用を比較すると、Gemini の料金 > Claude の料金となった。Gemini の費用が高くなったのは、Gemini の出力トークン数が多くなったためとなる。各モデルの原文 100 万文字あたりの費用を算出し、ニューラル機械翻訳の費用と比べたところ、費用の高い順から、Gemini > ニューラル機械翻訳 > Claude の順番になった。ただし、ニューラル機械翻訳の料金をはじめとする各社の料金体系は、大量の文献翻訳を行う場合は別の料金体系となることがあるので注意が必要である。

Claude 利用単価
-------------

入力: 3 USD / 100 万トークン
-----------------------

<sup>33</sup> Investigating Length Issues in Document-level Machine Translation Ziqian Peng Mon, 28 Apr 2025 13:52:53 UTC

<sup>34</sup> [https://japio.or.jp/service/files/price\\_AI.pdf](https://japio.or.jp/service/files/price_AI.pdf)

出力: 15 USD / 100 万トークン
Gemini 利用単価
入力: 7.50 USD / 100 万トークン
出力: 15 USD / 100 万トークン
米国ドル/円レート
155 円/USD

表 9.2-1 大規模言語モデルによる翻訳費用

言語/モデル	入力トークン数	出力トークン数	入出力トークン料金 (円)	原文文字数	原文 100 万文字あたりの料金 (円)
Claude					
cj	50,183	34,636	104	25,873	4,014
ej	43,248	34,587	101	87,456	1,149
je	57,285	23,253	81	34,438	2,343
kj	64,950	36,670	115	41,562	2,778
Gemini					
cj	39,983	1,031,936	2,446	25,873	94,528
ej	39,913	1,079,614	2,557	87,456	29,232
je	42,098	856,201	2,040	34,438	59,225
kj	46,828	871,956	2,082	41,562	50,087

### 9.2.2 1 年分あたりの翻訳費用

本項では英中韓それぞれ 1 年分の公報翻訳に係る費用を机上計算する(表 9.2-2)。具体的な算出方法は、以下の算出式を用いた。

#### ・翻訳費用の算出式

1 年あたりの翻訳対象公報数 × 1 公報に含まれる平均文字数 (タイトル、要約、発明の詳細な説明、特許請求の範囲の和) ÷ 100 万 × 100 万文字あたりの翻訳費用

(9.1.2 での「原文 100 万文字あたりの料金」を利用)

- ・ 1 公報に含まれる平均文字数：  
2024 年 7 月～9 月の各国公開特許公報から算出。

表 9.2-2 1 年分あたりの翻訳費用

言語	調査対象国	公報の件数	1 文献あたり文字数	Claude	Gemini
英語	米国	772,239 件	99209.1	¥88,028,487	¥2,239,555,038
中国語	中国	4,440,982 件	29728.7	¥529,946,614	¥12,480,018,320
韓国語	韓国	340,777 件	47415	¥44,886,776	¥809,303,075

## 10 まとめと提言

本事業では、国内外で提供される大規模言語モデルの特性や翻訳への活用手法を調査するとともに、大規模言語モデルを用いて特許文献及び審査書類の機械翻訳を実施し、その精度を評価、分析した。

調査結果については「2.」から「5.」において、大規模言語モデル、前後処理、プロンプト、および翻訳精度評価手法の各観点から整理して記述した。また、「8.」では翻訳結果の評価を、「9.」ではその他の観点からの評価を行い、それぞれの結果を報告した。本項では、特許庁における将来的な機械翻訳システムのあり方を検討するためにまとめと提言を行う。

### 10.1 まとめ1：LLMを用いた翻訳とNMTを用いた翻訳との比較

将来的な機械翻訳システムのあり方を検討するにあたり、これまでの調査・分析結果をもとに、LLMを用いた翻訳と従来のNMTを用いた翻訳とを対比した。(表10-1)。

表 10-1 LLMを用いた翻訳とNMTを用いた翻訳の比較

観点	LLMを用いた翻訳	NMTを用いた翻訳	備考
文全体の翻訳品質	◎～○	○	ほぼ同等の品質(LLMのGemmaを除く)であるが、微差でLLMの方がNMTより翻訳の品質は高い 項番 8.4.1
技術用語/特許用語の訳質	○	○	LLM、NMTともほぼ同等の品質(項番 8.2)、両者とも一定数の誤訳が見られたことから、共通の課題と考えられる。
技術用語/特許用語の訳ゆれ	○	○	LLMのほうが若干訳ゆれは少ないがほぼ同等の品質

長文への対応	◎	○	公開情報調査(項番 2)より、LLM コンテキスト長が大きいことを確認。コンテキスト長が大きいことが、翻訳品質に寄与するかは、引き続き調査が必要。NMT は制限を超える長文は翻訳の前処理、後処理で文の分割、結合で対応することが多い。
タグの保持	○	◎	LLM は一部、タグが保持できないものが見られた、NMT は翻訳の前後処理の対応によりタグの保持がされていた。
カスタマイズ性	◎	○	LLM, NMT とも追加学習によるカスタマイズを備える。LLM は、さらに本調査の前後処理(項番 7.2)や プロンプトの工夫(項番 7.3)で使ったようなプロンプトによりカスタマイズ可能。本調査では中日の前後処理①やプロンプト②で訳語改善が見られた。
費用	○	○	ボリュームディスカウントがなければ、費用の観点では LLM, NMT とも同程度(項番 9)。
処理速度	△	◎	処理速度の観点では NMT が圧倒的に優れている(項番 9)。

## 10.2 まとめ 2 : LLM を用いた評価と自動評価の比較

本調査では、大規模言語モデルによる評価を調査した。既存の自動評価手法と比較を以下に示す(表 10-2)。

表 10-2 大規模言語モデルによる評価と既存の自動評価との比較

	LLM による評価	既存の自動評価 (BLEU, RIBES)	備考
翻訳品質評価	○	△	どちらも人手評価との間で大きな相関は得られなかったが、LLM 評価の方が既存の自動評価に比べ、人手評価との相関係数は高かった。LLM 評価では、文の単位で評価が得られる点、評価結果に対する理由が得られることに特徴がある。

### 10.3 将来的な機械翻訳システムについて

表 10-1 に記した大規模言語モデルによる翻訳と、NMT による翻訳の特徴を踏まえ、将来的な機械翻訳システムのありかたの一案を検討した。

<検討要素>

(1) 文全体の翻訳品質の向上を行うこと：人手評価の結果から、現状においても微差ではあるが LLM の方が NMT より翻訳の品質は高いといえる。将来的には、LLM による翻訳の品質はさらに向上する可能性が高い。特に人名といった固有名詞に対する翻訳能力は語彙の大きさから、LLM に軍配が上がる。LLM を利用することで文全体の翻訳品質向上を図れる。

(2) 長文への対応性向上：LLM のコンテキスト長は非常に大きく、長文への対応が期待できる。LLM を利用することで特許文献で多用される長文への適応力を高められる。なお、LLM を利用した翻訳は、文単位翻訳でなく、段落単位、また、文字数の少ない文献であれば、文献単位の翻訳が可能となる可能性があり、さらに同一語の揺れへの耐性向上も見込める。

(3) カスタマイズ性の活用：LLM による翻訳にあたり、前後処理処理や、プロンプトの工夫により、訳質に改善の余地が大きいものについては一定の効果がある。

(4) 特許文書タグの保持：LLM を用いた翻訳を行う場合、タグを保持するための処理が必要である。具体的には比較対象とした NMT のような前後処理、プロンプトの工夫等の手当てが有効。

(5) 処理速度：本事業による調査の結果では、NMT は LLM3 モデルで最も高速であっ

た Gemma と比べても 1 文あたりの翻訳速度は 85.5 倍であった（1 文翻訳時間 総計ベースで Gemma は 1.71 秒に対し、NMT は 0.02 秒）。NMT モデルが翻訳に専用に最適化された小型モデルといえる一方、大規模言語モデルは文章生成のための汎用大型モデルであることから、処理速度の問題は直ちに改善することは難しいと考えられる。特にパラメータ数が 100Billion を超えるクローズドモデルとなればその傾向は著しく、その語彙の多さに起因した訳語の確率分布計算の負荷は原理的に NMT の比ではないといえる。

（6）機密性：機密性の高い原文の翻訳が要件である場合は、クローズドモデルでなく、オープンモデルの利用が望ましい。

※費用について：9.2.2 で述べたとおり、ボリュームディスカウント等を考慮せず、LLM API の利用料金を基に各国公報の翻訳費用（1 年分）を単純に試算すると、米国・中国・韓国の合計で年間 6 億円を超える。この点は NMT を用いた場合（ボリュームディスカウント未考慮）においても同様であるため、本章では費用面を提言の観点には含めていない。

なお、外国特許文献を翻訳する事業である「外国特許文献の日本語機械翻訳データ作成及び機械翻訳精度向上に関する調査事業」は、英語・中国語・韓国語の合計 2919 万件を 4 年間で NMT を用いて翻訳する事業であり、契約金額は合計 4.3 億円である。この金額は、LLM による各国特許公報の翻訳費用（米中韓、年間約 155.4 万件を Claude Sonnet を用いて翻訳すると年間 6.62 億円、Gemini pro を用いて翻訳すると年間 155.29 億円）と比較すると、非常に低コストである。

大量翻訳用途の機械翻訳システムは、訳質はもちろんのこと、限られた時間内に大量の特許公報の翻訳を行う必要があるため、翻訳速度が求められるといった特性を有している。当該特性を勘案すると、＜要素事項＞のうち、特に（5）の点から大量翻訳を LLM 翻訳のみで行うことは現実的ではない。訳質の向上と翻訳速度の担保のために以下の手法を提案する。

（a）ハイブリッド翻訳：NMT 翻訳を中心に翻訳の改善が必要な文について LLM 翻訳を補うことで、処理速度が速いという NMT 翻訳のメリット（＜要素事項＞（5））と、LLM 翻訳のメリット（＜要素事項＞（1）、（2））の両方を活かすことができる。

特に、LLM 翻訳はカスタマイズが比較的容易であることから（＜要素事項＞（3））、基本的なプロンプトでは翻訳改善を図ることができないような文についても比較的容易に翻訳結果を調整できる。

（b）LLM 翻訳の処理速度向上策 1：段落単位、ドキュメント単位の翻訳といったように翻訳対象文書を長文化することで、文単位で翻訳する際のオーバーヘッドの改善を試みる（＜要素事項＞（2）を考慮）。

（c）LLM 翻訳の処理速度向上策 2：パラメータ数が小さいオープンモデル LLM をベー

スに大量のコーパスで学習することで、LLM のコンテキスト長の長さを活かしつつも高速なモデルを作成する（＜要素事項＞（5）、（6）を考慮）。

また、リアルタイム翻訳においても、翻訳速度が求められることは同様であるが、UX 設計の標準指標として引用されることが多い「Usability Engineering」（著者 Jakob Nielsen）の応答時間の指標<sup>35</sup>によると Website のレスポンスにおいては 10 秒以内のフィードバックを行うのが望ましく、リアルタイム翻訳結果を逐次応答して画面表示するといったシステムの作り込みが求められる。

## 10.4 大規模言語モデル等による評価の活用

表 10-2 のとおり、従来の機械評価よりも LLM を用いた評価の方が人手評価との相関性が高い。下記の例は、人手評価の調査で用語の誤訳例として例示されたもの（表 10-3）と、この文について大規模言語モデルによる評価結果に付されたコメントである（表 10-4）。

表 10-3 人手評価で指摘された用語に関する誤訳の例

原文	麦角硫因、抗坏血酸 2-葡糖苷、抗坏血酸和它们的组合的成肌分化促进作用
正解訳文	エルゴチオネイン、アスコルビン酸 2-グルコシド、アスコルビン酸、およびそれらの組み合わせの筋分化促進作用
人手内容伝達 レベル：1	麦角硫因、抗坏血酸 2-グルコシド、抗坏血酸およびそれらの組み合わせの筋分化促進作用

表 10-4 大規模言語モデルによる評価結果

評価手法および評価観点	評価結果に付されたコメント
-------------	---------------

<sup>35</sup> <https://www.nngroup.com/articles/website-response-times/>

<p>評価手法① 語の誤訳 major</p>	<p>The translation uses the Chinese kanji for chemical compounds instead of the standard Japanese katakana terms.</p> <p>'<b>麦角硫因</b>' should be '<b>エルゴチオネイン</b>', and '<b>抗坏血酸</b>' should be '<b>アスコルビン酸</b>'.</p> <p>In a scientific or technical context, using the correct, standard terminology is crucial for clarity and accuracy.</p>
<p>評価手法②正確性（用語）スコア 3</p>	<p>(Acceptable translation): The translation uses Chinese Kanji terms such as 「<b>麦角硫因</b>」 and 「<b>抗坏血酸</b>」 instead of the standard Japanese terms 「<b>エルゴチオネイン</b>」 and 「<b>アスコルビン酸</b>」. While the meaning is understandable, these are not the commonly used technical terms in Japanese patent documents. The translation for 「成肌分化促進作用」 as 「筋分化促進作用」 is perfectly suitable.</p>
<p>評価手法③ スコア 45</p>	<p>This translation has significant terminology errors. The key chemical compounds "<b>麦角硫因</b>" (ergothioneine) and "<b>抗坏血酸</b>" (ascorbic acid) have been left in their Chinese Kanji form instead of being translated into the standard Japanese Katakana terms "<b>エルゴチオネイン</b>" and "<b>アスコルビン酸</b>" respectively. In a patent document, using correct and standard terminology is critical. While the grammatical structure and the latter half of the phrase ("... およびそれらの組み合わせの筋分化促進作用") are correct, the failure to translate the main subjects of the phrase is a major flaw that severely impacts the quality and professionalism of the translation. The meaning is partially obscured and relies on the reader's knowledge of Chinese.</p>

この例では、大規模言語モデルの評価結果を見ると各手法で誤訳の検出ができていたことが分かる。特に評価手法①はその評価の際に正訳を使用せず評価を行っている。

このように、LLM 評価では従来の機械評価と異なり、誤訳箇所といった、具体的な評価の理由を得ることができる。ただし LLM 評価には、評価に用いるモデルを変更すると結果が変わる、再現性が不安定といったデメリットがある。

また、従来の機械評価、例えば BLEU では log+幾何平均による圧縮により値を算出す

ることから、スコアが高くなるにつれ差がみえにくくなる特徴があり、近年の訳質が高い翻訳文どうしの評価手法としては限界を迎えつつあるといえる。

以上から、今後の特許翻訳文の訳質評価の仕組みとして、従来の機械評価に代わる以下の手法を提案する。

(a) LLM 評価の採用及び特許文章の特徴を考慮した評価ルールの検討：特許文特有の表現（請求項の末尾の体言止め等）を前提とした評価の仕組みの検討を行う。

(b) 特許ドメイン特化の評価手法の構築：例えば、ニューラルネットワーク（主に多言語 Transformer）を使った学習型翻訳評価指標である COMET<sup>36</sup>は Unbabel 社によって開発された評価の仕組みであり、原文、機械翻訳文、参照訳、人手評価を入力することで「人間評価とどれだけ相関するか」を学習するものである。LLM 評価の不安定さ及び BLEU といった従来の機械評価の限界を補完すべく、ニューラルネットワークを使った学習翻訳評価指標を特許ドメインに特化させた評価モデルの検討を行う。

---

<sup>36</sup> <https://unbabel.github.io/COMET/html/index.html>