日本語の文書レベルのテキスト平易化の自動評価

山下伊織　木山朔　Bian Kexin　小町守
一橋大学
{iori, hajime, bian, komachi}@scl.sds.hit-u.ac.jp

概要

本研究は，日本語の文書レベルにおける，テキスト平易化の大規模言語モデル（LLM）の自動評価能力を検証するものである．データセットとして JADOS-eval [1] を用い，ソース文書に対応する複数の平易化文書について，必要性・十分性・文平易度・文書平易度の 4 指標を対象に，人手評価と LLM の分類結果を比較した．その結果，GPT-4o と Gemini 2.5 Pro は 0-shot でも高い一致率を示し，従来指標 D-SARI [2] より人手評価との相関が高い傾向が確認された．一方，多くのモデルは文書平易度を体系的に過小評価し，特に漢字出現率など表層的難度が判断に影響する可能性が示された．

1　はじめに

テキスト平易化は文単位の研究が中心であったが [3]，文レベルの平易化では文書全体の平易度が必ずしも向上しないことが指摘されてきた [4]．一方，文書レベルテキスト平易化では，情報の取捨選択や段落間の結束性の保持といった，単文処理とは異なる課題が生じる [2, 5]．こうした背景を受け，近年は SWiPE [6] や SIMSUM [7] など英語を中心に文書レベルテキスト平易化のための資源が整備されてきており，日本語においても JADOS [8] やそれに平易度に関するメタデータを付与した JADOS-eval [1] の公開によって基盤が整いつつある．

しかし，文書レベルテキスト平易化の「自動評価」に着目した研究は依然として少なく，文単位指標 SARI [9] を文書レベルに拡張した D-SARI も談話的一貫性の評価には十分ではないことが報告されている [2]．加えて，近年は LLM を評価器とする LLM-as-a-Judge が注目されているものの [10]，その評価はモデルや言語に依存して不安定になりうることが指摘されている [11, 12]．これらの経緯から，本研究では日本語文書レベルテキスト平易化において，人手評価と LLM 評価を体系的に比較し，

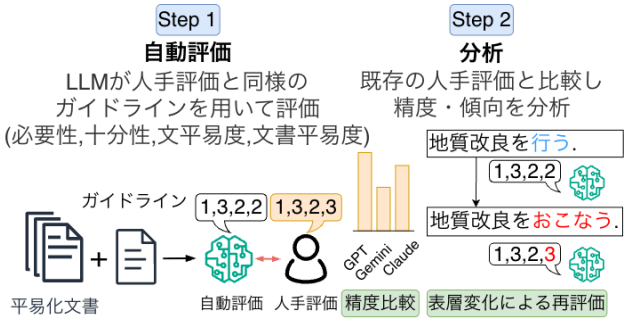

図 1: 日本語における文書レベルのテキスト平易化を対象とした人手評価と自動評価の比較の概要．

LLM-as-a-Judge による自動評価の有効性を検討する（表 1）．また，従来の評価指標である D-SARI との比較を行い，LLM-as-a-Judge の有効性を調査する．

2　深層学習による分類評価

本研究では，日本語の文書レベルのテキスト平易化の自動評価として LLM-as-a-Judge およびチューニングしたモデルを用いて人手評価と同様の必要性・十分性・文平易度・文書平易度の 4 指標を予測する．

手法 1: LLM-as-a-Judge（GPT 系モデル）　ソース文書とそれを平易化したターゲット文書を入力とし，人手評価と同様に付録表 4 に示す評価ガイドラインに基づいて，必要性・十分性・文平易度・文書平易度の 4 指標の離散値スコアを LLM に直接出力させる．出力は各評価指標をキーとする JSON 辞書形式で返すように日本語で指示した．この出力を自動評価結果とみなし，人手ラベルとの一致率を比較することで，LLM-as-a-Judge の有効性を検証する．

評価設定として，0-shot，1-shot（ソース文書 1 件と，対応する平易化文書 1 件を例示），Few-shot（ソース文書 1 件と，対応する平易化文書 6 件を例示）の 3 条件を設定する．1-shot および Few-shot の条件では，例示に用いない 187 件のソース文書と 1,122 件の平易化文書を評価対象とする．

さらに，通常の単一出力による評価 (Baseline) に加え，プロンプト設計の違いによる影響を分析する

ため，以下を検討する．

- Majority: 出力の揺れを抑制するため，各入力に対して 5 回評価を生成し，最頻値を最終判定とする多数決方式の設定
- Reasoning: 出力に理由づけを含めさせることで，推論過程を明示的に促す設定
- English: プロンプトを英語に切り替え，指示言語の違いが与える影響を分析する設定 [11]. [1)]

手法 2: 教師あり分類（BERT 系モデル） 比較対象として，ソース文書と平易化文書の組を入力とし，各指標のスコアをあらかじめ定義した離散的なクラスに分類する分類モデルを学習する．生成モデルに依存しない判別器としての性能を測定し，LLM-as-a-Judge と対比する．

3 実験設定

3.1 データセット

本研究では，日本語文書平易化コーパス JADOS [8] に対し，人手評価を付与した JADOS-eval [1] を用いた．本データセットは，日本語文書平易化コーパス JADOS [8] のうち Wikipedia の「良質な記事」「秀逸な記事」に分類される記事の概要部分から作成された validation セットのうち 188 件をソース文書として選定し，人手によって作成されたターゲット文書に加え，LLM 等 [2)] を用いて作成された 5 通りのターゲット文書が収録されており，5 名の日本語母語話者（大学生・大学院生）がガイドライン（付録表 4）に基づいて必要性・十分性・文平易度・文書平易度の 4 指標で人手によるラベルが付与されている．

3.2 モデル

GPT 系モデル 本研究では，GPT-4o [3)]，Claude 4 Sonnet [4)]，および Gemini 2.5 Pro [5)] を用いた．推論パラメータは共通して temperature = 0.2, max_tokens = 512（その他はデフォルト設定）とした．

BERT 系モデル 日本語 BERT 系モデルとして `sbintuitions/modernbert-ja-130m` [6)] および `cl-tohoku/bert-base-japanese-v3` [7)] を採用した．学習およびテストには 10-fold cross-validation を採用し train:dev:test=8:1:1 で分割，エポック数は 20，バッチサイズは 16，学習率は 3×10^{-5} とした．

3.3 評価方法

自動評価 それぞれの指標ごとに JADOS-eval [1] の人手評価の最頻値を正解ラベルとし，モデルによる出力の一致率 (Accuracy) を次式で算出した．[8)]

$$\text{Accuracy} = \frac{\text{正解と一致した評価数}}{\text{総評価数}}$$

メタ評価 Accuracy による比較に加え，人手作成を除いた 5 モデルによるターゲット文書に対し，既存の評価指標である D-SARI と人手評価との相関，および ModernBERT，GPT-4o 0-shot (Baseline) 設定の LLM-as-a-Judge による文書平易度の出力と人手評価との相関をそれぞれ算出し，両者を比較した．

4 実験結果

4.1 評価

自動評価 表 1 に LLM による自動評価の結果を示す．[9)] GPT および Gemini は，0-shot 設定でも 70%程度の一致率を示すことがわかった．対して，Claude は 0-shot では文平易度・文書平易度において顕著に低い一致率を示し，タスクに適さない傾向が見られた．また shot 数の変化による一致率への影響については，GPT，Gemini は変化があまり見られず，Claude では文平易度および文書平易度で改善が見られたが，他の GPT 系と同程度であった．BERT 系モデルは学習データによるファインチューニングによって十分性を除く 3 指標では GPT 系とほぼ同程度の一致率を示したが，十分性の一致率はやや GPT 系に劣る結果となった．

プロンプト変更による一致率の変化を見ると（表 1），GPT 系は一貫して，多数決および理由づけ

1) 1-shot と Few-shot で与える例示は日本語のままとした．
2) Enc-Dec モデルによる出力は bart-large-japanese を微調整したモデルを，Dec-only 型 LLM 出力として Llama-3.1-Swallow-8B-Instruct-v0.2 と gemma-2-9b-it，GPT-4o-2024-11-20 の 0-shot と 1-shot を用いて作成された．
3) gpt-4o-2024-08-06
4) claude-sonnet-4-20250514-v
5) 2025.06.17 release
6) https://huggingface.co/sbintuitions/modernbert-ja-130m
7) https://huggingface.co/tohoku-nlp/bert-base-japanese-v3
8) データセットの評価ラベルの粒度が粗く，ラベル分布に偏りがあるため，評価指標として相関係数ではなく Accuracy を採用した．
9) Gemini 2.5 Pro の 1-shot Reasoning，1-shot English でソース文書 1 件，Few-shot Reasoning で 2 件が PROHIBITED CONTENT を理由にレスポンスが拒否されたため，評価から除いた．

例示数	設定	必要性	十分性	文平易度	文書平易度
0-shot	Baseline	0.640	0.714	0.742	0.844
	Majority	0.642	0.723	0.735	0.846
	Reasoning	0.707	0.723	0.730	0.826
	English	0.781	0.721	0.681	0.915
1-shot	Baseline	0.750	0.694	0.750	0.865
	Majority	0.749	0.693	0.751	0.867
	Reasoning	0.758	0.675	0.709	0.889
	English	0.780	0.690	0.665	**0.924**
Fewshot	Baseline	**0.856**	0.397	**0.771**	0.773
	Majority	0.756	**0.725**	0.734	0.879
	Reasoning	0.777	0.711	0.725	0.857
	English	0.802	0.692	0.650	0.922

(a) GPT-4o

例示数	設定	必要性	十分性	文平易度	文書平易度
0-shot	Baseline	0.836	0.646	0.381	0.174
	Majority	0.840	0.633	0.379	0.176
	Reasoning	0.840	0.627	0.666	0.412
	English	**0.848**	0.573	0.707	0.481
1-shot	Baseline	0.811	**0.683**	0.730	0.548
	Majority	0.814	0.682	0.734	0.552
	Reasoning	0.824	0.663	0.703	0.487
	English	0.725	0.619	0.725	0.689
Fewshot	Baseline	0.769	0.607	0.788	0.589
	Majority	0.770	0.610	0.791	0.588
	Reasoning	0.783	0.524	**0.840**	0.717
	English	0.832	0.635	0.821	**0.751**

(b) Claude 4 Sonnet

例示数	設定	必要性	十分性	文平易度	文書平易度
0-shot	Baseline	0.697	0.676	**0.822**	0.699
	Majority	0.688	0.689	0.809	0.691
	Reasoning	0.757	0.652	0.810	0.711
	English	**0.850**	0.692	0.810	0.755
1-shot	Baseline	0.664	0.671	0.814	0.742
	Majority	0.644	0.677	0.820	0.733
	Reasoning	0.658	0.682	0.816	0.731
	English	0.802	0.677	0.812	**0.815**
Fewshot	Baseline	0.747	0.693	0.797	0.708
	Majority	0.743	0.693	0.801	0.689
	Reasoning	0.751	**0.704**	0.794	0.684
	English	0.824	0.688	0.787	0.758

(c) Gemini 2.5 Pro

モデル	必要性	十分性	文平易度	文書平易度
ModernBERT	0.690	0.547	0.637	**0.835**
BERT-base	**0.730**	**0.614**	**0.729**	0.814

(d) 教師あり分類（BERT 系）

表 1: 自動評価結果．各モデル内で各指標の一致率が最も高かったものを太字で示す．

による一致率改善は見られなかった．対して，英語プロンプトによる自動評価では文書平易度において顕著な改善を示した．これは，日本語の文書レベルのテキスト平易化の評価ではプロンプトの指示文を事前に翻訳することで精度が向上する事前翻訳が有効であることを示唆している [11]．

対象データ	評価項目	D-SARI	BERT	GPT
BART	必要性	**0.096**	0.087	0.078
	十分性	0.019	0.302	**0.513**
	文平易度	-0.032	0.054	**0.205**
	文書平易度	0.058	0.212	**0.218**
Llama	必要性	-0.109	0.077	**0.156**
	十分性	-0.195	0.042	**0.371**
	文平易度	0.020	0.279	**0.325**
	文書平易度	0.039	**0.336**	0.320
Gemma	必要性	-0.130	0.112	**0.403**
	十分性	-0.083	0.111	**0.324**
	文平易度	0.002	-0.049	**0.207**
	文書平易度	0.010	-0.069	**0.511**
GPT-4o (0-shot)	必要性	-0.067	-0.009	**0.146**
	十分性	**-0.165**	0.117	0.069
	文平易度	**0.096**	0.038	–
	文書平易度	-0.012	**-0.029**	–
GPT-4o (1-shot)	必要性	-0.076	0.038	**0.161**
	十分性	**-0.235**	0.114	0.202
	文平易度	-0.081	**0.094**	–
	文書平易度	–	–	–

表 2: 人手評価と 3 種の自動指標（D-SARI，ModernBERT による評価，および GPT-4o 0-shot による LLM-as-a-Judge）のスピアマン相関 ρ．各行で絶対値が最大の相関係数を太字で示す．“–” は分散が 0 のため相関が未定義のケース．

メタ評価　表 2 は，人手評価と 3 種の自動指標（D-SARI と ModernBERT，GPT-4o 0-shot (Baseline) 設定の LLM-as-a-Judge）のスピアマン相関を示す．D-SARI は多くのモデル・基準で相関が低い一方，LLM-as-a-Judge は概して高い正の相関を示した．この結果は，参照との一致に基づく評価する D-SARI と比べて，LLM-as-a-Judge の方が高い性能を持つ傾向を示唆する．一方で，十分性に関しては他 2 種が正の相関を示すのに対し D-SARI で負の相関が見られた．これは，平易度と十分性がトレードオフの関係にあることを示唆する．

4.2　分析

人手評価の揺れと評価精度　主観性の高い評価タスクでは，アノテータの合意度が低下するにつれて LLM の分類性能も低下する傾向が指摘されている [13]．そこで本研究では，主観性が高いと思われる本タスクにおいても同様の傾向が見られるか検証するため，文書平易度の評価精度が 0-shot で

	相関	ロジスティック回帰	
評価指標	Spearman ρ	係数	標準誤差
必要性	−0.220	−0.53	0.06
十分性	−0.183	−0.39	0.07
文平易度	−0.409	−0.81	0.07
文書平易度	−0.359	−0.64	0.07

表 3: 人手評価の分散と，GPT-4o 0-shot (Baseline) の判定が人手評価と一致したか否か（一致=1，不一致=0）との関係．Spearman 順位相関係数およびロジスティック回帰の結果を示す．

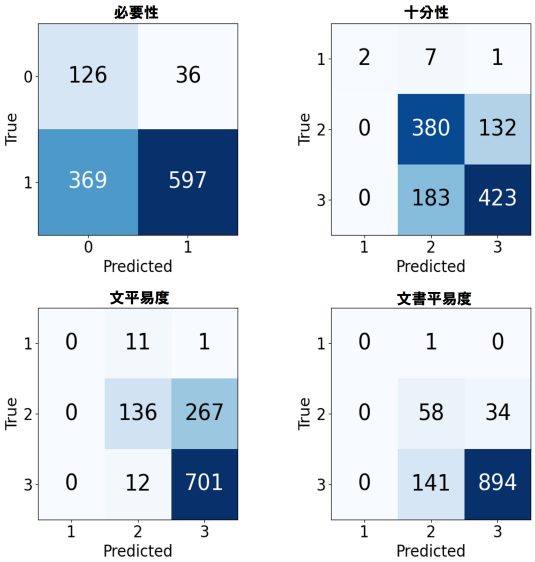

図 2: 人手評価と GPT-4o 0-shot (Baseline) の比較による各指標ごとの混同行列．

最も高かった GPT-4o (Baseline) による自動評価結果を用い，評価項目別にターゲット文書ごとの人手評価の分散と LLM の判定が人手評価と一致したか否か（一致=1，不一致=0）との Spearman 順位相関係数を算出し，ロジスティック回帰分析を行なった（表 3）．[10] その結果，それぞれの係数は，いずれも 4 つの評価項目すべてにおいて負の値を示し，$p < 0.001$ と統計的に有意であった．このことから，文書レベルのテキスト平易度の自動評価タスクにおいても，人手評価の揺れが大きい文書ほど，LLM による評価精度が低下する傾向が確認された．

文書平易度の過小評価傾向　人手評価と比較して，LLM はモデルにより必要性，十分性，文平易度の評価傾向には差異が見られる一方で，文書平易度に関しては，いずれのモデルにおいても一貫して過小評価する傾向を示した．図 2 には，この傾向が顕著に現れた代表的なモデルとして GPT-4o 0-shot (Baseline) 設定の結果を示す．形式的な特徴として，過小評価された文書は，人手評価と一致した文書に比べて漢字の出現割合が高く，ひらがなの割合が低いという傾向が観察された（図 3）．

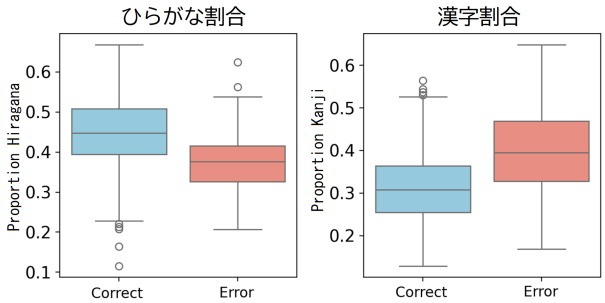

図 3: GPT-4o 0-shot (Baseline) が文書平易度を過小評価した事例のひらがな/漢字割合の箱ひげ図．

そこで，GPT-4o 0-shot (Baseline) 設定において文書平易度が過小評価されていた文書を対象とし，内容的な平易度を保持したまま表層的な平易度のみを変更する操作ののち再評価を行った．具体的には，Sudachi [11] による形態素解析を用い，漢字表記された接続詞・形状詞・形容詞・副詞・動詞をひらがな表記に変換した [12] 上で，GPT-4o 0-shot 設定で再評価した．その結果，29.4%の文書において過小評価が回復することが確認された．この結果は，漢字表記がひらがな・カタカナ表記に比べて難解な印象を与えやすいという日本語特有の性質が，LLM による文書平易度評価に影響を与えている可能性を示唆している．

5　おわりに

本研究では，日本語における文書レベルのテキスト平易化における人手評価と LLM による自動評価を比較した．主要な貢献として，自動評価においては GPT-4o および Gemini は，0-shot 設定でも人手評価と高い一致率を示し，日本語の文書レベルのテキスト平易化の自動評価器として有用であることを実証した．加えて，漢字の表層的な難しさが LLM の文書平易度に影響を与えることを検証した．

10）　なお，ロジスティック回帰では，評価指標間で分散のスケールが異なるため，説明変数である人手評価の分散を評価項目ごとに標準化した．

11）　形態素解析には SudachiPy（version 0.6.10）を用い，辞書には SudachiDict-core（version 20250825）を使用した．分割モードは SplitMode.A とした．

12）　表層的な操作の具体例は Appendix の表 5 に示す．

謝辞

本研究成果は国立研究開発法人情報通信研究機構 (NICT) の委託研究 “自動翻訳の精度向上のための「マルチモーダル情報の外部制御可能なモデリング」の研究開発” によって得られたものである.

参考文献

[1] 田中日加吏, Zhousi Chen, Kexin Bian, 小町守. 文書レベルの日本語平易化の評価基準の提案とデータセット構築. NLP2025 ワークショップ: LLM 時代のことばの評価の現在と未来, 2025.

[2] Renliang Sun, Hanqi Jin, and Xiaojun Wan. Document-level text simplification: Dataset, criteria and baseline. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7997–8013, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[3] Matthew Shardlow. A survey of automated text simplification. **International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Natural Language Processing 2014**, Vol. 4, No. 1, 2014.

[4] Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. Cross-sentence transformations in text simplification. In Amittai Axelrod, Diyi Yang, Rossana Cunha, Samira Shaikh, and Zeerak Waseem, editors, **Proceedings of the 2019 Workshop on Widening NLP**, pp. 181–184, Florence, Italy, August 2019. Association for Computational Linguistics.

[5] Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. Document-level text simplification with coherence evaluation. In Sanja Štajner, Horacio Saggio, Matthew Shardlow, and Fernando Alva-Manchego, editors, **Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability**, pp. 85–101, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria.

[6] Philippe Laban, Jesse Vig, Wojciech Kryscinski, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. SWiPE: A dataset for document-level simplification of Wikipedia pages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 10674–10695, Toronto, Canada, July 2023. Association for Computational Linguistics.

[7] Sofia Blinova, Xinyu Zhou, Martin Jaggi, Carsten Eickhoff, and Seyed Ali Bahrainian. SIMSUM: Document-level text simplification via simultaneous summarization. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 9927–9944, Toronto, Canada, July 2023. Association for Computational Linguistics.

[8] Yoshinari Nagai, Teruaki Oka, and Mamoru Komachi. A document-level text simplification dataset for Japanese. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 459–476, Torino, Italia, May 2024. ELRA and ICCL.

[9] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. **Transactions of the Association for Computational Linguistics**, Vol. 4, pp. 401–415, 2016.

[10] Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 238–255, Vienna, Austria, July 2025. Association for Computational Linguistics.

[11] Itai Mondshine, Tzuf Paz-Argaman, and Reut Tsarfaty. Beyond English: The impact of prompt translation strategies across languages and tasks in multilingual LLMs. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, **Findings of the Association for Computational Linguistics: NAACL 2025**, pp. 1331–1354, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[12] Xiyan Fu and Wei Liu. How reliable is multilingual LLM-as-a-judge? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Findings of the Association for Computational Linguistics: EMNLP 2025**, pp. 11040–11053, Suzhou, China, November 2025. Association for Computational Linguistics.

[13] Junyu Lu, Kai Ma, Kaichun Wang, Kelaiti Xiao, Roy Ka-Wei Lee, Bo Xu, Liang Yang, and Hongfei Lin. Is LLM an overconfident judge? unveiling the capabilities of LLMs in detecting offensive language with annotation disagreement. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Findings of the Association for Computational Linguistics: ACL 2025**, pp. 5609–5626, Vienna, Austria, July 2025. Association for Computational Linguistics.

A 付録

Necessity **（必要性，0/1）**	ターゲット文書に，ソース文書の主要な情報要素（Who，What，When，Where）がすべて保持されているかを評価する．なお，ソース側に存在しない要素がターゲットに含まれていない場合は減点の対象としない． **1:** すべての必要な要素が含まれている． **0:** 一つ以上の必要な要素が欠落している．
Sufficiency **（十分性，1–3）**	ターゲット文書が，ソース文書の主要な内容や全体的な意味をどの程度保持しているかを評価する． **3:** 原文の主要な内容と意図を完全に保持している． **2:** 主要な内容を部分的に保持しているが，一部の詳細が省略または歪曲されている． **1:** 原文の主要な内容を十分に伝えられていない．
Sentence Simplicity **（文平易度，1–3）**	各文の語彙的・構文的な平易さを，小学 6 年生程度の読解レベルを基準として評価する．この評価は，ソース文書を参照せず，ターゲット文書のみを基に行う． **3:** 文が短く，平易な語彙を用い，複雑な構文が最小限である． **2:** 全体的に平易だが，一部に難解な語彙や複文が含まれる． **1:** 難解な語彙や長く複雑な構文が頻繁に現れる．
Document Simplicity **（文書平易度，1–3）**	文書全体の読みやすさおよび理解しやすさを，小学 6 年生程度の読者を基準として評価する．この評価は，ソース文書を参照せず，ターゲット文書のみを基に行う． **3:** ほとんどの小学 6 年生読者にとって容易に理解できる． **2:** およそ半数の小学 6 年生読者が理解できる． **1:** 多くの小学 6 年生読者にとって理解が難しい．

表 4: JADOS-eval データセットで人手評価の際に用いられた平易化評価ガイドライン． 本研究では GPT 系の自動評価の際にプロンプトとして与えた．

B 一部漢字をひらがなにする表層操作の具体例

	文書例
変更前	ビザンティン建築とは 330 年から 1453 年までのほぼ 1100 年間にも**及ぶ**時代を**指します**．ローマ帝国では国教となったキリスト教の礼拝空間が形成され，初期キリスト教建築と**呼ばれます**．しかし，イスラム帝国や異民族の侵入による国土の縮小，帝国の政治機構の転換などに**伴って**ビザンチン建築も変容し，**特有**の建築形態を獲得しました．
変更後	ビザンティン建築とは 330 年から 1453 年までのほぼ 1100 年間にも**およぶ**時代を**さします**．ローマ帝国では国教となったキリスト教の礼拝空間が形成され，初期キリスト教建築と**よばれます**．しかし，イスラム帝国や異民族の侵入による国土の縮小，帝国の政治機構の転換などに**ともなって**ビザンチン建築も変容し，**とくゆう**の建築形態を獲得しました．

表 5: 漢字の表層的な難解さが LLM の文書平易度の評価に与える影響の検証に用いたデータの具体例．