人手評価データセットの統計的類似性と代替可能性

沼屋 征海 [1]　赤間 怜奈 [1,3,4]　守屋 彰二 [1]　佐藤 志貴 [1,2*]　鈴木 潤 [1,4,5]
[1] 東北大学　[2] サイバーエージェント　[3] 国立国語研究所
[4] 理化学研究所　[5] 国立情報学研究所 LLMC
is-failab-research@grp.tohoku.ac.jp

概要

コストの高い人手評価において，1 設定で得たモデル評価を他設定の評価として近似的に代替できる可能性がある．本研究では，人手評価の統計的性質に着目し，既存のモデル評価から代替可能性を推定する枠組みを提案する．具体的には，サンプル内標準偏差および平均スコアの分布をヒストグラム化し，分布間距離で類似度を定量化する．実験では 33 設定を対象に 10 モデルの応答を収集し，LLM 評価で人手評価を再現して順位相関を算出した．結果，標準偏差分布の類似度と順位相関の間に正の相関が確認され，評価者間のばらつきが近い設定ほどモデルの相対的優劣が一致しやすいことが示唆された．

1　はじめに

大規模言語モデル (LLM) の活用領域が拡大する中で，タスクごとに求められる各指標に対してモデルを評価する必要がある．評価手法として人手評価が広く用いられており，指標ごとにスコア付けを行う手法 [1, 2] や Chatbot Arena [3] のようなモデル順位による相対性能を測る手法が存在する．人手評価は自動評価と比較して信頼性が高い一方で，評価者の雇用費や事前説明を含む金銭的，時間的コストが高い．先行研究では，得られたモデル順位が異なるタスク間で相関することが示されており [4]，特定の条件下では，一方の設定で得られたモデル順位から他方の設定における結果を近似できる可能性がある．入出力形式や評価指標を変えた多様な評価設定が存在する中，モデル順位を既存の結果で近似的に代替できれば評価コスト削減に有用である．

従来までの代替可能性の議論では，一般に多数モデルに対する評価値を各設定で取得し順位相関を算出する必要があった [4]．これは評価の代替によってコスト削減を目指す一方で，代替可能性の推定自体に評価コストが発生するという実用上の制約となりうる．そこで本研究では，代替可能性を既存の少数モデル評価から推定することを目的とする．本研究では「統計的特性における類似度が高いほど代替可能性が高い」という仮説のもと，評価設定の統計的特性として評価者間のスコアのばらつきにおける分布または平均スコア分布の類似性に着目し，これらが設定間の代替可能性と対応するか (図 1) を検証する．低コストで代替可能な評価設定を特定できれば，新たなタスクに対する近似的な評価結果として代用することが期待できる．

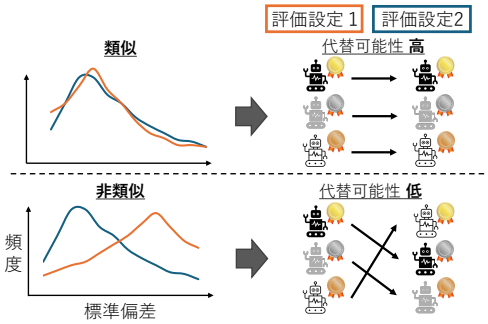

図 1　統計的類似性と代替可能性の対応．

実験では，5 つの文生成タスク，17 のデータセット，評価指標を区別した計 33 の設定を対象に 10 モデルの応答を収集し，LLM 評価により人手評価を近似的に再現することで [5] 設定間のモデル順位相関を算出する．同サンプル内のスコア標準偏差および平均スコアの分布間の距離を算出することで，設定間の統計的性質の類似度を定量化する．結果として，標準偏差分布の類似度とモデル順位相関との間に正の相関が確認され，評価者間のばらつきという統計的性質が近い評価設定同士ほど，モデルの相対的な優劣が一致しやすいことが示された．

2　関連研究

人手評価によるモデル順位　人手評価では，指標ごとのスコア付けによる絶対評価だけでなく [1, 2],

* 東北大学の学術研究員としての成果

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

順位付けによる相対評価も広く用いられている [3, 6]．代表的なプラットホームである Chatbot Arena では，多様なトピックにおける 2 モデルの応答に対するユーザ投票を収集することでモデル順位を測定する [3]．一般に人手評価は高コストであるため，LLM を評価者（LLM-as-a-judge）として人手の順位またはスコアを近似的に定量化する研究も進められており，適切な設定の下では人手評価と一定の相関が得られることが報告されている [5]．

評価の冗長性　評価の冗長性とは，データセット中のサンプル間や異なる評価設定間において測定される能力が重複している状態を指す [7]．ベンチマークが異なってもモデルの順位が一定程度一致することが報告されていると同時に [4]，多数のベンチマークにおける成績が少数の潜在能力で説明できるという結果も報告されており [8]，設定間で評価を代替できる可能性を示唆する．本研究では，1 つの手法として評価者間のスコアのばらつきにおける分布と平均スコア分布それぞれの類似度に着目して代替可能性を推定することを考える．

3　代替可能性推定

本研究では，「統計的性質における評価設定間の類似度が高いほど代替可能性が高い」という仮説を置き，これに基づき既存のモデル評価から低コストで代替可能性を推定することを目指す．本節では，この仮説を検証するための手法を定義する．

3.1　評価設定の定義

本研究における比較および分析はデータセット D と評価指標 c の組からなる評価設定 (D, c) を単位とする．ここで，データセット D は特定の入力と出力の形式を，評価指標 c はその出力を評価する観点を表す．本研究では，タスク設定や評価指標を区別せず，あらゆる評価設定同士において人手評価における代替可能性を推定できるかを検証する．

3.2　統計的類似度

信頼性を担保するため，人手評価では複数評価者の平均を最終スコアとすることが多い [9, 10]．1 サンプル（入力とモデル出力からなる評価単位）内の評価のばらつきは主観性を [11]，平均スコアの分布形状はタスク設定の難しさなどを反映すると考えられる．本研究では，各評価設定 (D, c) の統計的性質として，サンプルごとに付与された複数スコアの標準偏差 (サンプル内標準偏差) σ_i とサンプルごとの平均スコア μ_i を用いる．付与されたスコアを $[0, 1]$ の範囲になるように正規化した上で σ_i と μ_i を算出する．各設定について，μ_i および σ_i のヒストグラムおよび (μ_i, σ_i) の 2 次元ヒストグラムを作成し，確率分布となるよう正規化する．評価設定 (D_1, c_1) と (D_2, c_2) の統計的類似度は Jensen–Shannon divergence（JSD）により定量化し，(i) (μ, σ) の同時分布，(ii) μ の分布，(iii) σ の分布に対してそれぞれ JSD を計算する．JSD は値が小さいほど分布が類似していることを意味するため，p, q を分布として $1 - \mathrm{JSD}(p, q)$ で統計的類似度を定義する．

3.3　代替可能性の定義と測定

本研究では，評価設定間の代替可能性を「複数の対象モデルの評価値に基づく順位がどの程度一致するか」として捉える．各評価設定 (D, c) において複数モデルのスコアをサンプル平均で集約し，モデル順位を得る．評価設定 (D_1, c_1) と (D_2, c_2) のモデル順位間のスピアマン相関係数を算出し，相関が高いほど代替可能性が高いと解釈する．

4　実験

4.1　データセット

主要な文生成タスクのうち要約，対話，翻訳，物語生成，コード生成を対象とし，合計 17 の人手評価データセット，33 評価設定を用いた．なお，標準偏差と平均スコアの取りうる値を揃えるため，基本の対象を 5 段階評価であるものとし，同指標に対して各サンプルで 3 つの評価値が付与されている状態とした．詳細は付録 A に示す．

4.2　モデル順位の作成

評価対象モデル　LMArena (旧 Chatbot Arena [3]) の Text Arena[1)] で，2025 年 12 月時点で上位に位置する 10 モデルを採用した．具体的には gpt-5.2 [12], o3 [13], gemini-3-pro-preview [14], gemini-2.5-pro [15], claude-opus-4.5 [16], claude-sonnet-4.5 [17], deepseek-v3.1 [18], deepseek-r1 [19], mistral-medium-3[2)], qwen3-235b-a22b-instruct [20] を用いた．

文生成プロンプト　各設定に対するモデル指示を作成するため，Honovich らが導入した 5 つの入出力

1) https://lmarena.ai/ja/leaderboard/text
2) https://mistral.ai/news/mistral-medium-3

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

表 1　統計的類似度と代替可能性とのスピアマン相関係数 ($^{*}p < .05, ^{**}p < .01, ^{***}p < .001$).

ヒストグラム	全体	異データ	同データ
2D	0.238***	0.151***	0.529**
1D 標準偏差	**0.351****	**0.294****	**0.540****
1D 平均スコア	0.188***	0.102*	0.463**

を LLM に渡すことでインストラクションを作成する手法 [21] を用いた．プロンプトの構成は，インストラクション，参照応答による 5-shot，入力の順とした [22]．詳細は付録 B.1 に示す．

LLM 評価　LLM-as-a-judge によって人手評価を近似的に再現し，モデル順位を作成する．評価モデルとして gpt-5.2 を用いた．評価プロンプトは Murugadoss らの研究において人手評価と最大の相関を示した評価基準を明記するテンプレートを採用した [5]．詳細は付録 B.2 に示す．

4.3　実験結果

評価設定組ごとに第 3 節で定義した統計的類似度および代替可能性を算出した後，これらのスピアマン相関係数を算出した．さらに，評価設定組を異なるデータセット間の組と同一データセット内で指標のみが異なる組で切り分けることで，タスクまたは指標横断的に議論することを可能とした．結果を表 1 に示す．共通して統計的に有意な正の相関を示し，特にサンプル内標準偏差分布の類似度で高い相関を示した．また，同一データセット間では相対的に高い正の相関を示した．これらの結果から，今回の設定において，統計的類似度が高いほど代替可能性も高い傾向にあることが示唆された．統計的性質としては，サンプル内標準偏差に基づく類似度のほうが代替可能性と高い相関を示した．

5　分析

実験では，サンプル内標準偏差分布の類似度が代替可能性と高く相関した．この原因を議論し，標準偏差がどのような性質を表すかを考察する．

5.1　タスク間比較

サンプル内標準偏差は評価者間のばらつきを表し，評価設定の主観性を反映する [11]．図 2 に，サンプル内標準偏差分布の類似度を可視化したヒートマップを示す．この図から，物語生成は他の評価設定との類似度が低く，性質が異なることがわかる．

この要因を詳細に調べるため，サンプル内標準偏差の分布の偏りをタスク間で比較する．可視化のためにヒストグラムをカーネル密度推定により連続関数で表現したグラフを図 3 に示す．物語生成の分布が他タスクに比べて右に偏り，高い標準偏差をとるサンプルの割合が高かった．つまり，物語生成では評価者間で判断が分かれやすい可能性が高く，これは物語生成が嗜好性や主観性と関連するためであると考えられる．また，要約 (Coherence) は低い標準偏差に分布が偏っている．これは，要約文に対して原文という明確な評価基準が示されているためだと考えられる．さらに，対話と翻訳は考慮すべき評価基準の多い包括的な指標であるため，要約より標準偏差が高く非常に似た分布となったと考えられる．以上より，サンプル内標準偏差分布は評価設定がもつ評価基準の明確度を反映し，嗜好性や主観性を含む設定で高い値に分布が偏る傾向が示された．したがって，サンプル内標準偏差によって評価基準の明確度や嗜好性，主観性を捉え，その類似度に基づきモデル順位を近似的に代替できる評価設定の組を特定できる可能性が示唆された．

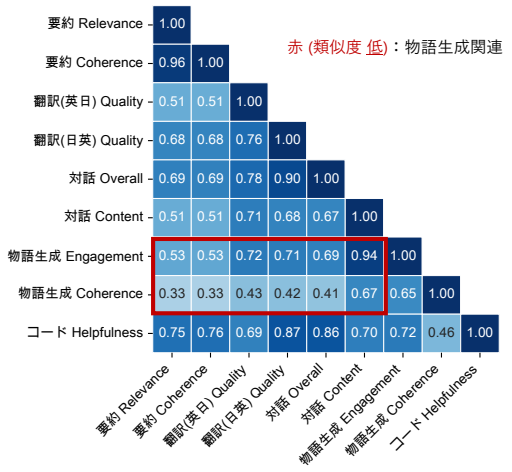

図 2　標準偏差分布のタスク間類似度．

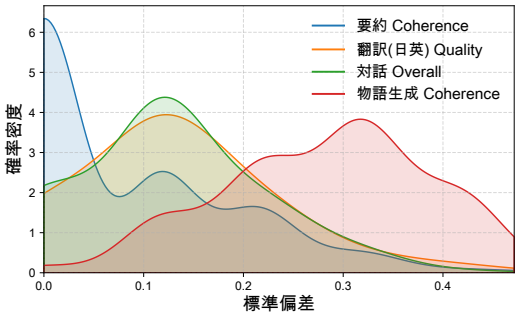

図 3　複数タスクの標準偏差のカーネル密度関数．

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

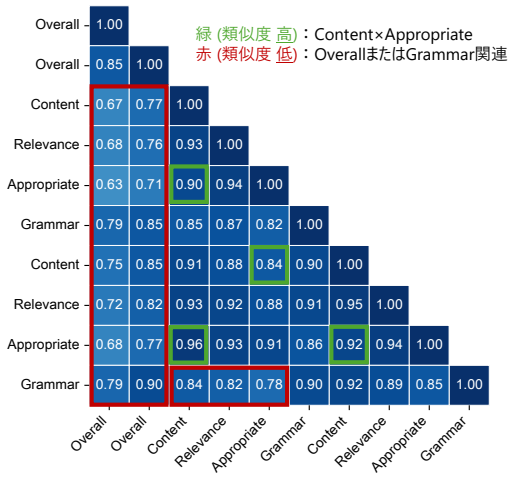

図 4　対話タスクにおける標準偏差分布の指標間類似度.

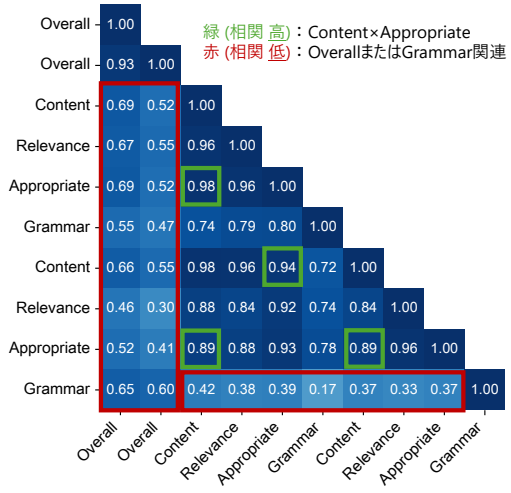

図 5　対話タスクにおける指標間順位相関.

5.2　指標間比較

次に，タスク間だけでなく同一タスク内での評価指標間でも同様の傾向が確認されるかを調べた．対話タスクにおけるサンプル内標準偏差分布の類似度，順位相関のヒートマップをそれぞれ図 4，5 に示す．対話タスクでは全体として高い類似度，順位相関を示した．個別に見ると，包括的な指標である総合評価 (Overall) では他と比較して低い類似度，順位相関を示した．また，評価基準が明確である文法性 (Grammar) は類似度，順位相関がやや低い傾向を示した．さらに意外にも，内容の充実度 (Content) や適切さ (Appropriate) のような評価者の嗜好等によって基準が変動しうる指標は互いに類似した．これは一見評価基準が異なる指標同士でも代替できる可能性を示唆する．以上より，対話を例とした同一タスクにおける評価指標間の分析においても，サンプル内標準偏差は評価基準の明確度や嗜好性の関連度を表す統計的特性となりうること，異なる指標間での代替可能性が存在することが示された．これは複数指標で評価するタスクにおいて，指標間でモデル順位を近似的に代替できる可能性を示唆する．

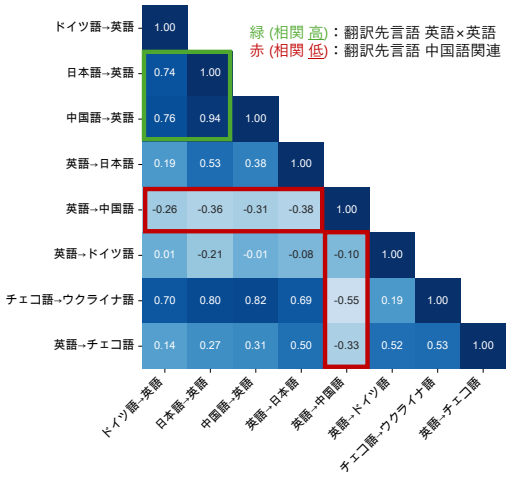

図 6　同一指標の翻訳タスクにおける言語間順位相関.

5.3　翻訳言語間比較

最後に，翻訳言語間の代替可能性を比較した．図 6 に同一評価指標の翻訳タスクにおける順位相関ヒートマップを示す．翻訳タスクでは翻訳先言語によって代替可能性が変動すること，類似タスクかつ同一指標でも代替可能性が低くなる場合があることが示唆された．たとえば，翻訳先言語が英語の設定間では高い順位相関を示した．一方で，翻訳先言語が中国語の場合，他の設定と負の順位相関を示した．これは中国語翻訳における DeepSeek 系モデルの評価スコアが特に高かったことに起因していると考えられる．代替可能性の議論において，タスクや指標の類似性だけでなくモデル集合の特性を考慮することの重要性が示唆された．

6　おわりに

評価設定間の代替可能性を推定するために，サンプル内標準偏差と平均スコアを人手評価の統計的性質として定義し，これらの分布類似度がモデル順位相関と対応するかを検証した．実験の結果，サンプル内標準偏差分布の類似度が高い評価設定間ほどモデル順位が近くなることが示唆される結果が得られた．したがって，対象設定に対して標準偏差分布の類似度が高い設定を特定できれば，そのモデル順位を近似的に代替できる可能性がある．

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

謝辞

本研究は，JSPS 科研費 JP25K21263，JST BOOST JPMJBY24A1，JST ムーンショット型研究開発事業 JPMJMS2011-35 (fundamental research)，および，文部科学省の補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の助成を受けたものです．

参考文献

[1] Emre Can Acikgoz, Carl Guo, Suvodip Dey, Akul Datta, Takyoung Kim, Gokhan Tur, and Dilek Hakkani-Tur. TD-EVAL: Revisiting task-oriented dialogue evaluation by combining turn-level precision with dialogue-level comparisons. In **Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, 2025.

[2] Dongqi Liu, Chenxi Whitehouse, Xi Yu, Louis Mahon, Rohit Saxena, Zheng Zhao, Yifu Qiu, Mirella Lapata, and Vera Demberg. What is that talk about? a video-to-text summarization dataset for scientific presentations. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics**, 2025.

[3] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: an open platform for evaluating llms by human preference. In **Proceedings of the 41st International Conference on Machine Learning**, 2024.

[4] Yotam Perlitz, Ariel Gera, Ofir Arviv, Asaf Yehudai, Elron Bandel, Eyal Shnarch, Michal Shmueli-Scheuer, and Leshem Choshen. Benchmark agreement testing done right: A guide for LLM benchmark evaluation. In **NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling**, 2025.

[5] Bhuvanashree Murugadoss, Christian Poelitz, Ian Drosos, Vu Le, Nick McKenna, Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. Evaluating the evaluator: measuring llms' adherence to task evaluation instructions. In **Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence**, 2025.

[6] Wafaa Mohammed, Sweta Agrawal, Amin Farajian, Vera Cabarrão, Bryan Eikema, Ana C Farinha, and José G. C. De Souza. Findings of the WMT 2024 shared task on chat translation. In **Proceedings of the Ninth Conference on Machine Translation**, 2024.

[7] Zicheng Zhang, Xiangyu Zhao, Xinyu Fang, Chunyi Li, Xiaohong Liu, et al. Redundancy principles for MLLMs benchmarks. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics**, 2025.

[8] Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance. In **Proceedings of the 38th International Conference on Neural Information Processing Systems**, 2024.

[9] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. Benchmarking large language models for news summarization. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 39–57, 2024.

[10] Cyril Chhun, Fabian M. Suchanek, and Chloé Clavel. Do language models enjoy their own stories? prompting large language models for automatic story evaluation. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 1122–1142, 2024.

[11] Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating summarization evaluation. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 391–409, 2021.

[12] OpenAI. Update to gpt-5 system card: Gpt-5.2. Technical report, OpenAI, 2025.

[13] OpenAI. Openai o3 and o4-mini system card. Technical report, OpenAI, 2025.

[14] Google DeepMind. Gemini 3 pro model card. Technical report, Google DeepMind, 2025.

[15] Google DeepMind. Gemini 2.5 pro model card. Technical report, Google DeepMind, 2025.

[16] Anthropic. Claude opus 4.5 system card. Technical report, Anthropic, 2025.

[17] Anthropic. Claude sonnet 4.5 system card. Technical report, Anthropic, 2025.

[18] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, et al. Deepseek-v3 technical report, 2024. https://arxiv.org/abs/2412.19437.

[19] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. https://arxiv.org/abs/2501.12948.

[20] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report, 2025. https://arxiv.org/abs/2505.09388.

[21] Or Honovich, Uri Shaham, Samuel R. Bowman, and Omer Levy. Instruction induction: From few examples to natural language task descriptions. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, 2023.

[22] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, et al. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, 2022.

[23] Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 391–409, 2021.

[24] Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, et al. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In **Proceedings of the Eighth Conference on Machine Translation**, 2023.

[25] Shikib Mehri and Maxine Eskenazi. USR: An unsupervised and reference free evaluation metric for dialog generation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, 2020.

[26] Chen Zhang, João Sedoc, Luis Fernando D'Haro, Rafael Banchs, and Alexander Rudnicky. Automatic evaluation and moderation of open-domain dialogue systems, 2021. https://arxiv.org/abs/2111.02110.

[27] Mario Rodríguez-Cantelar, Chen Zhang, Chengguang Tang, Ke Shi, Sarik Ghazarian, João Sedoc, Luis Fernando D'Haro, and Alexander I. Rudnicky. Overview of robust and multilingual automatic evaluation metricsfor open-domain dialogue systems at DSTC 11 track 4. In **Proceedings of the Eleventh Dialog System Technology Challenge**, 2023.

[28] Mikhail Evtikhiev, Egor Bogomolov, Yaroslav Sokolov, and Timofey Bryksin. Out of the bleu: How should we assess quality of the code generation models? **J. Syst. Softw.**, Vol. 203, No. C, 2023.

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

表 2　用いたデータセットと評価指標.

タスク	データセット	評価指標	引用
要約	Summeval	Coherence / Relevance / Fluency / Consistency	[23]
	Benchmarking LLM for News Summarization (CNN/DM) / (XSUM)	Coherence / Relevance	[9]
翻訳	WMT23 News Systems and Evaluations (cs→uk) / (de→en) / (en→cs) / (en→ja) / (en→zh) / (en→de) / (ja→en) / (zh→en)	Translation Quality	[24]
対話	USR (PersonaChat) / (Topical-Chat)	Overall	[25]
	Topical-DTSC10 / Persona-DSTC10	Appropriateness / Content Richness / Grammatical Correctness / Relevance	[26, 27]
物語生成	HANNA Benchmark	Relevance / Coherence / Empathy / Surprise / Engagement / Complexity	[10]
コード生成	Evaluation of metrics for code generation (CoNaLa)	Helpfulness	[28]

A　データセット詳細

A.1　対象データセットの条件と前処理

統計的類似度では標準偏差を計算するため，1 サンプルに対して複数の人手評価スコアが付与されている必要がある．また JSD の性質上，標準偏差や平均スコアの取りうる値が一致するほど類似度が高くなる傾向にあるためスコア粒度を揃える必要がある．これらを考慮し，以下の条件を満たすデータセットを本実験の対象とした．

- 同評価指標に対して 1 サンプルに 3 つ以上の評価スコアが付与かつ公開されている．
- 評価が 5 段階である．

なお，翻訳タスクは 0-100 の Direct Assessment が用いられることが多い [24]．よって標準偏差や平均スコアの取りうる値を揃えるため，5 段階に畳み込むことで評価スコアを用いた．さらに，同評価指標に対して付与された評価スコアが 3 つ未満のサンプルは対象外とし，3 つを超える場合は無作為に 3 つとすることによって全対象サンプルで同数のスコアが付与されている状態とした．

A.2　用いたデータセット

表 2 に実験で用いたデータセットと評価指標を示す．本研究では，データセットを入力と出力の形式として定義したため，Benchmarking LLM for News Summarization における CNN/DM と XSUM，WMT23 News Systems and Evaluations における複数の言語ペアなどは異なるデータセットとしてみなした．

B　プロンプト

B.1　応答生成プロンプト

モデル応答を得るために用いたプロンプトを図 7 に示す．`{instruction}`はデータセットごとに作成した．

B.2　LLM 評価プロンプト

応答に対して LLM 評価を付与するためのプロンプトを図 8 に示す．`{instruction}`は応答生成時と同様である．また，評価基準 (Rubrics) は元論文を基に作成した．

```
## Instruction
{instruction}

## Examples
### Example 1
Input:
{input example1}
Output:
{output example1}

## Evaluation Instance
Input:
{input}
Output:
```

図 7　応答生成のためのプロンプト.

```
# Task to evaluate
Your tasks is to evaluate the interaction between a user and an AI assistant. I want you to evaluate the assistant's response. Evaluate the response for the given rubric below. Use the rubric to guide you evaluating and base all your evaluation decision on the rubric.
The AI responses are for: {instruction}

# Sample to evaluate
User:
{input}
Assistant:
{output}

# Rubrics
{rubrics 1}: {rubrics 1 instruction}
Score 1: {score1 criteria}
...
Score 5: {score5 criteria}

# Instructions
Evaluate the quality of the response from the sample and return a score between 1 and 5 as JSON:
## Score: [JSON]
Return only:
{"<metric names from Rubrics>": <score 1-5>, ...}
```

図 8　LLM 評価付与のためのプロンプト.

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