アライメントが大規模言語モデルの数値バイアスに与える影響

佐藤郁子 [1] 金輝燦 [1] 陳宙斯 [2] 三田雅人 [3,1] 小町守 [2]
[1] 東京都立大学 [2] 一橋大学 [3] 株式会社サイバーエージェント
{sato-ayako, kim-hwichan}@ed.tmu.ac.jp, mita_masato@cyberagent.co.jp
{zhousi.chen, mamoru.komachi}@r.hit-u.ac.jp

概要

大規模言語モデル（LLM）を評価者として用いる “LLM-as-a-judge” は，機械翻訳品質推定（MTQE）をはじめとする，多くの評価タスクで効果を示している．しかし，指示チューニングや強化学習によってアライメントされた LLM では，特定の評価スコアを頻繁に生成する数値バイアスが確認されている．この現象は，アライメントにより出力の多様性が減少するという既存の知見と一致しており，多様性の欠如は評価スコアの偏りを引き起こす可能性がある．その結果，入力の微細な変化に対する評価の頑健性が損なわれる懸念がある．本研究では，LLM のアライメントが数値バイアスおよびタスク性能に与える影響を調査する．

1 はじめに

大規模言語モデル (LLM) をシステム出力の評価者として利用する手法は，“LLM-as-a-judge” [1] と呼ばれる．この手法は，BLEU [2] や ROUGE [3] のような表層評価指標と比較して，参照なしでシステム出力を直接評価させるため，人間のアノテータの時間と労力を削減し，よりスケーラブルな評価プロセスを可能にする．また，機械翻訳品質推定 (Machine Translation Quality Estimation; MTQE) [4, 5, 6] のような多言語タスクでも，その有効性が示されている．

評価者として用いられるモデルは，指示チューニングや RLHF [7] などのアライメント手法が適用された後の LLM (post-alignment LLM) が主流である．**アライメント**とは，モデルが人間の意図や価値観に沿った応答を生成できるように調整するプロセスであり，これによってモデルは優れた指示追従能力を獲得し，高い評価性能を実現している．しかし，このようにアライメントが施されたモデルには新たな課題が浮上している．その一例として，図 1 に示すような，入力に寄らずモデルが特定の数値トークン

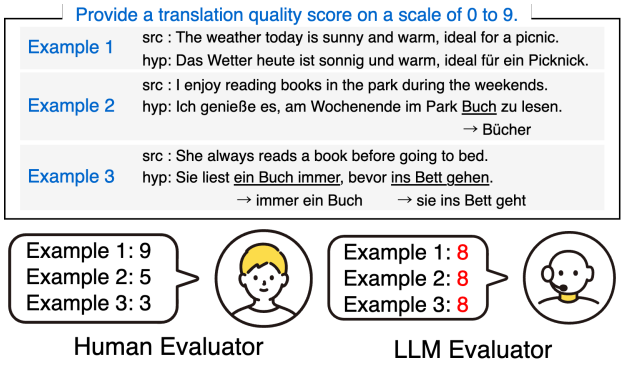
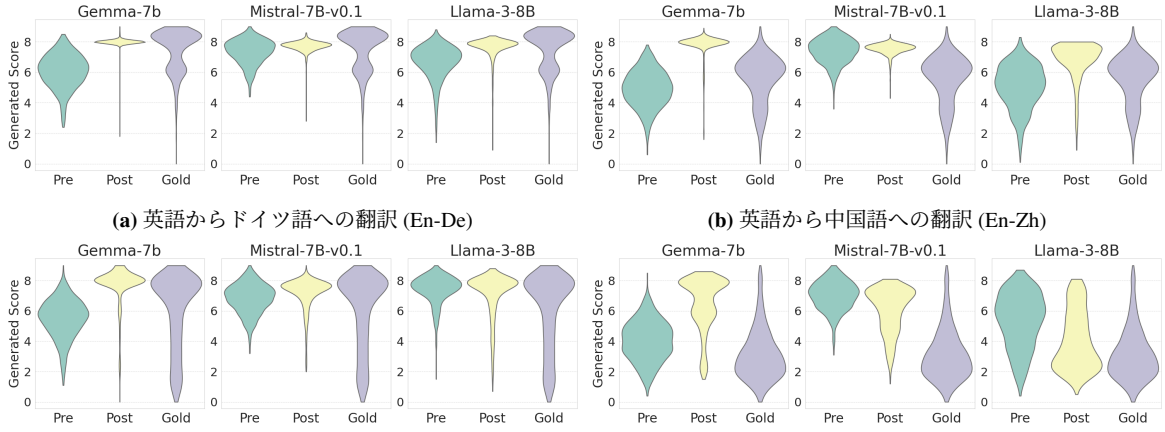


図 1: MTQE タスクのスコアベース評価における LLM 評価者のバイアスの例．LLM 評価者は，入力に関係なく特定の数値を頻繁に使用する傾向がある．

（例えば，0–9 の範囲で 8）を頻繁に生成する現象が確認されている．本研究では，この現象を**数値バイアス**と呼ぶ．特定のトークンへの極端な集中は，モデルが異なる出力間の差異を適切に評価する能力を低下させる可能性がある．そのため，数値バイアスの原因の特定と対処が不可欠である．

我々は，数値バイアスの原因を考える上で，LLM のアライメントがモデル出力に影響を与える可能性に注目する．先行研究では，アライメントによってモデルの多様性 [8] や創造性 [9] が損なわれることが指摘されており，これが出力内容を特定のパターンに偏らせる要因になりうると示唆されている．こうした偏りが，数値トークンの生成にも影響を及ぼし，不自然な分布を生む可能性がある．アライメントの過程で出力可能な数値トークンの分布が不自然に偏る副作用が発生する場合，これは評価タスクにおいて望ましい影響ではない．さらに，多言語が関与する MTQE のようなタスクでは，こうしたバイアスが言語ごとに異なる形で現れる可能性があり，評価性能のばらつきを引き起こす要因となりうる．

これらの背景を踏まえ，本研究では次の 2 つの研究課題に基づいて，LLM のアライメントが数値バイアスに与える影響についてより深い分析を行う．



(a) 英語からドイツ語への翻訳 (En-De)

(b) 英語から中国語への翻訳 (En-Zh)

(c) ロシア語から英語への翻訳 (Ru-En)

(d) ネパール語から英語への翻訳 (Ne-En)

図 2: En-De，En-Zh，Ru-En，Ne-En の MTQE スコア分布.

RQ1　LLM のアライメントは数値バイアスにどのような影響を与えるのか？ また，その影響は言語ごとに異なるのか？

RQ2　数値バイアスを緩和し，評価タスク性能を向上するには，どのような評価設定が有効か？

RQ1（アライメントの影響）について，pre-alignment モデルと post-alignment モデルの評価スコア分布を比較した結果（図 2），post-alignment モデルは特定スコアに偏る傾向が顕著であり，言語によってアライメントの影響の大きさが異なることが示唆された．また，数値バイアスの強化による評価精度低下も確認された．RQ2（数値バイアスの緩和）に対しては評価スコア範囲と温度パラメータの 2 つの設定の影響を調査し，指定するスコア範囲でバイアスの強度が異なること，特定のモデルでは温度パラメータの増加でバイアスが緩和することが示された．

2　関連研究

Chiang ら [10] は LLM が人間の代替として評価者になれるかについて，物語生成と敵対的攻撃サンプル生成の 2 つのタスクの品質評価で実験を行い，LLM の評価は人間の専門家の評価と一致することを示した．また論文中で，LLM による評価のメリットとして，再現性の高さ，コストの低さなどを挙げている．同時に，事実知識に基づいた評価には適さないこと，有害な回答のリスクなどのデメリットも指摘し，目的に沿って評価手段を選択する重要性を強調している．この研究では，LLM 評価者の利点と課題を広く議論しているが，バイアスの定量化やその影響については十分に分析されていない．

Sturebor ら [11] は，LLM-as-a-judge について次の 3 種類のバイアスを提案し，要約タスクでこれらのバイアスを確認した．(i) 有用なテキストより低難易度のテキストを好む．(ii) プロンプトで指定されたスコア範囲内の出力可能な値が少ないほど良い評価性能を示す．(iii) 複数の事例を評価するとき，以前に評価した事例の影響を受ける．これらのバイアスを定量化し，LLM を評価に用いるためのガイドラインを提案している．この研究では様々なバイアスを定量化し，複数の評価設定で比較しているが，数値バイアスについては議論されていない．

Ohi ら [12] は，LLM 評価者を用いて，data-to-text と 文法誤り訂正の 2 種類のタスクでスコアリングによる評価を行った．著者らは，LLM は尤度の高いテキストを好むという尤度バイアスの存在を指摘した．尤度バイアスは，バイアスの強いインスタンスを few-shot 例として使用することで緩和され，評価性能が向上した．この研究と同様に，我々の研究もスコアリングにおけるバイアスを対象としているが，特に数値バイアスに着目し，その発生要因をアライメントの影響と仮定して分析する点で異なる．

3　分析手順

3.1　評価タスク

本研究では，機械翻訳品質推定 (MTQE) に取り組む．MTQE とは，原文とシステム出力を入力として受け取り，翻訳の品質を表す数値スコアを推定するタスクである．すなわち，入力から数値を得る回帰タスクであり，出力の分布を直接可視化できる．本研究では，WMT 2020 Shared Task on Quality Estimation のテストデータ [13] から，7 つの言語ペア (En-De，En-Zh，Et-En，Ne-En，Ro-En，Ru-En，Si-En)

における文レベルの Direct Assessment (DA) [14] スコアを採用する [1]．従って，指定した範囲内のスコアを LLM に出力させ，数値バイアスの有無を検証する．Gold スコアの分布は言語ペアごとに異なり，各言語特有の翻訳の難しさや品質のばらつきを反映している．この多様性は，モデルが翻訳の差異を正確に捉えられるかを検証する上で重要な基準となる．

3.2 LLM を用いたスコアリング

手順 1: プロンプト設計 本研究では，付録 A.3 の表 4 に示すプロンプトテンプレートを使用する．GEMBA [5] を参考に，データセットごとに以下の引数を設定し，プロンプトテンプレートに埋め込む．

- ソース言語の名称: {{source language}}
- ターゲット言語の名称: {{target language}}
- スコア範囲の最小値: {{min score}}
- スコア範囲の最大値: {{max score}}
- ソース言語文: $src_{1..N}$
- ターゲット言語文: $hyp_{1..N}$

スコア範囲は {0–9，1–5，1–100} のいずれかを指定する．作成したプロンプトで数値を 10 回生成する．

手順 2: 評価スコアの計算 10 回の生成結果のうち，非数値トークンは除外し，範囲外の数値は範囲内に収めるためにクリッピング処理を適用し，上限または下限の値に置き換える．処理後の生成スコアの平均を最終的な評価スコアとする．

3.3 研究課題

RQ1: LLM のアライメントは数値バイアスにどのような影響を与えるのか？ また，その影響は言語ごとに異なるのか？ RQ1 を調査するために，まず，pre-alignment モデルと post-alignment モデルの評価スコアの分布をヴァイオリン図で可視化する．さらに，数値バイアスを定量化する指標として尖度を採用[2] し，2 つのモデルそれぞれで計算する[3]．また，評価精度の指標として，人間とモデルの評価スコア間のケンドールの相関係数 τ を計算する．

RQ2: 数値バイアスを緩和し，評価タスク性能を向上するには，どのような評価設定が有効か？ MTQE タスクの数値バイアスに影響する可能性がある設定項目として，スコア範囲と温度パラメータがある．スコア範囲では，評価スコアとして出力可能な数値トークンが変わり，確率分布に影響が生じる．温度パラメータはモデルのランダム性を制御するもので，値が大きいほど出力が多様になる．

1） MTQE のスコアの種類は付録 A.1 で詳細に説明する．
2） 分散も数値バイアスの指標となり得るが，ヴァイオリン図から大まかに判断できるため，本論文では尖度を報告する．
3） Gold のスコア範囲は 0–100 のため，モデルに指定したスコア範囲にスケール変換し，尖度の計算を行う．

言語	モデル	尖度 gold	尖度 pre	尖度 post	τ pre	τ post
En-De	gemma	1.48	**0.27**	128.17	**0.21**	0.05
	mistral	1.48	**0.41**	52.92	0.04	**0.09**
	llama	1.48	**2.53**	21.05	0.21	**0.26**
En-Zh	gemma	0.20	**−0.06**	21.57	**0.20**	0.18
	mistral	0.20	**0.73**	11.33	0.01	**0.13**
	llama	0.20	**−0.06**	3.96	0.17	**0.25**
Et-En	gemma	−1.24	**−0.11**	11.64	**0.31**	0.27
	mistral	−1.24	**0.53**	4.93	0.15	**0.37**
	llama	−1.24	**0.98**	−1.09	0.40	**0.50**
Ne-En	gemma	0.97	**−0.34**	0.45	0.14	**0.28**
	mistral	0.97	0.58	**−0.19**	0.13	**0.28**
	llama	0.97	**−0.64**	−0.95	0.19	**0.28**
Ro-En	gemma	−0.16	**−0.59**	2.73	0.38	**0.43**
	mistral	−0.16	**0.71**	1.12	0.20	**0.49**
	llama	−0.16	**0.56**	−1.45	0.46	**0.60**
Ru-En	gemma	−0.57	**0.40**	9.70	0.20	**0.23**
	mistral	−0.57	**0.17**	6.33	0.13	**0.25**
	llama	−0.57	3.95	**3.11**	0.09	**0.34**
Si-En	gemma	−0.72	**−0.37**	1.97	0.15	**0.24**
	mistral	−0.72	1.38	**0.03**	0.07	**0.26**
	llama	−0.72	**−0.36**	−1.13	0.26	**0.32**

表 1: MTQE スコア分布の尖度と τ．モデルごとに最も絶対値が小さい尖度，最も大きい τ を **太字** で示す．

4 実験

実験では付録 A.2 の 表 3 に示す 3 つのオープン LLM を使用する．非数値出力を抑制するため，*max_token=5* を設定する．サンプリング時の温度パラメータは {0.4, 0.7, 1.0, 1.3} のいずれかを指定する．RQ1 の実験では 0.7 に設定する．

4.1 RQ1 に対する実験の結果

スコア範囲 0 – 9 における En-De，En-Zh，Ru-En，Ne-En の評価スコアの分布を図 2 に示す[4]．pre モデルではスコア分布が広範囲にわたり，モデルや言語ペアによって多様性が見られる．一方，指示追従能力が不十分なため，モデルがランダムな挙動を示す可能性も考えられる．En-De，En-Zh，Ru-En において，post モデルの評価スコアは 8 付近に偏っており，アライメントによる過度な一貫性向上が影響している可能性がある．注目すべきは，Gold データのスコア分布が言語ごとに異なるにも関わらず，post モデルの評価スコア分布が似た形状を示している点である．この結果は，post モデルが翻訳品質の差異を十

4） 他の言語の結果は，付録 B.1 の図 4 に記載する．

分に反映できておらず，入力事例にかかわらず特定のスコアに偏るバイアスが存在することを示唆している．一方，Ne-En では，post モデル間で分布が異なり，特定の数値への過度な集中は見られなかった．すなわち，アライメントによる数値バイアスへの影響は言語ペアによって異なることがわかる．

評価スコア分布の尖度と τ を表 1 に示す．尖度が大きいほど分布形状が鋭く，数値バイアスが強い．En-De，En-Zh，Et-En，Ro-En，Ru-En は post モデルのバイアスが顕著である一方で，Ne-En と Si-En は pre モデルの方が尖度が大きい傾向にある．ネパール語 (Ne) と シンハラ語 (Si) は低資源言語であり，アライメントデータに含まれる割合がアライメントの影響に差異をもたらす可能性がある．モデル別の観点では，gemma は特にアライメントの影響が大きい．

一般に，アライメントによる指示追従能力の向上は，評価精度を改善すると考えられる．しかし，尖度が高い言語ペアやモデルでは τ が小さく，評価精度が低下している．尖度の差 ($kurtosis_{\text{post}} - kurtosis_{\text{pre}}$) と評価精度の差 ($\tau_{\text{post}} - \tau_{\text{pre}}$) の間のピアソンの相関係数は –0.68 と強い負の相関を示しており，バイアスが評価精度を損なう要因であることを示唆する．pre 時点での評価精度はモデルによって異なるものの，強いバイアスは指示追従能力向上による恩恵を損なう恐れがある．これらの結果より，アライメントは特に高資源言語 (例: English，German，Chinese) の品質推定で数値バイアスを強め，評価精度に悪影響を与えると示唆される．

4.2 RQ2 に対する実験の結果

バイアスが顕著な En-De を分析対象とする．

4.2.1 スコア範囲による影響

{0–9, 1–5, 1–100} の 3 つの値に設定した場合のそれぞれの尖度と τ を表 2 に示す．尖度はスコア範囲によって大きく異なり，ほぼすべてのモデルで，0–9 のスコア範囲で尖度が最も大きいことが確認できる．特に，post モデルは尖度の変化が大きい．尖度が大きいほど τ も小さい傾向があり，過度な偏りは評価精度の悪化に繋がることが示唆される．すなわち，post モデルで評価を行う場合は評価スコアの範囲にも注意を払う必要がある．

4.2.2 温度パラメータによる影響

スコア範囲を 0–9 に固定し，{0.4, 0.7, 1.0, 1.3} の 4 つの値に設定した場合の尖度と τ を図 3 に示す [5)]．

5) mistral モデルにおける結果は付録 B.2 の 図 5 に示す．

モデル		尖度 1–5	尖度 0–9	尖度 1–100	τ 1–5	τ 0–9	τ 1–100
gemma	pre	**–0.09**	0.27	0.19	**0.22**	0.21	0.20
	post	**71.79**	128.17	87.45	0.05	0.05	**0.13**
mistral	pre	**0.14**	0.41	–0.24	**0.07**	0.04	0.05
	post	**11.95**	52.92	42.00	**0.13**	0.09	0.11
llama	pre	**–0.03**	2.53	3.23	**0.23**	0.21	0.21
	post	**10.19**	21.05	23.96	0.22	0.21	**0.24**

表 2: 異なる評価スコア範囲を指定した場合の尖度と τ の変化．各スコア範囲，モデルごとに最も絶対値が小さい尖度，最も大きい τ を **太字** で示す．

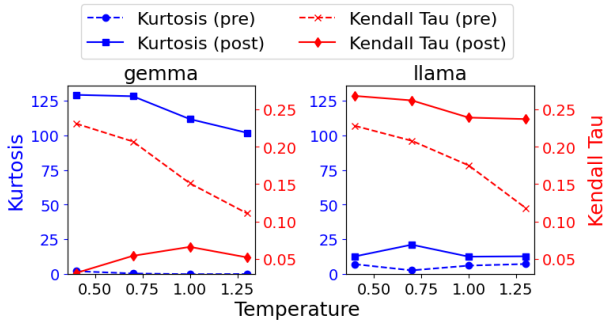


図 3: 異なる温度パラメータ (Temperature) を指定した場合の尖度 (Kurtosis) と τ (Kendall Tau) の変化

全ての pre モデルと llama の post モデルは温度パラメータの影響をほとんど受けない．一方で，gemma と mistral では，温度パラメータが大きいほど，post モデルの尖度は小さい．つまり，アライメントによる数値バイアスへの影響を抑えたい場合は，温度パラメータを高く設定することが緩和策として挙げられる．一方で，gemma の post モデルの結果に見られるように，温度パラメータが高すぎるとモデルの評価精度が低下する可能性があるため，この場合は 1.0 が適切な値であると考えられる．

5 おわりに

本研究では，LLM のアライメントが数値バイアスに与える影響を調査し，特定のスコアへの過度な集中を引き起こす数値バイアスが観測された．数値バイアスは特に，高資源言語の評価で顕著であった．また，数値バイアスの強化が評価精度の低下につながることが示された．これらの結果は，アライメントによる過剰な一貫性の向上が，品質評価タスクでの実用性を制約する可能性を示唆している．今後は，アライメント強度ごとの検証やバイアス緩和手法の開発などが必要である．本研究の知見は，LLM を評価者として利用する際の課題解決に向けた指針となることが期待される．

参考文献

[1] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In **Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2023.

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

[3] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[4] Tom Kocmi and Christian Federmann. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, **Proceedings of the Eighth Conference on Machine Translation**, pp. 768–775, Singapore, December 2023. Association for Computational Linguistics.

[5] Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In **Proceedings of the 24th Annual Conference of the European Association for Machine Translation**, pp. 193–203, 2023.

[6] Ayako Sato, Kyotaro Nakajima, Hwichan Kim, Zhousi Chen, and Mamoru Komachi. TMU-HIT's submission for the WMT24 quality estimation shared task: Is GPT-4 a good evaluator for machine translation? In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, **Proceedings of the Ninth Conference on Machine Translation**, pp. 529–534, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[7] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, **Advances in Neural Information Processing Systems**, Vol. 35, pp. 27730–27744. Curran Associates, Inc., 2022.

[8] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, **Proceedings of the 40th International Conference on Machine Learning**, Vol. 202 of **Proceedings of Machine Learning Research**, pp. 29971–30004. PMLR, 23–29 Jul 2023.

[9] Behnam Mohammadi. Creativity has left the chat: The price of debiasing language models, 2024.

[10] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? 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. 15607–15631, Toronto, Canada, July 2023. Association for Computational Linguistics.

[11] Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators, 2024.

[12] Masanari Ohi, Masahiro Kaneko, Ryuto Koike, Mengsay Loem, and Naoaki Okazaki. Likelihood-based mitigation of evaluation bias in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics ACL 2024**, pp. 3237–3245, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics.

[13] Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. Findings of the WMT 2020 shared task on quality estimation. In Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors, **Proceedings of the Fifth Conference on Machine Translation**, pp. 743–764, Online, November 2020. Association for Computational Linguistics.

[14] Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. MLQE-PE: A multilingual quality estimation and post-editing dataset. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 4963–4974, 2022.

[15] Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 1460–1474, 2021.

A　詳細な実験設定

A.1　MTQE のスコアの種類

MTQE には Multidimensional Quality Metric（MQM）[15] と Direct Assessment（DA）[14] の 2 つの異なるアノテーション手法が存在する. MQM は DA よりも複雑な品質推定手法であり，誤りの検出および重要度評価から構成される．アライメント前のモデルの指示追従能力が低いため，本研究においては，この手法の実現は困難である．よって，本研究では，DA スコアを採用する. アライメント後の GPT モデルにおける MQM 性能は [4] で報告されている.

A.2　実験で使用した評価者 LLM

本研究で, 評価者として使用するオープンソース LLM の一覧を 表 3 に示す.

Series	Pre-alignment	Post-alignment
Gemma	gemma-7b	gemma-7b-it
Mistral	Mistral-7B-v0.1	Mistral-7B-Instruct-v0.1
Llama 3	Meta-Llama-3-8B	Meta-Llama-3-8B-Instruct

表 3: 実験で使用した LLM 評価者一覧.

A.3　プロンプト

本研究で用いる MTQE タスクのプロンプトテンプレートを 4 に示す．このテンプレートは，タスクの説明，スコア範囲，および元のデータに基づいた評価基準を含むように設計されている．出力を数値のみに限定するために，“Do not provide any explanations or text apart from the score.” と明示的に記述した.

Please analyze the given source and translated sentences and output a translation quality score on a continuous scale ranging from {{min score}} to {{max score}}.

Translation quality should be evaluated based on both fluency and adequacy.

A score close to {{min score}} indicates a low quality translation, while a score close to {{max score}} indicates a high quality translation.

Do not provide any explanations or text apart from the score.

{{source language}} Sentence: {{src_i}}

{{target language}} Sentence: {{hyp_i}}

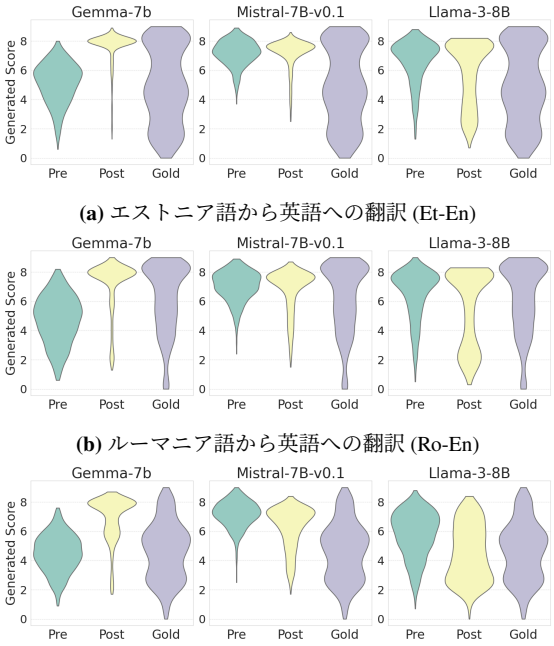
Score:

表 4: MTQE のプロンプトテンプレート.

B　実験結果の補足

B.1　他の言語ペアにおける分布

4.1 節の実験における，Et-En，Ro-En，Si-En での MTQE スコア分布を図 4 に示す.



(a) エストニア語から英語への翻訳 (Et-En)

(b) ルーマニア語から英語への翻訳 (Ro-En)

(c) シンハラ語から英語への翻訳 (Si-En)

図 4: Et-En，Ro-En，Si-En の MTQE スコア分布.

B.2　mistral における温度パラメータの影響

4.2.2 節の実験における，mistral モデルでの実験結果を 5 に示す.

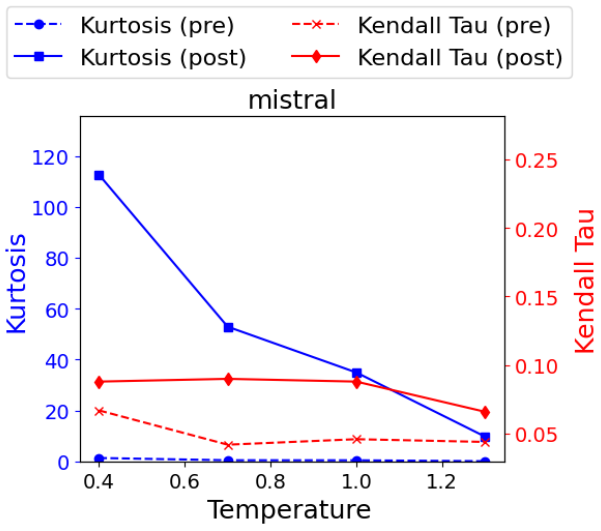


図 5: 異なる温度パラメータ (Temperature) を指定した場合の尖度 (Kurtosis) と τ (Kendall Tau) の変化