大規模言語モデルにおける評価バイアスの尤度に基づく緩和

大井聖也 [1] 金子 正弘 [2,1] 小池 隆斗 [1] Mengsay Loem [1] 岡崎 直観 [1]
[1] 東京工業大学 [2] MBZUAI
{masanari.ohi@nlp., ryuto.koike@nlp., mengsay.loem@nlp., okazaki@}c.titech.ac.jp
masahiro.kaneko@mbzuai.ac.ae

概要

大規模言語モデル (Large Language Model; LLM) は言語生成タスクの評価器として用いられている。ところが、ある文章の意味を変えずに語順や構造を変更した文章を作ると、LLM が計算する尤度が大きく変化することがある。そのため、LLM 評価器には、尤度が低い文章を不当に低く、尤度が高い文章を不当に高く評価する**尤度バイアス**が存在すると考えられる。本研究では、尤度バイアスが LLM 評価器の性能を低下させることを明らかにし、Few-shot によるバイアス緩和手法を提案する。実験では、複数の LLM が data-to-text タスクと文法誤り訂正タスクで尤度バイアスを持つ可能性を示し、その緩和に成功した。

1 はじめに

LLM は優れた言語理解能力と文章生成能力を示し、最近は文生成タスクの自動評価手法としても活用されている [1, 2, 3, 4]。例えば、評価対象の文章の尤度を LLM に計算させ、評価スコアとして使用する方法 [2, 5] や、LLM に文章の評価スコアを直接出力させる方法 [1, 3] が提案されている。BLEU [6] や ROUGE [7] などの従来の自動評価手法と比べ、LLM による自動評価は多くのタスクで人間の評価とより高い相関を示すことが報告されている。LLM の学習は膨大な事前学習データ [8, 9] と指示学習データ [10, 11] の尤度最大化であり、文章生成も尤度に基づいている。ゆえに、尤度を直接的に評価スコアとする方法だけではなく、評価スコアを生成させる方法においても、評価対象の文章の尤度が評価結果に影響を与えると考えられる。

ところが、LLM が計算する尤度は文章の流暢性や文法性、意味などの良し悪しを捉えているとは限らない。例えば、ある文章の語順や構造を変更して言い換えると、LLM の尤度が変動することが報告されている [12]。この場合、文章の意味に関する評価を行う際に、尤度の変動が LLM の評価結果に悪影響を及ぼしうる。言い換えると、LLM の尤度と文章の良し悪しのずれが、さまざまな評価項目で評価バイアスを引き起こしている可能性がある。本研究では、LLM が文章の評価スコアを出力する際、尤度の低い文章を（人間の評価よりも）不当に低く評価し、尤度の高い文章を不当に高く評価するという評価バイアスの存在を仮定し、これを**尤度バイアス**と呼ぶ。尤度バイアスの例を図 1 に示す。この図ではデータから文を生成するタスク（data-to-text）において、人間の採点であれば同スコアになる出力のうち、尤度の低い出力（上側）が尤度の高い出力（下側）よりも不当に低く評価されていることを表す。

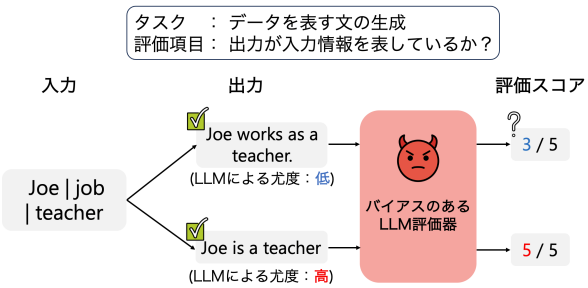

図 1: 尤度バイアスの例。人間の採点であれば同じスコアをつけられるべき出力のうち、尤度の低い出力（上側）が尤度の高い出力（下側）よりも不当に低く評価されていることを表す。

この問題に対処するため、我々は尤度バイアスを (1) 定量的に測定し、(2) 緩和する手法を提案する。本研究は LLM 評価器における評価時のバイアスを緩和する初めての試みである。まず、LLM の尤度と、LLM と人手評価のスコアの差の相関に着目し、尤度バイアスの定量的な測定を行う。実験の結果、GPT-3.5 と Llama2 13B [13] の 2 つの LLM が data-to-text と文法誤り訂正タスク (GEC) の 2 つのタスクにおいて尤度バイアスを持つ可能性を示した。

次に、訓練データから尤度バイアスの強い事例（タスクの入出力のペア）を特定し、それらの事例に人手評価スコアを付与し、Few-shot 事例として LLM 評価器に与えることで、評価時の尤度バイアスを緩和する。提案手法により、ほとんどのモデル・タスクにおいて LLM の尤度バイアスが緩和され、評価性能（人手評価スコアとの順位相関係数）も向上することが分かった。

2 提案手法

先行研究 [1, 4] に倣い、本研究では LLM に文章の評価を指示するプロンプトを与え、評価スコアを計算する。また、Liu ら [1] に倣い、LLM に評価スコアを直接出力させるのではなく、評価スコアの候補値（例: $\{1, \ldots, n\}$）を予測させ、その尤度からスコアの期待値を計算し、最終的な評価スコアとする（これを Score_m と書く）。先行研究では、タスクの説明・評価項目・評価対象文章の 3 つでプロンプトを構成していたが、我々はこれに加えて Few-shot 事例をモデルに与えることで出力を安定化させ、より正確に尤度バイアスを測定・緩和することを目指す[1)]。

2.1 尤度バイアスの測定

本研究では、**人間の評価と比較して LLM が尤度の低い文章を不当に低く評価し、尤度の高い文章を不当に高く評価する**という評価のバイアスを**尤度バイアス**と呼ぶ。まず、その定量的な測定方法を提案する。t_i を入力文章、t_o を出力文章とし、これらをまとめて評価対象の事例 $t = (t_i, t_o)$ と書く。d をタスクの説明、θ をモデルのパラメータ、P をモデルによる尤度とすると、LLM の尤度による指標 LS (likelihood score) は次式で計算される。

$$\text{LS}(t) = \log P(t_o \mid t_i, d; \theta) \tag{1}$$

次に、LLM がどれだけ不当な評価をしているかを表す指標 US (unfairness score) として、LLM による評価スコア (Score_m) と人手評価スコア (Score_h) の差を次式で計算する[2)]。

$$\text{US}(t) = \text{Score}_\text{m}(t; \theta) - \text{Score}_\text{h}(t) \tag{2}$$

LLM による評価スコア Score_m を計算するときは、訓練データからランダムに選んだ Few-shot 事例をプロンプトに含める。

1) Score_m を計算するために用いた実際のプロンプトや計算式、Few-shot の事例数などを付録 A に記載する。
2) LLM と人手によるスコアの値域が異なる場合を考慮し、Score_m と Score_h は同じ値域になるように正規化する。

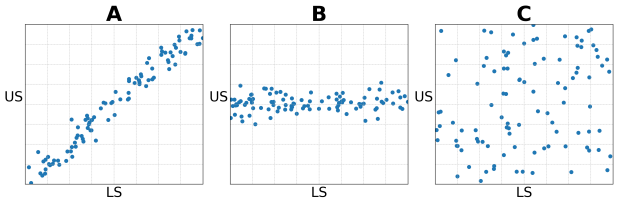

図 2: 仮想の評価器における尤度バイアスを可視化した図。A はバイアスがある、B はバイアスがなく高性能、C はバイアスがなく低性能な評価器を表す。

LS、US を用いて LLM の持つ尤度バイアスの強さを示す指標である **BiasScore** を計算する。BiasScore は、事例の集合であるデータセット $D = \{t^{(1)}, t^{(2)}, \ldots, t^{(n)}\}$ 全体における LS と US のスピアマンの順位相関係数 ρ として計算する。

$$\text{LS}_\text{D} = [\text{LS}(t^{(1)}), \text{LS}(t^{(2)}), \ldots, \text{LS}(t^{(n)})] \tag{3}$$

$$\text{US}_\text{D} = [\text{US}(t^{(1)}), \text{US}(t^{(2)}), \ldots, \text{US}(t^{(n)})] \tag{4}$$

$$\text{BiasScore} = \rho(\text{LS}_\text{D}, \text{US}_\text{D}) \tag{5}$$

BiasScore は [-1, 1] の範囲を取り、1 は最も強いバイアスを表し、0 はバイアスがないことを表す。

2.2 尤度バイアスの緩和

図 2 に複数の仮想の評価器による尤度バイアスを可視化した散布図を示す。横軸は LS (式 1)、縦軸は US (式 2)、点は評価事例、点の回帰直線の傾きが BiasScore を表す。各図は次のように解釈される。

- A の散布図は尤度バイアスを持つ評価器を表す。評価器は尤度の高い事例に不当に高いスコアを与えており（右上部分）、尤度の低い事例に不当に低いスコアを与えている（左下部分）。バイアスの緩和前の LLM 評価器はこの状態であると想定する。
- B の散布図は尤度バイアスを持たない評価器を表す。US が 0 に近いため、評価性能は高いと考えられる。
- C の散布図も尤度バイアスを持たない評価器を表すが、US の値がランダムに分布しているため、評価性能は低いと考えられる。

提案手法による尤度バイアスの緩和では、バイアスのある状況 (A) から、評価性能が高くバイアスのない状況 (B) に変化させることを目指す。その際、評価性能が低くバイアスのない状況 (C) にならないように、バイアスが強い事例（A の図の右上・左下）を特定し、その評価を集中的に是正する。そのため、事例 t のバイアスの強さを示す指標として、RS

を導入する。

$$\mathrm{RS}(t) = |\mathrm{LS}^*(t) + \mathrm{US}^*(t)| \tag{6}$$

ここで、LS^* と US^* はそれぞれ、データセット D に対して平均値が 0、値域が $[-1, 1]$ になるように LS と US を正規化したものである。$\mathrm{RS}(t)$ は事例 t が図 2 の A において右上か左下に近づくほど大きな値を取る。評価バイアスの軽減のため、訓練データから $\mathrm{RS}(t)$ が大きな事例、つまりバイアスの強い事例を抽出し、人手評価のスコアを付与したうえで Few-shot 事例として用いることで、評価における尤度バイアスを緩和する。

3 実験結果

3.1 使用したタスク・データ

実験では、data-to-text と GEC の 2 つのタスクにおける尤度バイアスの測定・緩和を行う。data-to-text タスクでは、WebNLG+ [14] をデータセットとして用いた。2846 個の英語の各事例に対して、text structure, relevance, fluency, correctness, data coverage の 5 つの評価項目に関する人手評価スコア $\mathrm{Score_h}$ が付与されている。GEC では、TMU-GFM-Dataset [15] をデータセットとして用いた。4221 個の英語の各事例に対して、grammar, fluency の 2 つの評価項目に関して人手評価スコア $\mathrm{Score_h}$ が付与されている[3)]。

これらのデータはそれぞれ、4:1 の割合で訓練・評価データに分割した。また、タスク全体での傾向を調べるために、すべての評価項目のマイクロ平均を新たな評価項目 total として導入した。

3.2 使用した LLM

実験には、OpenAI 社が API として提供する GPT-3.5[4)] と、オンプレミスで動作する Llama2 13B (L-13B) [13] を LLM として用いた。ただし、GPT-3.5 はトークンの尤度を出力しないため、代わりに Llama2 13B で尤度を計算する。まず、これらの LLM の評価器としての性能を確かめるために、人手と LLM による評価スコアのスピアマンの順位相関係数を計算した。結果を表 1, 2 の「評価性能」の「緩和前」列にそれぞれ示す。全体的な傾向として、data-to-text では GPT-3.5 が Llama2 13B よりも高い性能を示し、GEC ではどちらのモデルも同程度の性能を示した。

3.3 尤度バイアスの測定

2.1 節で述べた手法を用い、data-to-text と GEC の評価データにおける尤度バイアスを測定した。

data-to-text での結果 表 1 の「BiasScore」の「緩和前」列における数値はほとんどのモデル・評価項目で 0.20 を超える結果を示している。BiasScore は相関係数であり、0.20 は弱相関を表すため、この結果は尤度バイアスの存在を示唆していると言える。評価項目 total においては GPT-3.5 (0.38) が Llama-2 13B (0.17) よりも大きな BiasScore を示しており、他の評価項目においても同じ傾向が観察される。また、評価項目 relevance がどちらのモデルにおいても全ての評価項目の中で最も大きな値を示した。

GEC での結果 表 2 の「BiasScore」・「緩和前」列における数値はすべてのモデル・評価項目で 0.20 を超える結果になっており、data-to-text と同様に尤度バイアスの存在を示唆している。また、評価項目 total においても同様に GPT-3.5 (0.43) が Llama2 13B (0.21) よりも大きな BiasScore を示しており、他の評価項目についても同じ傾向が観察される。

評価項目ごとの尤度バイアスの比較 data-to-text の評価項目ごとの尤度バイアス（表 1 の「BiasScore」・「緩和前」列）に着目すると、どちらのモデルにおいても fluency と text structure が比較的小さな BiasScore を示している。これらの項目は、入力によるタスク固有の制約は考慮せず、生成された文章に内在する自然さや文構造のような特性のみに基づいて評価する。そして、どちらのモデルにおいても relevance と data coverage は、BiasScore が小さい内在的評価の項目とは対照的に大きな BiasScore を示している。これらの項目は、生成された文章に対して外在する入力との関連性や情報の有無などのタスク固有の制約を用いて評価する。これらのことから、内在的評価の項目は、外在的評価の項目よりも小さい尤度バイアスを持つといえる。出力の尤もらしさは内在的な特性に強く依存することを鑑みると、文章の尤度と内在的評価による文章の優劣には正の相関があると予想される。よって、内在的評価においては、LLM 評価器が尤度に影響を受けていることが、過小または過大評価のように必ずしも悪い結果につながらないと考えられる。一方で、評価に寄与する特性が異

3) データセットやその評価項目の説明を付録 B に示す。GEC のデータセットでは 3 つ目の評価項目として meaning が存在するが、データセットを作成した研究 [15] で meaning における結果が全体の評価にほとんど寄与しないことが示されているため、本実験からは除外した。

4) `gpt-3.5-turbo-instruct` を API 呼び出しに用いた。

表 1: data-to-text での尤度バイアス緩和前後の BiasScore と評価性能。緩和後の値における太字はバイアスの緩和が狙い通りに作用したことを、* はバイアス緩和前後での並べ替え検定による有意な差 （ $p < 0.05$) を、† は有意な傾向の差 ($p < 0.06$) を表す。

	BiasScore				評価性能 ρ			
	緩和前		緩和後		緩和前		緩和後	
評価項目	GPT-3.5	L-13B	GPT-3.5	L-13B	GPT-3.5	L-13B	GPT-3.5	L-13B
text structure	.36	.17	**.23** *	**.02** *	.46	.34	**.53** †	**.36**
relevance	.43	.28	**.31** *	**.15** †	.35	.25	**.38**	.23
fluency	.26	.20	.29	**.00** *	.41	.33	**.55** *	**.52** †
correctness	.36	.21	**.32**	**-.01** *	.44	.37	**.47**	**.43**
data coverage	.40	.24	**.32** *	**.16**	.20	.24	**.30** †	**.25**
total (マイクロ平均)	.38	.17	**.32** †	**.02** †	.48	.40	**.58** *	**.46**

表 2: GEC での尤度バイアス緩和前後の BiasScore と評価性能。太字、 * 、 † は表 1 と同じ意味で使用した。

	BiasScore				評価性能 ρ			
	緩和前		緩和後		緩和後		緩和後	
評価項目	GPT-3.5	L-13B	GPT-3.5	L-13B	GPT-3.5	L-13B	GPT-3.5	L-13B
grammar	.46	.24	**.37** †	.24	.48	.45	**.54**	**.46**
fluency	.36	.16	**.29**	**.09**	.40	.49	**.47**	.48
total (マイクロ平均)	.43	.21	**.37**	**.18**	.45	.48	**.52**	**.52**

なるため、外在的評価では尤度からの影響が LLM 評価器に悪影響を及ぼしていると考えられる。

3.4 尤度バイアスの緩和

尤度バイアス緩和のため、2.2 節で説明した手法を用いて訓練データからバイアスを持つ事例を 8 つ取得し、Few-shot 事例として用いた。表 1, 2 の「緩和後」列にバイアス緩和後の BiasScore と評価性能を示す。太字はバイアスの緩和が狙い通りに作用した（BiasScore の絶対値が減少した・評価性能が向上した）ことを表す。また、$R = 100000, \alpha = 0.05$ として並べ替え検定を行い、バイアス緩和前後で有意な差 ($p < 0.05$) が確認されたものを* 、有意な傾向の差 ($p < 0.06$) が確認されたものを † で表した。

data-to-text での結果 表 1 の「BiasScore」と「評価性能 ρ」の「緩和後」列における数値から、提案手法によってほとんどのモデル・評価項目で BiasScore の絶対値が減少し、同時に評価性能が向上したことがわかる。GPT-3.5 では text structure (-0.13), relevance (-0.12), data coverage (-0.08) において、Llama2 13B では text structure (-0.15), fluency (-0.20), correctness (-0.20) において BiasScore の絶対値が有意に減少している。同時に、GPT-3.5 では fluency (+0.14), total (+0.10) において評価性能が有意に向上している。また、GPT-3.5 で text structure (+0.07), data coverage (+0.10) において、Llama2 13B で fluency (+0.19) において評価性能の向上に有意な傾向が確認された。したがって、モデル・タスク全体を通して尤度バイアスの緩和が狙い通り成功し、それによって評価性能が向上していることが確認できた。

GEC での結果 表 2 の「BiasScore」と「評価性能 ρ」の「緩和後」列における数値から、提案手法によってほとんどのモデル・評価項目で BiasScore の絶対値が減少し、同時に評価性能が向上したことがわかる。有意な傾向のバイアス緩和が確認されたのは GPT-3.5 の grammar (-0.09) のみだったが、少なくとも提案手法がバイアスの緩和を促進し、評価性能を向上させる影響を与えたと考える。

以上の結果から、尤度バイアスを緩和する提案手法が data-to-text と GEC における LLM 評価器の尤度バイアスを緩和し、同時に評価性能を向上させることに成功したと考える。

4 おわりに

本研究では LLM が尤度の低い文章を不当に低く、尤度の高い文章を不当に高く評価する傾向を尤度バイアスとして定義し、定量化する方法を提案した。また、我々はバイアスの強い事例を特定し、Few-shot 事例として用いることで尤度バイアスを緩和する方法を提案した。実験の結果、data-to-text、GEC の 2 つのタスクにおいて複数の LLM が尤度バイアスを持つ可能性を示した。さらに、尤度バイアスを緩和し評価性能を向上することに成功した。今後は、他モデルでの尤度バイアスの測定や、バイアスが強い事例を抽出する別手法の検討、ファインチューニングを用いるバイアスの緩和などに取り組みたい。

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（22501）により得られたものです。また、本研究は JSPS 科研費 19H01118 の助成を受けました。論文執筆にあたっては、ケンブリッジ大学の Simone Teufel 先生から助言をいただきました。

参考文献

[1] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023.

[2] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire, 2023.

[3] Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, **Proceedings of the 24th Annual Conference of the European Association for Machine Translation**, pp. 193–203, Tampere, Finland, June 2023. European Association for Machine Translation.

[4] 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.

[5] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, **Advances in Neural Information Processing Systems**, Vol. 34, pp. 27263–27277. Curran Associates, Inc., 2021.

[6] 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.

[7] 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.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[9] OpenAI. Gpt-4 technical report, 2023.

[10] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In **International Conference on Learning Representations**, 2022.

[11] 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.

[12] Tatsuki Kuribayashi, Takumi Ito, Jun Suzuki, and Kentaro Inui. Language models as an alternative evaluator of word order hypotheses: A case study in Japanese. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 488–504, Online, July 2020. Association for Computational Linguistics.

[13] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[14] Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina, editors, **Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)**, pp. 55–76, Dublin, Ireland (Virtual), 12 2020. Association for Computational Linguistics.

[15] Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, and Mamoru Komachi. SOME: Reference-less sub-metrics optimized for manual evaluations of grammatical error correction. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 6516–6522, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

A　LLM 評価器の設定

尤度の計算　式 1 に示す通り、我々はモデルの文章に対する尤度を、タスク説明 d、タスク入力 t_i で条件付けしたタスク出力 t_o の対数尤度として計算する。単純に条件付けを行わず $\log P(t_o;\theta)$ を計算するよりも、条件付けを行うことでタスクと入力に関する情報を考慮した尤度が得られることが期待できる。タスク説明は以下の文章を用いた。

- data-to-text: Please generate a description of the following xml data.
- GEC: Please modify the following English text to make it grammatically correct.

Score_m の計算方法　先行研究 [1, 4] に倣い、我々はタスク説明・評価項目で構成されるプロンプト I と評価対象の事例 t を LLM に与えることで、LLM による評価スコア Score_m を計算する。我々はこれに加えて Few-shot 事例 F を LLM に与えることでモデルの出力を安定させることを狙う。F は、バイアス測定の際はランダムに、バイアス緩和の際はバイアスの強い事例を訓練データから 8 つ抽出して用いる。また、先行研究 [1] に倣い、LLM がスコアを出力する確率を用いてスコアの期待値を計算し、それを Score_m として用いる。LLM に直接スコアを出力させる代わりにスコアの期待値を計算することで、スコアが 1 つに集中せず、より詳細な値が得られることが期待できる。スコアの候補を $\{1, 2, \ldots, n\}$、スコア i を LLM が出力する確率を $Q(i \mid t, F, I; \theta)$ とすると、Score_m は以下のように計算される。

$$\text{Score}_\text{m}(t;\theta) = \frac{\sum_{i=1}^{n} i \times Q(i \mid t, F, I; \theta)}{\sum_{j=1}^{n} Q(j \mid t, F, I; \theta)} \tag{7}$$

また、タスク説明と評価項目を含んだプロンプトの例を以下に示す。

data-to-text で correctness を評価するプロンプト

You will be given an xml data and an English sentence that represents xml data. Your task is to rate the sentence that represents xml data on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria: Correctness: (1-5) - does the text describe predicates with correct objects and does it introduce the subject correctly? 1 is the lowest score, 5 is the highest.

GEC で fluency を評価するプロンプト

You will be given an English sentence that may have grammatical errors and a sentence that is the corrected version of the sentence. Your task is to rate the corrected sentence on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria: Fluency: (0-4) - How natural the sentence sounds for native speakers; 4: Extremely natural, 3: Somewhat natural, 2: Somewhat unnatural, and 1: Extremely unnatural, and 0: Other.

B　データセット

data-to-text　data-to-text では、WebNLG+ [14] (CC BY-NC-SA 4.0) から人手評価スコアが付与されている事例を抜き出し、データセットとして用いた。2846 個の英語の各事例に対して、以下の 5 つの評価項目における人手評価スコア Score_h が 0 から 100 までの 1 点刻みで付与されている。

- text structure: 出力が文法的に正しく構成されているかどうか
- relevance: 出力が入力データに基づいているかどうか
- fluency: 出力が自然な文章かどうか
- correctness: 出力が入力データを正しく説明しているか
- data coverage: 出力が入力データの情報を全て含んでいるか

GEC　GEC では、TMU-GFM-Dataset [15] (CC BY 4.0) をデータセットとして用いた。4221 個の英語の各事例に対して、以下の 3 つの評価項目における人手評価スコア Score_h が 0 から 4 までの 1 点刻みで付与されている。

- grammar: 出力が文法的に正しいか
- fluency: 出力が自然な文章かどうか
- meaning: 出力が入力と同じ意味を持つか

3.1 節の脚注で述べたように、データセットを作成した研究 [15] で meaning はタスク全体の評価スコアにほとんど寄与していないことが明らかにされているので、本実験からは除外した。