長文生成の多面的評価: 人手評価と自動評価の向上を目指して

鴨田豪 [1] 浅井明里 [2] Ana Brassard[3,1] 坂口慶祐 [1,3]
[1] 東北大学 [2] ワシントン大学 [3] 理化学研究所
go.kamoda@dc.tohoku.ac.jp akari@cs.washington.edu
ana.brassard@riken.jp keisuke.sakaguchi@tohoku.ac.jp

概要

大規模言語モデル (LLM) は幅広いタスク目覚ましい成長を遂げているが，情報検索クエリに対する長文応答の評価は依然として困難である．本研究ではそのような長文生成に対して網羅的な評価基準を定め，人手評価と自動評価の質を高める．評価には 4 つの軸を設け，それぞれに対し絶対評価を行う．この枠組みに従い，人間の回答と LLM が生成した回答に対して総計 3,600 件の人手評価を収集し，総合評価に最も影響を与える重要な評価軸を明らかにする．更に，従来の手法を人手評価との相関で凌駕する，LLM を利用した自動評価手法を確立する．

github.com/gokamoda/LFQA-MultiAspectEval

1 はじめに

質問応答タスクは大きく Factoid 型と Non-Factoid 型に分けられる．「日本の首都は？」のような回答が名詞 1 単語など一言で回答できるようなタスクを Factoid 型とよび，「なぜ空は青い？」のような複数文での説明を要するようなタスクを Non-Factoid 型と呼ぶ．Factoid 型の質問応答タスクは特に Mask 言語モデルで研究の対象とされてきた [1, 2, 3]．一方で，大規模言語モデル (LLM) は様々なタスクで著しい成長を遂げており [4]，情報検索クエリに対して流暢でもっともらしい長文回答が生成できる様になった [5]．これより，Non-Factoid 型質問応答タスクへの注目が集まるようになった [6, 7]．現在の言語モデルは高品質な回答を出力できるとされるが，それらの評価は依然として課題の一つである [8]．広く使われている自動評価指標である ROUGE は [9] 模範解答との使用語彙の一致度を利用するが，簡単にだしぬけてしまう [10]．信頼できる評価を行うために，昨今では 2 つの回答の候補を見せて全体的な良さを比較する一対比較型の人手評価が行われてきた [11, 12, 13, 14, 8, 15]．しかしこの手法では相対的評価にしかならず，また評価対象がどの点で優れているか (劣っているか) は明らかにならない．

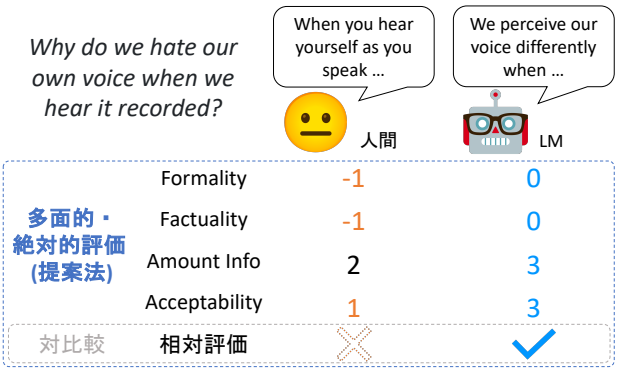

図 1 LFQA タスクではこれまで一対比較型の評価が行われてきた．本研究では多面的な絶対評価手法を行う．

本研究では網羅的な評価基準を新たに定め，人手評価と自動評価のそれぞれに対してその妥当性を示す．まず，これまで行われてきた一対比較型の相対評価に替わり，LFQA タスクに特化した 4 つの評価軸 (Formality, Factuality, Amount Info, Acceptability) を設け，事前に定義された評価尺度にしたがって絶対的評価を行う (図 1)．

新たな評価基準に従い，LFQA タスクで広く使われている ELI5 データセットの一部と ChatGPT を用いて生成した回答の総計 1,200 の回答に対して人手評価を行う．結果，相対評価では 83%のケースで ChatGPT の回答が好まれたが，これらの回答が絶対評価において満点をとっている割合は半数にとどまった．また，ChatGPT の回答でも知識の正しさの面で改善の余地があることが明らかになり，多面的絶対評価を行うことの妥当性が示された．

自動評価では LLM を使用した LLM-Mate (**M**ulti-**A**spec**T** **E**valuation) を提案する．LLM-Mate では評価軸ごとの評価を LLM で予測し，その重み付き和

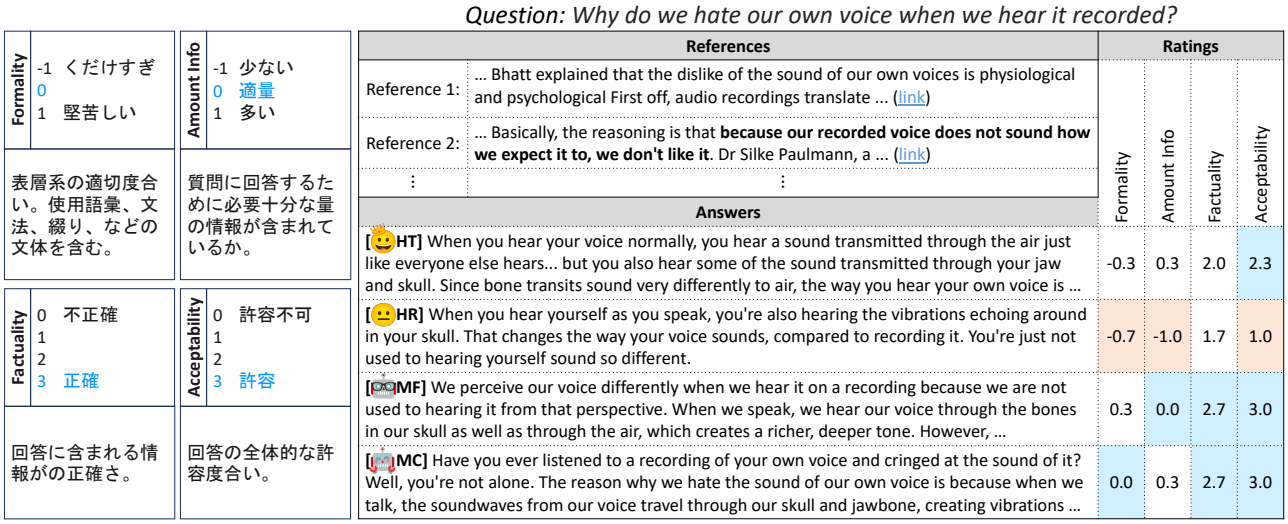

図 2 評価軸と尺度を左に示す．右に評価を行う質問・回答と，それに対する評価の例を示す．青いセルは理想スコアを示し，赤いセルは悪いスコアを示す．HT, HR, MF, and MC はそれぞれ Human Top, Human Random, Model Formal, Model Casual を指す．一つの回答に対して 3 件の評価データを収集する．

で最終評価を決定する．重みは人手評価結果をもとに計算され，GPT-4 と追加学習させた Llama2-7B を使用した結果，人手評価との相関は 0.7 を超え，従来の評価指標を凌駕した．

2 LFQA タスクの評価

多面的絶対評価 従来の一対比較型の評価に代わり，多面的かつ絶対的な評価を行う．

多面的人手評価は説明生成タスク [16] や文書要約タスク [17] で採用されている．これらを参考に，LFQA タスク特化の多面的評価軸を決定する．本研究では Formality, Factuality, Amount Info, Acceptability の 4 軸を用意する (図 2 左)．Formality は回答の表層に着目し，使用語彙や文構造，綴の正しさ等の適切度合いを測る．回答の内容を評価する軸としては Factuality と Amount Info を用意し，それぞれ回答に含まれる知識や事実の正しさと量を測る．最後に回答の総合的な評価を Acceptability で評価する．ここで総合的な高評価を得る回答は最低限，「自己完結的であり，かつ専門外の読み手にも伝わらなければならない」とした．これらの評価軸は．それぞれの回答がどの面で優れているか (伸びしろがあるか) の分析を可能にする．

絶対的評価では，複数の回答を比較した相対評価ではなく，評価尺度に応じた絶対値での評価を行う．Factuality と Acceptability では 0 から 3 の整数値を取る尺度を採用する．一方 Amount Info と Formality ではそのバランスが重要なため -1 から 1 の整数値を取る尺度を採用する．

評価対象 広く使われている LFQA データセットである ELI5 から 300 の質問を使用する．それぞれの質問に対して 4 つ，計 1,200 件の回答を評価対象とする．4 つの回答のうち 2 つは ELI5 に収録されている，回答最もスコアの高い回答 (**HT; Human-Top**) と，それ以外の回答をランダムに選択したもの (**HR; Human-Random**) を使用する．残り 2 つは ChatGPT による回答で，一つは簡単なプロンプトで生成させ (**MF; Model-Formal**)，もう一つはより砕けた表現をするように生成させる (**MC; Model-Casual**).

3 人手評価

2 節に従い，Amazon Mechanical Turk 上で人手評価を行った．1,200 の回答について各 3 件，合計 3,600 件の評価データを収集した．アノテータには多面的絶対評価の他に，既存研究が行ってきた相対的評価と合わせて，「評価の根拠」の記述を依頼した．依頼時には，特に Factuality の正確性向上のために Google Seach API [1] で収集した参考文献 10 件を見せた．アノテータ一致度の κ は 0.74 で，計$5,064 支払った．収集した評価データは公開する．このデータは 283 件の評価対象に対して 1 件ずつ評価した Krishna ら [10] や，260 件に対して 1-2 件ずつ評価した Xu ら [8] らのデータよりも大規模であり，参考文献を事前に与えた上で Nakano ら [15] のように外部ツールの利用を制限しなかった点で優れている．

1) https://serpapi.com/

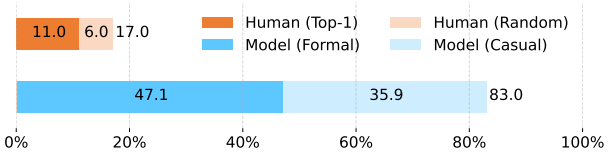

図 3　相対評価の結果．ChatGPT の回答が好まれるケースが 83%だった．

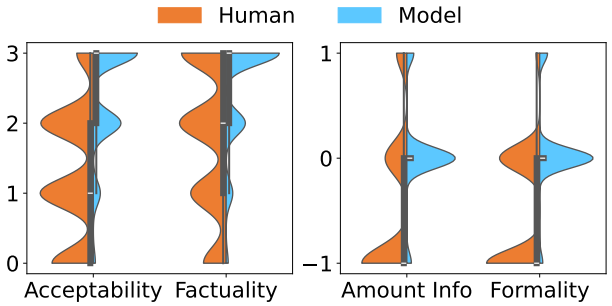

図 4　評価軸ごとの評価分布．理想のスコアは左図で 3，右図で 0 である．

3.1　結果

まず，既存研究に倣って相対評価の結果を見る．図 3 に示すように，83%のケースで ChatGPT の回答が好まれた．この結果は Xu らの結果 [8] と概ね一致する．次に多面的絶対評価の結果を見る．図 4 に多面的評価の分布を示す．Acceptability に着目すると，**HT**，**HR** 回答の平均は 1.38, 1.15 であったのに対し，**MF**，**MC** 回答は 2.46，2.36 だった．相対評価では全体の 83%で ChatGPT の回答が好まれたが，これらの回答が絶対評価において最高評価を受ける割合は 54.4%にとどまった．このことから多面的に絶対尺度で評価を行うことの重要性が示唆される．更に他の評価軸に着目すると，人間が作成した回答は情報量が少なく，砕けた文体で書かれる傾向があることがわかる．一方 ChatGPT の回答は，適度な情報量と文体で書かれることがわかる．また，ChatGPT の回答について，Factuality では他の評価軸に比べて評価の標準偏差が大きく，0.75 となった．

次に，総合的評価軸である Acceptability に，他の軸が与えている影響を測るため，線形回帰モデルを学習する．まず減点方式に変換するために入力を標準化する (式 1, 2)．次に回帰モデルの重み $\mathbf{w}$ を学習させる (式 3)．

$$f\left(\begin{bmatrix}\text{Formality}\\ \text{Amount Info}\\ \text{Factuality}\end{bmatrix}\right)\begin{bmatrix}(\text{Formality}-3)/3\\ -|\text{Amount Info}|\\ -|\text{Factuality}|\end{bmatrix} \tag{1}$$

$$y_{\text{Acce}} = \text{Acceptability} - 3 \tag{2}$$

$$y_{\text{Acce}} = \mathbf{w}\cdot f(\mathbf{x}) \tag{3}$$

結果，w_{Form}, w_{Info}, and w_{Fact} はそれぞれ 0.335, 0.739, 2.048 となり，Factuality が最も重要であり，Formality の軸で測った文体の影響は小さいことがわかる．モデルは全体の 80%のデータで学習され，残りの 20%では Pearson 相関係数が 0.853 となり，手法の妥当性が示された．

その他，収集した「評価の根拠」を分析すると，評価に使用した 3 軸の他に「完全さ」「わかりやすさ」「具体性」などの言葉が多く使用されていた．

4　自動評価

4.1　LLM-Mate

ELI5 データセットの評価指標としては，模範解答との類似度を計算する **ROUGE** が採用されているが [9]，人手評価との相関が低いことが知られている [10]．更に，模範解答の質に関しても問題視されている [15]．そこで，本論文では模範解答を必要としない新しい評価指標 LLM-Mate(**M**ulti-**A**spec**T** **E**valuation) を提案する．

LLM-Mate は昨今飛躍的発展を遂げている LLM を利用した多面的評価指標である．LLM を評価に使用する研究は行われているが，一対比較型の手法を採用している [8, 12, 13, 14]．本論文で提案する LLM-Mate では 2 節で紹介した多面的絶対評価の枠組みを使用し，LFQA タスクに最適化した指標である．LLM-Mate ではまず LLM に Factuality, Amount Info, Formality の 3 評価軸で独立に評価を出力させる．入力 Prompt は，各評価軸と尺度の説明 i，LFQA タスクの質問 q とその回答 r で構成する．それらを式 3 で学習した重みを使って Acceptability に変換することで総合評価を算出する[2)]:

$$\text{MATE\_score} = \mathbf{w}\cdot f(\mathbf{LLM}(i,q,r)) + 3 \tag{4}$$

評価 LLM としては，One-shot 形式で入力を与える GPT-4[18] と今回収集した人手評価データで追加学習した Llama2-7B[19] を使用する．Llama2-7B は，GPT-4 が決定的でないため用意した．

4.2　実験

LLM-Mate と他の既存評価指標とを今回収集した人手評価データとの Pearson 相関係数 ρ で比較する．

既存評価指標　前述の通り，**ROUGE** が広く使われている．模範解答としては HT 回答を使用

2)　関数 f は式 1 で定義されており，定数項 (+3) が式 2 の逆関数に対応する．

表 1 上段:🤖 MF (Model-generated Formal), 🤖 MC (Model-generated Casual), 😀 HT (Human-written Top-rated) と 😐 HR (Human-written Random-sampled) の回答の平均スコア．下段: 人手評価データの Acceptability との Pearson 相関係数．“Human Acce.” 列は，人手 Acceptability スコアの平均を示す．ROUGE 指標では，模範解答として HT 回答を使用する．

Type	Human	ROUGE (↑)			BS(↑)	GPT2-PPL (↓)			Len	GPT-4 (↑)		Llama2 (↑)	
	Acce.	1	2	L	F1	A	QA	RQA		OA	**Mate**	OA	**Mate**
🤖 **MF**	**2.46**	23.3	3.5	13.2	75.1	11.1	10.9	14.0	109	**2.94**	2.81	**2.72**	**2.99**
🤖 MC	2.36	**24.2**	**3.7**	**13.5**	**75.3**	**10.6**	**10.6**	**13.7**	107	2.91	**2.82**	2.55	**2.99**
😀 HT	1.38	-	-	-	-	31.6	27.1	33.9	**112**	2.07	1.96	1.48	1.81
😐 HR	1.16	21.0	3.1	12.4	75.2	40.2	31.1	41.1	88	1.67	1.55	1.26	1.54
Corr.	-	0.19	0.12	0.14	0.09	-0.38	-0.59	-0.54	0.22	0.70	0.72	0.72	**0.74**

する．その他の模範解答を使用する指標として，**BERT-Score (BS)** も報告する．また，文章の流暢さを測る **GPT2-PPL** も計測する．GPT2-PPL では A 設定，QA 設定と RQA 設定を用意する．A 設定では，評価対象のある質問に対する回答のみを入力として PPL を測る．QA 設定では回答の前にそれに対する質問を結合したものを，RQA 設定では回答の前にランダムな質問を結合したものを入力とする．更に，最も簡単なベースラインとして **Len** を用意し，比較対象とする．

LLM-Mate GPT-4 を使用した LLM-Mate では，出力形式を定めるために One-shot 形式でプロンプトを作成する．Llama2 を使用した設定では，3 節で収集した人手評価データで追加学習を行う．具体的には，192 (/300) の質問に対応する評価データを学習データとして使用する．多面的評価を行う **LLM-Mate** の有用性を示すために，直接総合評価 (Acceptability) を出力させる **OA** 設定も用意し，その結果を報告する．

4.3 結果

既存評価指標 これまで使用されてきた模範解答との一致度を示す ROUGE や BERT-Score では，人手評価データとの相関が 0.2 以下と低く，最も簡単な Len よりも相関がないことが示された．この結果は Krishna ら [10] の報告に合致する．また，GPT2-PPL は ROUGE 等よりも高い相関を示しているが，QA 設定と RQA 設定の差が小さい．RQA 設定では入力する質問と回答に脈略がないことからこの指標に対する疑念を抱かせる．

LLM-Mate GPT-4, Llama2 で人手評価との相関係数がそれぞれ 0.72, 0.74 となり最も強い相関を示している．この結果は，LLM で直接総合評価を出力する設定での結果である OA 設定での相関係数，0.70, 0.72 よりも高くなった．また，GPT-4 が決定的でないことを踏まえて，全体の 20%のデータについて評価を合計 3 回行い，その平均評価値を採用する方式を採った場合，OA 設定，MATE 設定での相関係数はそれぞれ 0.74, 0.78 となった．これらの結果は総じて自動評価であっても多面的評価を適応することの有用性を示している．

一方で，課題も残る．GPT-4 による軸ごとの評価の平均値は人手評価と比較して，Factuality, Amount Info, Formality でそれぞれ+0.24, +0.04, +0.05 の乖離がある．Llama2 では+0.43, -0.02, -0.02 の乖離があり，どちらも Factuality が過大評価になりやすいことがわかる．また，Llama2 を使用した LLM-Mate では，MF と MC 回答に対してほとんどの場合で満点を出力しており，過大評価になっていることがわかる．

5 終わりに

本研究では，LFQA タスクの多面的絶対評価の枠組みを作成した．それに従い，ELI5 データセットと ChatGPT で生成した回答に対して大規模な評価を行った．多くの場合，ChatGPT の回答が人間の回答よりも高品質であることが示された一方で，改善の余地は大きいことが明らかになった．我々が作成した多面的絶対評価枠組みは，人手評価で重要視される評価軸を明らかにし，Factuality が LFQA タスクにおいて特に需要であることがわかった．また，LLM を利用した新たな評価指標 LLM-Mate を提案し，人手評価データとの高い相関を示した．

今回作成した枠組みと LLM-Mate は，LFQA に限らずあらゆるタスクに適応可能である．また，作成した大規模な人手評価データは後続研究に大いに貢献すると考える．本枠組みと合わせて自動評価指標 LLM-Mate も幅広く利用されることを期待する．

謝辞

本研究は JSPS 科研費 JP21K21343，JP22H00524, the IBM Fellowship の助成を受けたものです．また，Stability AI より提供された計算資源をモデルの学習に使用しました．

参考文献

[1] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, 2019.

[2] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? **Transactions of the Association for Computational Linguistics**, 2020.

[3] Benjamin Heinzerling and Kentaro Inui. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, 2021.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In **Advances in Neural Information Processing systems**, 2020.

[5] Nelson F Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. **arXiv preprint arXiv:2304.09848**, 2023.

[6] Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W. Bruce Croft, and Mark Sanderson. A non-factoid question-answering taxonomy. In **Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval**, 2022.

[7] Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. ASQA: Factoid questions meet long-form answers. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2022.

[8] Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. A critical evaluation of evaluations for long-form question answering. **arXiv preprint arXiv:2305.18201**, 2023.

[9] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long form question answering. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. Association for Computational Linguistics, 2019.

[10] Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. Hurdles to progress in long-form question answering. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, 2021.

[11] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. **arXiv preprint arXiv:2107.03374**, 2021.

[12] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. **arXiv preprint arXiv:2305.11206**, 2023.

[13] Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary llms. **arXiv preprint arXiv:2305.15717**, 2023.

[14] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. **arXiv preprint arXiv:2304.03277**, 2023.

[15] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. **arXiv preprint arXiv:2112.09332**, 2021.

[16] Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing human-AI collaboration for generating free-text explanations. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. Association for Computational Linguistics, 2022.

[17] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. **arXiv preprint arXiv:2303.16634**, 2023.

[18] R OpenAI. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[19] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

A　ELI5 のサンプリング

ELI5 から 300 の質問とそれに対応する回答を抽出する過程を説明する．まず，テストセットの最初の 800 件から，URL などの外部参照がなされている回答を破棄する．次に残った回答が 2 件未満であるものを破棄する．残った質問に対して ChatGPT に回答を生成させる．ChatGPT の生成回答が “As an AI” などで始まる回答を，質問への回答を拒否したとみなし，元の質問を破棄する．残ったデータから，300 件の質問をランダムに選択する．

B　ChatGPT のプロンプト

MF，MC 回答の収集には以下を使用する．

表 2　MF 回答に使用したプロンプト

system	You are a helpful assistant who answers questions on a forum.
user	Answer the following question in 75-100 words: {question}

表 3　MC 回答に使用したプロンプト

system	You are a helpful assistant who answers questions on a forum.
user	Instruction: Answer the following question in 75-100 words. Requirements: The answer should not use difficult vocabularies. The answer should be understandable to people outside the field. The answer should be a little bit casual. Question: {question}

C　アノテーション

C.1　インターフェース

Amazon Mechanical Turk で使用した評価インターフェースの一部を図 5 に示す．

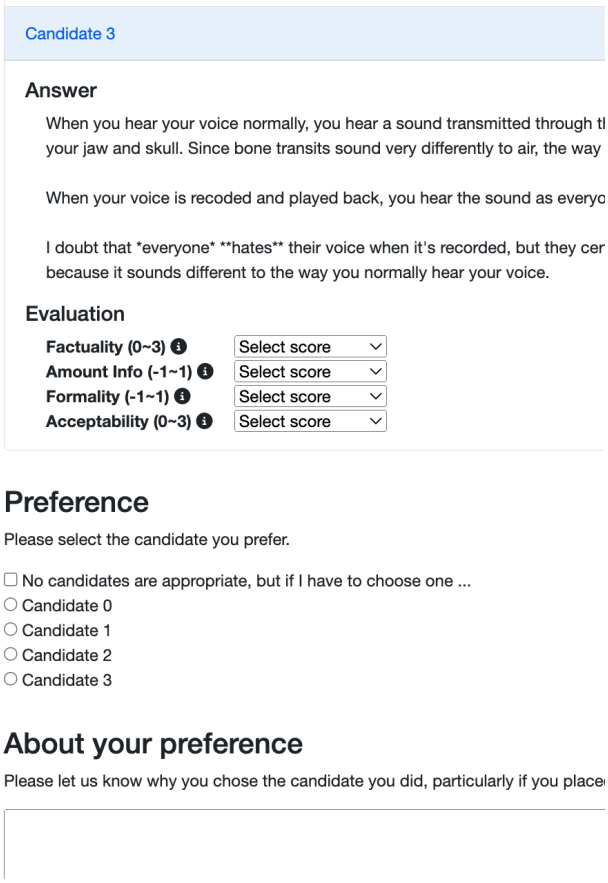

図 5　Amazon Mechanical Turk で使用した評価インターフェースの一部を示す．この他に，指示，Candidate0-2，参考文献，コメントがあるが紙面の都合で割愛する．

C.2　アノテーター一致度

図 6 は相対評価の一致度を示す．300 件中の 69 件では，3 人全員の相対評価 (4 つの回答のうちどれが一番良いか) が一致した．190 件では 2 人の評価が一致し，41 件では全員の票が割れた．

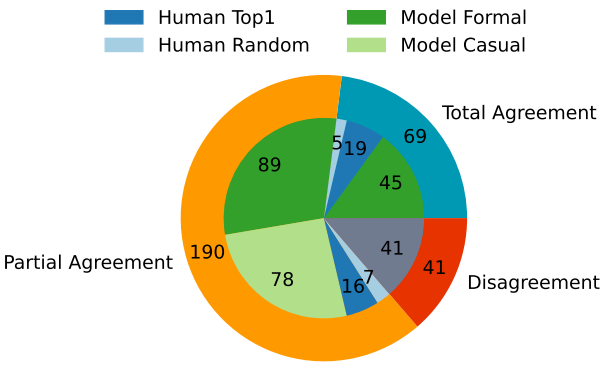

図 6　相対評価の一致度．“Total Agreement” は，3 人のアノテータが同じものを選んだことを意味し，“Disagreement” は 3 人が全員異なる回答を選択したことを意味する．内側の円グラフは，どの回答が過半数のアノテータによって選択されたかを示している．