日本語における LLM と人間の誤用傾向の差異の分析

中西純 [1]　牧野晃平 [1*]　佐々木裕 [1]
[1] 豊田工業大学
{sd21062,yutaka.sasaki}@toyota-ti.ac.jp, bear.kohei@gmail.com

概要

大規模言語モデル（LLM）は自然な文章を生成する一方で，人間の文章に一定確率で現れるはずの「誤用」が観察されることは稀である．本研究はこの現象を，日本語の母語話者における認知的取り違え（例：「危機一髪／危機一発」）に焦点を当てて定量化する．具体的には，(i) Wikipedia 編集履歴由来の正誤ペアから，認知的要因に起因する誤用事例のみを抽出して，日本語誤用データセットを構築した（81k 件）．(ii) アンケート（400 名）と LLM の出力分布から，同一指標として二択正規化確率を算出，比較し，LLM と人間の正用選好率の相関は約 0.3 と低いと確認された．(iii) SFT および DPO により，人間の誤用傾向を部分的に模倣できると示した．

1　はじめに

大規模言語モデル（Large Language Model; LLM）は，人間によって書かれた膨大な量の文章を学習することで，自然な文章を生成できる．近年の LLM は，多様な自然言語処理タスクで高い性能を示しており，生成された文章は人間の書いたものと容易に区別できない水準に達することから，LLM は人間の言語生成能力を再現しているように見える．

一方で，LLM による文章生成を観察すると，人間の文章では一定の確率で生じるはずの誤用がほとんど見られず，あまりにも文章が整いすぎているという違和感がある．誤用とは，例えば「危機一髪」と「危機一発」のように，本来の表現を取り違えて誤った意味・用法で用いることである．人間は誤用を含む記述することが多々あるが，図 1 のように，LLM が誤用を生成することは極めて稀である．これは，LLM が誤りの少ない文章に強く偏った，人間とは異なる言語分布を獲得している可能性を示唆する．

人間の言語生成における誤用は，単なるノイズではなく，認知や知識の不完全さに起因する自然な現象で，誤用の出現頻度や分布は人間らしい言語生成過程の特徴を反映している．人間と LLM の誤用生成傾向の差異は，LLM が生成する文章の人間らしさや表現の多様性に影響するため，LLM の適切な利用を考える上で重要な問題である．

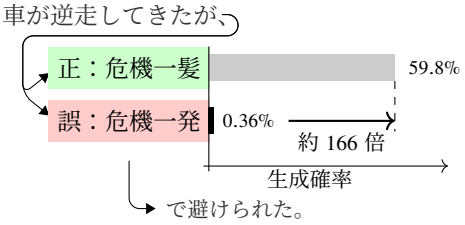

図 1　LLM（llm-jp/llm-jp-3-1.8b）に対し，文脈に続く正用・誤用候補の確率を teacher forcing で算出し，二択正規化した例．誤用側の相対確率が極端に低い．

本研究の目的は，日本語における人間と LLM の誤用傾向の差異を，同一の評価枠組みで定量化し，さらにその差異を学習により制御可能かを検証することである．具体的には，(i) 認知的誤用に焦点化した大規模データセットを構築し，(ii) 人間の誤用と正用の二択選好確率と LLM の二択正規化確率を直接比較し，(iii) Direct Preference Optimization (DPO) / Supervised Fine-Tuning (SFT) により誤用傾向を部分的に人間へ近づけることを試みる．

本研究の主な貢献は，以下のとおりである．

- データ：認知的誤用に焦点化した日本語誤用データセット（81,517 件）を構築する．
- 分析：人間の選好確率と LLM の内部確率を同一指標で比較し，誤用抑制と相関の低さを定量的に示す．
- 制御：DPO/SFT により誤用傾向（分布）を部分的に調整できることを示す．

2　関連研究

LLM の誤りを分析する先行研究の多くは，特定のタスクにおける誤答パターンに着目している．例えば Liu ら [1] は，多肢選択問題における学生と LLM

* 現在，株式会社エムニ所属

の誤答傾向が対象で，事実の真偽や推論の正しさといったコンテンツの誤りを対象としており，誤りの性質はドメインやタスクに強く依存する．これに対し，本研究が扱う誤用は，語や表現の選択という言語使用一般の現象である．

誤用事例を収集した資料としては，慣用的な誤りを集めた書籍が存在する [2, 3, 4]．しかし，これらは文脈情報が乏しく，デジタル化されていないため，大規模な分析や LLM との比較に適さない．他の誤用を含む既存コーパスは，日本語学習者の誤りを収集したコーパス [5, 6] であり，母語の影響を受ける [7, 8]．そのため，母語話者による認知的な誤用傾向を分析する本研究の目的には適合しない．

誤用データは，言語使用の分析にとどまらず，誤り検出・訂正モデルの学習や評価，文章作成支援などの応用的な研究でも重要な役割を果たす [9, 10]．本研究は，母語話者の認知的誤用に焦点化した大規模データを構築し，人間の二択選好と LLM の内部確率を同一指標で直接比較する点に特徴がある．

3 日本語誤用データセットの構築

本研究では，人間と LLM の誤用生成傾向を比較するために，多数の誤用事例が必要である．本研究で対象とする誤用事例には，成人の日本語母語話者による書き言葉の誤りで，キーボード操作などの物理的ミスではなく，認知や勘違いに起因する誤りであること，文字・単語レベルで誤り箇所が同定可能であること，そして十分な文脈長が求められる．

ベースとして日本語 Wikipedia 入力誤りデータセット（Japanese Wikipedia Typo Dataset; JWTD）[11] を用いる．JWTD は編集履歴から文単位の修正前後を収集した正誤ペアであり，本研究の枠組みに適合する．一方で，JWTD には誤変換・表記揺れ・打鍵ミスなど，本研究の対象外の事例も多く含まれる．対象件数は約 29 万件に及ぶため，全件の手作業分類は困難である．また，誤用判定は文脈依存性が高く，単純な辞書・ルールでは高精度な分離が難しい．

そこで，今回は LLM を用いた ICL（In-Context Learning）により，各事例を二値分類問題として解かせた．具体的には，誤りの原因が認知的な誤用 (`CognitiveError`) であるか単なる入力操作ミス (`KeystrokeError`) かを分類する．本研究では，認知的な誤用のみを対象として分析をしたいため，適合率 (Precision) が高い分類が重要視される．これを実施するために，開発用データとして 200 件のデータにラベル付けをして，複数のモデルとプロンプト構成の組み合わせから最も適合率が高いモデルを選択する．パラメタ数やモデル構成が異なる 4 種類のモデルと，異なる問題の定式化の組み合わせに対して，開発データに対する分類性能を評価した．問題の定式化は，`CognitiveError` を正例とした二値分類（0/1），ラベル名を生成させた場合（label），その語の定義とラベル名を生成した場合（def.+label）を比較して，その結果を表 1 に示した．最終的に，適合率 83.9%を示した，def.+label の設定の tokyotech-llm/Llama-3.1-Swallow-70B-Instruct-v0.3 を採用した．

表 1 モデル・応答形式別の分類精度（10-shot）

モデル	応答形式	Acc.	Prec.	Rec.	F1
llm-jp-13B	0/1	0.560	0.753	0.440	0.556
ELYZA-8B	0/1	0.570	0.636	0.728	0.679
Swallow-8B	0/1	0.440	0.659	0.216	0.325
Swallow-70B	0/1	0.630	0.642	0.920	0.757
llm-jp-13B	label	0.660	0.669	0.904	0.769
ELYZA-8B	label	0.610	0.626	0.936	0.750
Swallow-8B	label	0.610	0.669	0.744	0.705
Swallow-70B	label	0.640	0.635	**1.000**	0.776
ELYZA-8B	def.+label	**0.789**	0.794	0.896	**0.842**
Swallow-70B	def.+label	0.779	**0.839**	0.798	0.818

全事例の分類を実行した結果，`CognitiveError` として，161,103 件のデータを用意することができた．この分類結果の妥当性を検証するため，`CognitiveError` に分類された事例からランダムに 500 件を抽出し，手作業で正誤判定を行ったところ，426 件が `CognitiveError` で，適合率が 85.2%であることから，一定水準の信頼性があると確認できた．

さらに，各事例を精査したところ，そのまま使用するには問題のある事例が一部に含まれていることが確認された．具体的には，文脈が極端に短い事例や，固有名詞に起因する誤り，誤用箇所が文中で既出である場合などである．これらを除外する後処理として，ルールベースおよび MeCab+UniDic による形態素解析結果を用いたフィルタリングを行い，最終的に 81,517 件の誤用事例が得られた．

4 人間と LLM の誤用傾向比較

本研究では，人間と LLM における誤用生成傾向を，同一の評価指標に基づく評価を目指す．そこで，ある文脈 x が与えられたときに，正用表現が選択される相対的な確率 Q（以降「正用選好率」と呼ぶ）を以下の指標で評価する．ここで，$y_{\text{正}}$ は正用

表 2　設問ごとの人間と LLM の Q の比較．括弧内は Q の評価対象外とした．

誤	正	人間	sarashina 3b	ELYZA 8B	Swallow 70B Inst
一獲千金	一攫千金	0.340	0.945	0.986	0.974
(場内を) 湧かせた	沸かせた	0.435	0.679	0.984	0.920
若干	弱冠	0.481	0.818	0.755	0.245
木偶の棒	木偶の坊	0.583	0.947	0.971	0.818
激 (を飛ばし)	檄	0.603	0.018	0.005	0.026
(雪辱を) 晴らした	果たした	0.693	0.976	0.997	0.932
貪欲に	貪欲に	0.753	0.997	1.000	1.000
思いばかって	慮って	0.813	1.000	1.000	1.000

表現，$y_{\text{誤}}$ は対応する誤用表現，x は文脈とする．

$$Q = \frac{P(y_{\text{正}} \mid x)}{P(y_{\text{正}} \mid x) + P(y_{\text{誤}} \mid x)} \tag{1}$$

4.1　人間の誤用傾向の測定

人間における誤用傾向を測定するため，Yahoo!クラウドソーシングを用いて，2025/10/31・11/12 の 2 日間で 400 名を対象とするアンケート調査を実施した．各参加者には，誤用部分の周辺文脈と正用と誤用を提示し，二択のどちらが正しいと思うかを出題した．出題数は 1 人あたり 90 問とし，想定タスク完了時間を 15 分に設定した．各設問は平均 10 名，最小 3 名から回答を得た．不誠実回答（バッドワーカ）対策として，極端に簡単な問題を 10 問混在させ，正解率が 7 割未満の回答者，および，タスク全体の完了時間が 1 分未満の回答者（計 31 件）を除外した．得られた各設問における回答分布から，人間の正用選好率を以下のように推定した．

$$Q_{\text{human}} = \frac{P_{\text{正}}}{P_{\text{正}} + P_{\text{誤}}} = \frac{\frac{n_{\text{正}}}{n_{\text{正}}+n_{\text{誤}}}}{\frac{n_{\text{正}}}{n_{\text{正}}+n_{\text{誤}}} + \frac{n_{\text{誤}}}{n_{\text{正}}+n_{\text{誤}}}} = \frac{n_{\text{正}}}{n_{\text{正}} + n_{\text{誤}}} \tag{2}$$

4.2　LLM の誤用傾向の測定

LLM には文脈 x を入力し，続く候補表現 y の生成確率 $P(y \mid x)$ を，モデルの出力ロジットから求める．実際にデコードして文章生成を行うのではなく，**teacher forcing** で候補文字列のトークン列確率を計算する．本研究では，文脈 x の末尾に候補文字列 y をそのまま連結し，候補側トークンの対数確率を集計した．正用・誤用ペアは同一表現の取り違えが中心で，多くは同程度の長さであるため，本設定で相対的な選好の比較が可能である．

表 3　人間と LLM の正用選好率と相関

	平均 Q	ピアソン相関係数	スピアマン順位相関係数
Human	0.706	1.000	1.000
llm-jp-3.1-13b	0.776	0.110	0.202
+DPO	0.777↗	0.103↘	0.201↘
+SFT(誤)	0.791↗	0.197↗	0.282↗
+SFT(正)	0.808↗	0.150↗	0.280↗
llm-jp-3.1-13b-inst	0.746	-0.125	0.021
+DPO	0.750↗	-0.124↗	0.022↗
+SFT(誤)	0.781↗	0.047↗	0.147↗
+SFT(正)	0.791↗	-0.001↗	0.076↗
sarashina2.2-3b	0.781	0.487	0.425
+DPO	0.783↗	0.496↗	0.437↗
+SFT(誤)	0.730↘	0.498↗	0.364↘
+SFT(正)	0.788↗	0.462↘	0.309↘
sarashina2.2-3b-inst	0.777	0.523	0.479
+DPO	0.779↗	0.528↗	0.471↘
+SFT(誤)	0.706↘	0.409↘	0.327↘
+SFT(正)	0.780↗	0.413↘	0.266↘
Tanuki-8B-dpo-v1.0	0.779	0.256	0.322
+DPO	0.776↘	0.242↘	0.321↘
+SFT(誤)	0.736↘	0.166↘	0.237↘
+SFT(正)	0.762↘	0.108↘	0.190↘
Swallow3.1-8B	0.787	0.321	0.362
+DPO	0.786↘	0.331↗	0.360↘
+SFT(誤)	0.772↘	0.350↗	0.339↘
+SFT(正)	0.822↗	0.311↘	0.328↘
Swallow3.1-8B-Inst	0.805	0.251	0.317
+DPO	0.802↘	0.254↘	0.318↗
+SFT(誤)	0.783↘	0.316↗	0.358↗
+SFT(正)	0.820↗	0.200↘	0.297↘
Swallow3.3-70B	0.762	0.374	0.329
Swallow3.3-70B-Inst	0.749	0.263	0.261
ELYZA-JP-8B	0.800	0.260	0.270
+DPO	0.797↘	0.279↗	0.290↗
+SFT(誤)	0.756↘	0.317↗	0.341↗
+SFT(正)	0.803↗	0.256↘	0.285↗
Llama-3.1-8B	0.782	0.467	0.518
Llama-3.1-8B-Inst	0.825	0.483	0.410
Qwen2.5-14B	0.774	0.357	0.337
Qwen2.5-14B-Inst	0.757	0.409	0.342
gpt-oss-20b	0.678	0.405	0.328
gpt-oss-120b	0.621	0.181	0.201

y がトークン列 $(y_1, \ldots, y_T)$ からなるとき，

$$P(y \mid x) = \prod_{t=1}^{T} P(y_t \mid x, y_{<t}) \tag{3}$$

により算出する．日本語では空白を挿入せず，文脈 x の末尾に候補文字列 y をそのまま連結した系列に対し，候補側トークンの対数確率を集計する．最後に式 (1) により二択正規化し，Q_{LLM} を得る．

4.3 人間と LLM の比較

比較では次の二点を主要観点とする．

- **分布のずれ：** 平均 Q の差（LLM がどの程度誤用を抑制するか）
- **難しさの整合：** 設問ごとの Q の相関（人間が誤りやすい事例で LLM も誤るか）

人間および LLM について算出した正用選好率を比較した結果を，全体の平均正用選好率および設問ごとの相関を相関係数（ピアソン相関係数およびスピアマン順位相関係数）として表 3 に示す．全体として多くの LLM は人間より平均正用選好率が高く，誤用を抑制する傾向が確認された．また，人間と LLM の設問ごとの正用選好率の相関は概ね 0.3 前後に留まり，人間と LLM の誤用傾向の整合が低いと示された．表 2 に示した設問事の正用選考率は，人間の正用選好率が低い設問に対して，LLM が高い正用選好率を示す事例が複数存在することを示す．これは，LLM が人間の誤用分布を反映しておらず，異なる判断基準（あるいは学習分布の偏り）に基づいて誤用を回避している可能性を示唆する．

5 LLM の誤用傾向の調整

前節の結果は，多くの LLM が誤用を強く抑制し，人間の難しさと整合しないことを示した．そこで，人間の誤用選好を明示的に学習させることで，LLM の誤用傾向（分布）を人間に近づけられるかという疑問が生まれる．そのために，構築した誤用データを教師として，DPO および SFT を適用し，LLM と人間の誤用傾向のアラインメントを試みる．

ファインチューニング後の各モデルに対して，4.2 節と同様の手法を用いて，正用選好率および人間との相関係数を算出した．結果は，表 3 における+DPO および+SFT の行に示す．各数値の横に付した矢印は，ファインチューニング前のモデルに対する増減を表す．評価の結果，DPO または SFT によって，人間の正用選好率との相関係数がわずかに上昇する傾向が確認された．このことから，誤用に関する人間の選択傾向を明示的に学習させることで，LLM の誤用生成傾向を一定程度，人間に近づけられていることが確認できた．

本研究の結果は，多くの LLM が誤用を強く抑制し，人間の誤用分布とは異なる選好を持つことを示す．この乖離は，学習データにおける規範的表現の過多や生成時の暗黙的な「正しさ」バイアス，など複数要因が重畳した結果である可能性がある．また，DPO/SFT により誤用傾向を動かせることは，誤用が単なる偶然ではなく，モデル内で**学習可能な選好**として表現されていることを示唆する．一方で，誤用傾向を人間へ寄せることが，文章の自然性・有用性・安全性にどのような副作用を持つかは未解明であり，今後の検証が必要である．

6 おわりに

本研究は，日本語における認知的誤用に着目し，人間と LLM の誤用傾向の差異を同一指標で定量的に分析した．まず，日本語母語話者による認知的要因に起因する誤用に焦点を当てた，日本語誤用データセットを構築した．次に，人間を対象とするアンケート調査と，LLM の内部確率に基づく評価を通じて，誤用生成確率を同一指標で比較した．その結果，LLM は誤用を抑制する傾向があり，人間の選択確率分布と整合しない（相関が低い）ことを示した．事例を個別に分析したところ，人間にとって難易度の高い事例であっても，LLM が極端に高い正用選好率を示す例が複数確認された．さらに，誤用データを用いた DPO および SFT によるファインチューニングを行い，LLM の誤用生成傾向を人間に近づける試みを行った．その結果，相関係数がわずかに上昇し，誤用に関する人間の選好を部分的に模倣できる可能性が示唆された．本研究は，LLM の性能評価において正用選好率のみでは捉えられない，人間らしさや分布的な違いを明らかにした点に意義がある．今後は，誤用傾向の調整が生成文の自然性・下流タスク・安全性に与える影響を広範に検証し，「正確さ」だけでは捉えられない分布的評価の設計へ接続することが課題である．

謝辞

LLM の追加学習にあたり，産総研及び AIST Solutions が提供する ABCI 3.0 を「ABCI 3.0 開発加速利用」の支援を受けて利用した．ここに深く感謝する．

参考文献

[1] Naiming Liu, Shashank Sonkar, and Richard Baraniuk. Do llms make mistakes like students? exploring natural alignments between language models and human error patterns. In **Artificial Intelligence in Education**, pp. 364–377. Springer Nature Switzerland, 2025.

[2] 国広哲弥. 新編　日本語誤用・慣用小辞典. 講談社, 2010.

[3] NHK アナウンス室. NHK 間違いやすい日本語ハンドブック. NHK 出版, 2013.

[4] 三條雅人. ネットで見かけた信じられない日本語—うろ覚え・勘違い・言い間違い・誤植. 社会評論社, 2015.

[5] 市川保子, 浅山友貴, 荒巻朋子, 板井美佐, 太田陽子, 坂本まり子, 杉本ろここ, 副島昭夫, 田代ひとみ, 野田景子, 本郷智子. 日本語誤用辞典. スリーエーネットワーク, 2010.

[6] 迫田久美子, 小西円, 佐々木藍子, 須賀和香子, 細井陽子. 多言語母語の日本語学習者横断コーパス. 国語研プロジェクトレビュー, Vol. 6, No. 3, pp. 93–110, 2016.

[7] 石橋玲子. 第 2 言語習得における第 1 言語の関与. 風間書房, 2002.

[8] 迫田久美子. 改訂版 日本語教育に生かす 第二言語習得研究. アルク, 2020.

[9] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[10] Agnes Luhtaru, Taido Purason, Martin Vainikko, Maksym Del, and Mark Fishel. To err is human, but llamas can learn it too. In **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 12466–12481, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[11] 田中佑, 村脇有吾, 河原大輔, 黒橋禎夫. 日本語 wikipedia の編集履歴に基づく入力誤りデータセットと訂正システムの構築. 自然言語処理, Vol. 28, No. 4, pp. 995–1033, 2021.

[12] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In **International Conference on Learning Representations**, 2022.

[13] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In **Advances in Neural Information Processing Systems**, Vol. 36, pp. 10088–10115. Curran Associates, Inc., 2023.

A　分類に使用したプロンプト

3 章にて最終的に使用したプロンプトを図 2 に示す．指示文に続けて 10 件の例を提示し，続けて分類対象の文を与える．

あなたは誠実で優秀な日本人のアシスタントです。あなたにはこれから示す文に含まれる誤りの種類を分類してもらいます。
文には、誤りと修正が <修正前/修正後> の形式で含まれます。
分類の基準は次のとおりです。

CognitiveError: 思考の勘違いに起因する誤り
- 人間が文を考える過程での勘違いや迷いに起因する
- 類似した漢字の混同
- <線維/繊維>性植物
- 同一あるいは読みの言葉の混同
- <危機一発/危機一髪>
- 慣用的な表現の誤用
- 雪辱を<晴らす/果たす>

KeystrokeError: 指先などの運動に起因する誤り
- 単純なキーボードの押し間違い
- プログラムの<動作/動産>環境
- 手書きではありえないような変換ミス
- 雑誌の<地名度/知名度>
- 読みは一致しているが、取り違えると言語的に不自然になる場合
- 記事の<無いよう/内容>については...

修正前後での意味の類似性を検討し、分類結果は CognitiveError KeystrokeError のどちらかで出力してください。

(ここから few-shot)
その頃、東京市長でもあった衆議院議員尾崎行雄は、先の日露戦争の<講話/講和>に助力してもらったアメリカへの謝礼を考えていたところへ、水野からこのような計画があることを知らされた。

「講話」（講義形式で）わかりやすく説いて聞かせること。また、その話。
「講和」交戦国間の合意で、戦争を終結し、平和を回復すること。
意味の類似: False
Result: KeystrokeError

(同様に残り 9 例)

(対象の文)

図 2　ICL 用プロンプト

B　使用した LLM の正式名称

本文中では LLM のモデル名を短縮した形で表記していたが，表 4，表 5，表 6 に正式な名称を示す．https://huggingface.co/ の末尾にモデル名を加えることで，当該モデルにアクセスできる．

表 4　表 1 で使用した LLM

短縮表記	正式名称
llm-jp-13B	llm-jp/llm-jp-3-13b-instruct3
ELYZA-8B	elyza/Llama-3-ELYZA-JP-8B
Swallow-8B	tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.3
Swallow-70B	tokyotech-llm/Llama-3.1-Swallow-70B-Instruct-v0.3

表 5　表 2 で使用した LLM

短縮表記	正式名称
sarashina-3b	sbintuitions/sarashina2.2-3b
ELYZA-8B	elyza/Llama-3-ELYZA-JP-8B
Swallow-70B-Inst	tokyotech-llm/Llama-3.3-Swallow-70B-Instruct-v0.4

表 6　表 3 で使用した LLM

短縮表記	正式名称
llm-jp-3.1-13b	llm-jp/llm-jp-3.1-13b
llm-jp-3.1-13b-inst	llm-jp/llm-jp-3.1-13b-instruct4
sarashina2.2-3b	sbintuitions/sarashina2.2-3b
sarashina2.2-3b-inst	sbintuitions/sarashina2.2-3b-instruct-v0.1
Tanuki-8B-dpo-v1.0	weblab-GENIAC/Tanuki-8B-dpo-v1.0
Swallow3.1-8B	tokyotech-llm/Llama-3.1-Swallow-8B-v0.5
Swallow3.1-8B-Inst	tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5
Swallow3.3-70B	tokyotech-llm/Llama-3.3-Swallow-70B-v0.4
Swallow3.3-70B-Inst	tokyotech-llm/Llama-3.3-Swallow-70B-Instruct-v0.4
ELYZA-JP-8B	elyza/Llama-3-ELYZA-JP-8B
Llama-3.1-8B	meta-llama/Llama-3.1-8B
Llama-3.1-8B-Inst	meta-llama/Llama-3.1-8B-Instruct
Qwen2.5-14B	Qwen/Qwen2.5-14B
Qwen2.5-14B-Inst	Qwen/Qwen2.5-14B-Instruct
gpt-oss-20b	openai/gpt-oss-20b
gpt-oss-120b	openai/gpt-oss-120b

C　人間に対するアンケートの出題

Yahoo!クラウドソーシングにて出稿したタスク画面を図 3 に示す．回答者の平均年齢は 49.6 歳であった．

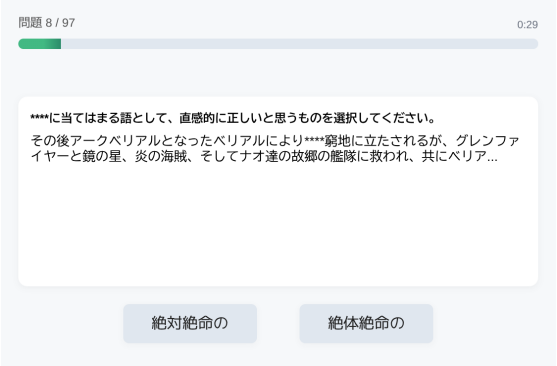

図 3　タスク画面．穴埋め 2 択形式で出題する．

D　FT 時の学習設定

DPO では，人間の調査において選択確率が高かった表現を `chosen`，選択されにくかった表現を `rejected` とし，人間の選好を反映するように学習を行った．一方，SFT では教師データを誤用表現/正用表現として，誤用の生成を促進/抑制する二つの設定を用意した．QLoRA[12, 13] を使用し，量子化は `nf4`，LoRA のランク $r = 64$ で `q_proj,k_proj,v_proj,o_proj` に追加したパラメータを学習させた．

E　ソースコード

- データセット構築，誤用傾向分析，FT
  https://github.com/JunSotohigashi/misusing-corpus-jp
- 誤用アンケート調査
  https://github.com/JunSotohigashi/misusing-survey