# 大規模言語モデルに含まれる社会集団間の感情の抽出

田中 邦朋 [1]　笹野 遼平 [2]　武田 浩一 [2]
[1] 名古屋大学情報学部　[2] 名古屋大学情報学研究科
tanaka.kunitomo.z3@s.mail.nagoya-u.ac.jp　{sasano,takedasu}@i.nagoya-u.ac.jp

## 概要

大規模言語モデル (LLM) は大量のテキストからモデルを学習することで、社会常識や偏見など、人間が潜在的に持つ知識や感情をある程度、獲得しているとされる。しかし、特定の社会集団が持つ感情を各種の LLM からどのくらい抽出可能かは明らかとなっていない。本研究では、国籍、宗教、人種/民族という観点でそれぞれ規定される社会集団を対象に、ある集団から別の集団への印象に関する質問を LLM に入力し、その応答に感情分析を行うことで、集団間の感情を LLM を用いてどの程度、抽出できるかの検証に取り組む。

## 1　はじめに

大規模言語モデル (**L**arge **L**anguage **M**odel; LLM) は、様々なタスクにおいて、あたかも人間が作成しているかのような質の高いテキスト生成を実現している [1, 2]。このような LLM が持つ人間の行動や対話を模倣する能力に着目し、人間を被験者とした社会実験を LLM により再現しようとする試みも多く行われている [3, 4, 5, 6, 7]。一方で、大規模コーパスから学習された言語モデルが持つ社会的に望ましくない偏った価値観 [8, 9] を修正するため、近年の LLM は、人間のフィードバックによる強化学習 (**R**einforcement **L**earning with **H**uman **F**eedback; RLHF) [10] を取り入れることで、社会的に望ましくない出力を抑制するように学習が行われている。このように RLHF をモデルの学習に取り入れた結果、社会実験の代替モデルとして LLM が本来習得した知見を出力しなくなる可能性が考えられる。

そこで本研究では、社会集団間の感情を対象に、RLHF を取り入れた各種の LLM から、これらの感情をどのくらい抽出可能かの検証に取り組む。具体的には、国籍、宗教、人種/民族の各属性で規定される、ある集団から別のある集団への感情を、各種の LLM の出力から再現できるかを検証する。概要を図 1 に示す。まず、(1) ある社会集団が別の社会集団に持つ感情を問う質問を LLM に入力する。次に、(2) LLM の出力に対し感情分析を適用してスコア化し、(3) 集団の全組み合わせのスコアを集計する。最後に、(4) 集計結果と実際の社会調査データとの相関係数を算出することで、LLM から集団間の感情がどの程度、抽出可能か評価する。

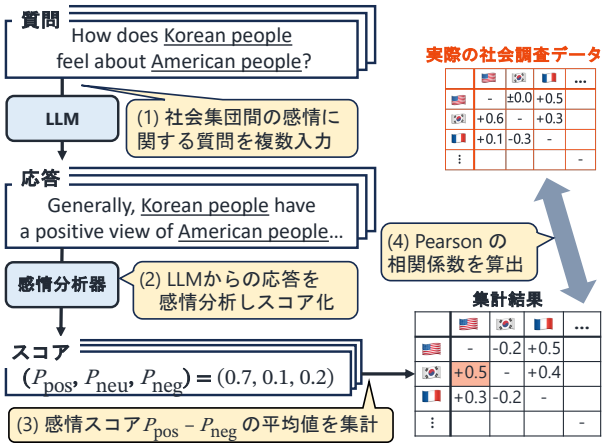


**図 1**　社会集団間の感情の抽出実験

## 2　対象とする社会集団とデータ

本研究では社会集団として、国籍、宗教、人種/民族の 3 つの属性で規定される集団を考える。各属性について、感情の主体となる集団 $G_{\text{from}}$ から、感情の対象となる集団 $G_{\text{to}}$ への感情を LLM から抽出し、実際の社会調査の結果から得られたデータ (実データ) との一致度を算出することで、LLM からどの程度、社会集団間の感情を抽出できているか評価する。各属性において考慮する社会集団の一覧を表 1 に示す。また、各属性に対応する実データの詳細は以下の通りである。

**国籍**　2022 年に行われた新聞通信調査会による社会調査 [11] のうち、各国籍の参加者に対して他国への感情を 4 択で質問し、ポジティブな 2 つの選択肢のいずれかを回答した割合を集計したデータ

**表 1**　属性ごとの社会集団の一覧

| 属性 | 社会集団 |
|---|---|
| 国籍 | Chinese (CN), French (FR), British (GB), Korean (KR), Thai (TH), American (US), Japanese* (JP), Russian* (RU) |
| 宗教 | atheist (ATH), Catholic (CTH), Evangelical (EVG), Jew (JEW), Mainline Protestant, (MPR), Mormon (LDS), Muslim* (MUS) |
| 人種/民族 | Asian (AS), Black (BL), Hispanic (SP), White (WH) |

*は $G_{\text{to}}$ としてのみ考慮した集団であることを示す

**宗教**　2022 年にアメリカで行われた Pew Research Center による社会調査 [12] のうち、各宗教の信者に他宗教への好感度を 6 択で質問し、ポジティブな 2 つの選択肢を回答した割合から、ネガティブな 2 つの選択肢を回答した割合を減じた値を集計したデータ

**人種/民族**　2019 年にアメリカで行われた Pew Research Center による社会調査 [13] のうち、各人種/民族に属する参加者が他人種/民族への感情を 100 点満点で回答した結果の平均値を集計したデータ

# 3　集団間の感情の抽出手順

## 3.1　社会集団間の感情に関する質問

本研究では、図 1 に示す手順で、LLM から社会集団間の感情をどのくらい抽出可能か評価する。まず、LLM に入力する社会集団間の感情に関する質問を表 2 に示すようなテンプレートを用いて作成する。応答が英語で出力されるように最初に “Always answer in English.” と指示を与え、テンプレートを用いて作成した質問を加えたプロンプトを LLM に入力する。質問テンプレートは大きく Yes/No 型質問と Wh 型質問の 2 タイプに分けることができ、各タイプ 6 種類、合わせて 12 種類で構成される。

各テンプレートの $G_{\text{from}}$ および $G_{\text{to}}$ の部分には、それぞれ感情の主体となる集団、感情の対象となる集団が入る。この際、集団を意味するテキスト表現として、国籍、人種/民族については “people” を後置し、宗教については複数形にして入力する。また、社会調査がアメリカで実施された宗教と人種/民族については、質問の前に “In the U.S., ” というフレーズを追加することで、アメリカの社会集団についての感情を取り出すことを明示する。例えば、アジア人から黒人への感情に関して、“Do $G_{\text{from}}$ like $G_{\text{to}}$?” というテンプレートを用いて生成された質問は、“In the U.S., do Asian people like Black people?” となる。

**表 2**　LLM へ入力する質問のテンプレート

| 質問タイプ | テンプレート |
|---|---|
| Yes/No 型 | Do $G_{\text{from}}$ like $G_{\text{to}}$?<br>Do $G_{\text{from}}$ love $G_{\text{to}}$?<br>Do $G_{\text{from}}$ appreciate $G_{\text{to}}$?<br>Do $G_{\text{from}}$ have good feelings on $G_{\text{to}}$?<br>Do $G_{\text{from}}$ have positive feelings on $G_{\text{to}}$?<br>Do $G_{\text{from}}$ have favorable feelings on $G_{\text{to}}$? |
| Wh 型 | How do $G_{\text{from}}$ feel about $G_{\text{to}}$?<br>How do $G_{\text{from}}$ perceive $G_{\text{to}}$?<br>How do $G_{\text{from}}$ view $G_{\text{to}}$?<br>What are $G_{\text{from}}$’s feelings on $G_{\text{to}}$?<br>What are $G_{\text{from}}$’s thoughts on $G_{\text{to}}$?<br>What are $G_{\text{from}}$’s impressions on $G_{\text{to}}$? |

LLM の応答の揺らぎによる影響を抑制するため、質問テンプレートで生成された各質問は独立に 3 回入力し、3 つの応答を得る。質問は 12 種類存在するので、各社会集団ペアに対し、36 種類の応答が得られることになる。

## 3.2　LLM の応答のスコア化と集計

LLM の応答に感情分析を適用することで、応答の感情をスコア化する。具体的には、与えられた入力に対し positive、neutral、negative である確率をそれぞれ出力する 3 値感情分析器に、LLM の各応答を入力し、positive である確率 $P_{\text{pos}}$ から、negative である確率 $P_{\text{neg}}$ を引いた値を算出する。続いて、集団間感情を算出するため、集団ペアごとに算出されたスコアを平均する。この際、質問タイプによる感情の抽出性能の違いを分析できるように、Yes/No 型質問 6 種類、計 18 応答のスコアのみを平均する設定、Wh 型質問 6 種類、計 18 応答のスコアのみを平均する設定、全質問 12 種類、計 36 応答のスコアを平均する設定の 3 つの設定を比較する。

各属性について、集団の全組み合わせの感情のスコアを算出、集計し、社会調査の結果から得られた実データと比較する。LLM を用いて得られたスコアと、実データのスコアの分布は異なると考えられるため、スコアの絶対的な差ではなく、相関係数により一致度を評価する。具体的には、本研究では、LLM を用いて得られたスコアと調査データの Pearson の相関係数を算出することで、LLM から集団間の感情がどの程度、抽出可能か評価する。

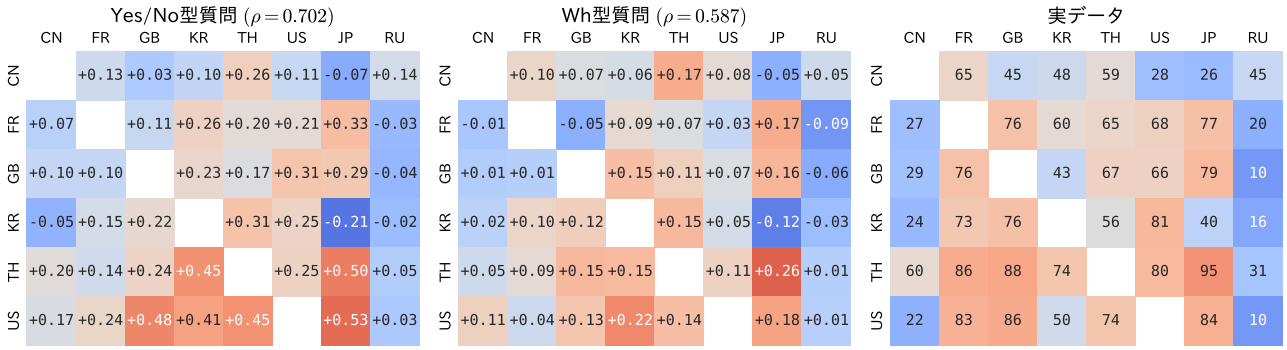


**図 2**　GPT-4 から得られた国籍間の感情スコアの集計結果と国籍の実データ．縦軸が $G_{\text{from}}$，横軸が $G_{\text{to}}$．

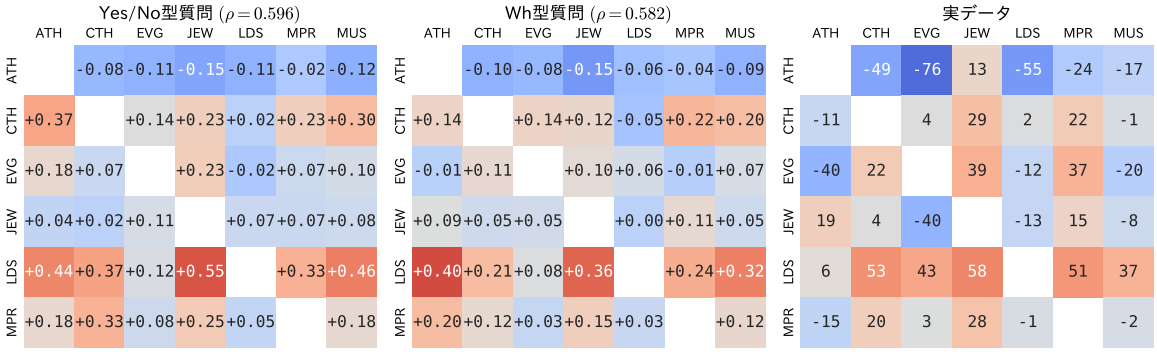


**図 3**　GPT-4 から得られた宗教間の感情スコアの集計結果と宗教の実データ．縦軸が $G_{\text{from}}$，横軸が $G_{\text{to}}$．

# 4　実験

## 4.1　実験設定

評価対象の LLM として次の 5 つの LLM をデフォルト設定で用いた。

- GPT-3.5 Turbo (`gpt-3.5-turbo-0613`[1])
- GPT-4 (`gpt-4-preview-1106`[2])
- Llama 2-Chat 13B[3]
- Llama 2-Chat 70B[4]
- Vicuna 13B v1.5[5]

感情分析器には TweetNLP [14] の Sentiment Analysis を用いた。最大入力長が 512 トークンであることから、512 トークンずつ、直前の 256 トークンを重複させた上で入力し、各分割での感情の生起確率を平均し、最終的な感情スコアを算出した。また、LLM の応答を感情分析器を入力する際、出力されるスコアが感情分析モデルの持つ集団へのバイアスに影響されないよう、社会集団を示す語を `[MASK]` トークンに置き換えた上で感情分析器に入力した。

1）https://platform.openai.com/docs/models/gpt-3-5
2）https://platform.openai.com/docs/models/gpt-4
3）https://huggingface.co/meta-llama/Llama-2-13b-chat
4）https://huggingface.co/meta-llama/Llama-2-70b-chat
5）https://huggingface.co/lmsys/vicuna-13b-v1.5

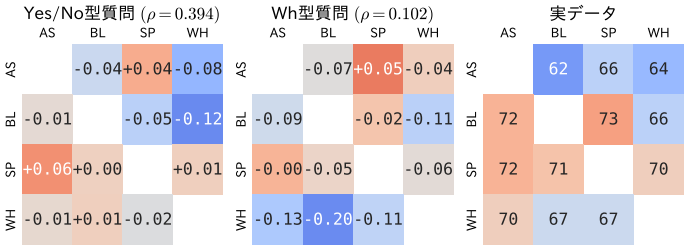


**図 4**　GPT-4 から得られた人種/民族間の感情スコアの集計結果と人種/民族の実データ．縦軸が $G_{\text{from}}$，横軸が $G_{\text{to}}$．

## 4.2　実験結果

LLM として GPT-4 を用いた場合の、国籍間、宗教間、人種/民族間の感情スコアの集計結果と対応する実データを図 2、図 3、図 4 にそれぞれ示す。各図の左側が Yes/No 型質問に対する計 18 応答のスコアをまとめた結果、中央が Wh 型質問に対する計 18 応答のスコアをまとめた結果となっており、$\rho$ の値は右側に示す実データとの相関係数を示している。

国籍間、および、宗教間の感情については、Yes/No 型質問、Wh 型質問、いずれに対しても相関係数は 0.58 から 0.7 程度の値となっており、ある程度、高い相関で LLM から集団間の感情を抽出できていると言える。一方、人種/民族間の感情については、相関係数は Yes/No 型質問に対して 0.394、Wh 型質問に対して 0.102 と、大幅に低い値となった。

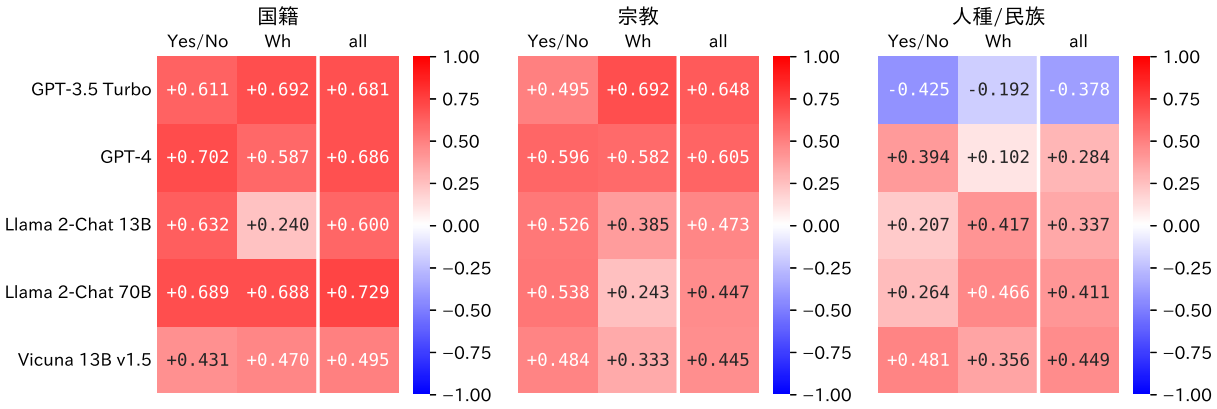


**図 5**　言語モデルと質問タイプの組み合わせごとの LLM から抽出した感情スコアと実データの相関係数.

5 つの LLM それぞれについて、Yes/No 型質問、Wh 型質問、および、その両方 (all) を用いて感情を抽出した場合の、実データとの相関係数をまとめた結果を図 5 に示す。国籍間、および、宗教間の感情については比較的大きな相関係数が得られることが確認できる。LLM による差はほとんど確認できなかった一方で、質問タイプについては Yes/No 型質問の方が安定した結果が得られた。Yes/No 型質問、および、all に対しては相関係数は最低でも 0.431 と、中程度の相関が確認できたことから、Yes/No 型質問を用いることで、国籍間、宗教間の感情は LLM からある程度、抽出可能であると言える。

一方、人種/民族に関しては、国籍、宗教ほど高い相関は確認できなかった。特に GPT-3.5 Turbo については負の相関が見られた。人種/民族について、相対的に低い相関しか得られなかったのは、RLHF により社会的バイアスが軽減された結果、社会的に好ましくない可能性がある入力に対して明示的な回答を拒否するようになったためである可能性が考えられる。そこで、質問に対して適切に回答しているかという観点から LLM の応答を分類した。まず、LLM の応答には、既存の社会調査の結果を引用したものが存在しており、それらの応答の大半は “%” を含んでいたことから “%” を含むものは回答拒否ではない応答として機械的に判定した。続いて、残りの応答に対し、Vicuna 13B v1.5 を用いた zero-shot 分類[6] により、回答拒否とみなせる応答と、そうでない応答に分類し、全応答に占める回答拒否となる応答の割合を算出した。結果を図 6 に示す。

国籍や宗教に比べ、人種/民族は回答の拒否率がいずれのモデルにおいても高いことが確認できる。このような結果となる理由として、特にアメリカにおいて、人種/民族に関するバイアスが大きな社会問題として認識されており [9]、モデル構築の際に RLHF を取り入れることで、LLM が人種/民族に関する質問に対して回答を拒否するようになったことが考えられる。また、全体的な傾向として、高い相関係数が得られた属性・モデルの組み合わせの方が回答拒否率が低い傾向があることが確認できる。このことは、LLM が質問に対し明示的な応答を避ける傾向が強くなると、LLM を社会実験の代替モデルとして利用することが困難となることを示している。

| | 国籍 | 宗教 | 人種/民族 |
|---|---|---|---|
| GPT-3.5-turbo | 0.13 | 0.10 | 0.52 |
| GPT-4 | 0.02 | 0.02 | 0.33 |
| Llama 2-Chat 13B | 0.10 | 0.13 | 0.59 |
| Llama 2-Chat 70B | 0.04 | 0.03 | 0.14 |
| Vicuna 13B v1.5 | 0.24 | 0.17 | 0.41 |

**図 6**　LLM の応答に占める回答拒否の割合.

## 5　おわりに

本研究では、国籍、宗教、人種/民族の 3 つの属性で規定される集団間の感情を LLM から抽出できるかの検証に取り組んだ。検証の結果、LLM から抽出された、日常的なテキスト表現により緩やかに規定された国籍や宗教に関する社会集団間の感情は、社会調査によって得られた集団間の感情との高い相関を示すことがわかった。一方で、LLM が関連する質問への回答を拒否する割合が高い人種/民族は、国籍や宗教に比べ、集団間の感情の抽出が難しいことが判明した。今後の展望として、社会集団間以外の集団間の感情抽出や、英語以外の言語での入力が感情抽出へ与える影響について検証を検討したい。

6)　zero-shot 分類の詳細および性能については付録 A に示す。

# 参考文献


[1] OpenAI. GPT-4 Technical Report. 2023.

[2] 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.

[3] Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In **Proceedings of the 40th International Conference on Machine Learning (ICML)**, pp. 337–371, 2023.

[4] John J Horton. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? **arXiv preprint arXiv:2301.07543**, 2023.

[5] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of One, Many: Using Language Models to Simulate Human Samples. **Political Analysis**, Vol. 31, No. 3, pp. 337–351, 2023.

[6] Junsol Kim and Byungkyu Lee. AI-Augmented Surveys: Leveraging Large Language Models for Opinion Prediction in Nationally Representative Surveys. **arXiv preprint arXiv:2305.09620**, 2023.

[7] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose Opinions Do Language Models Reflect? **arXiv preprint arXiv:2303.17548**, 2023.

[8] Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models. pp. 2611–2624, 2021.

[9] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In **Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)**, p. 610–623, 2021.

[10] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training Language Models to Follow Instructions with Human Feedback. In **Advances in Neural Information Processing Systems (NeurIPS)**, pp. 27730–27744, 2022.

[11] 新聞通信調査会. 第 9 回 諸外国における対日メディア世論調査, 2023.

[12] Pew Research Center. Americans Feel More Positive Than Negative About Jews, Mainline Protestants, and Catholics, 2023.

[13] Pew Research Center. Views on Race in America 2019, 2019.

[14] Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa-Anke, Fangyu Liu, Eugenio Martínez-Cámara, et al. TweetNLP: Cutting-Edge Natural Language Processing for Social Media. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations**, 2022.

## A　Zero-shot 分類の詳細

Vicuna 13B v1.5 を用いた LLM の応答の分類は図 7 に示すプロンプトを用いて行った。[Question]、[Answer] の部分に対応する質問、応答を入力し、“Label:” に続く出力の先頭 3 トークンに含まれる数字を予測ラベルとして扱った。

```
Reading a question-answer pair, classify the answer along the criterion below.
Label 0:  the answer only discusses the danger of generalization or biases.
Label 1:  the answer provides even just a little information to help understand the relationship between two groups.

Question: [Question]
Answer: [Answer]
Label:
```

**図 7**　Zero-shot 分類に用いたプロンプト

また、分類性能を評価するため、国籍間の感情に関する LLM の応答 200 例に対し人手によるアノテーションを行い、そのデータを用いて分類性能を評価した。アノテーションの際は、まず社会調査の結果を引用したと考えられるものをラベル 2 とし、残りの応答例を回答拒否 (ラベル 0) とそれ以外 (ラベル 1) に分類した。アノテーションは 2 名で実施し、2 名のラベルが一致した 179 例を用い、Vicuna 13B v1.5 による分類性能を評価した。分類性能の評価結果を表 3 に示す。

**表 3**　アノテーション済の回答の分類結果

| 正解 \ 予測 | ラベル 0 | ラベル 1 | ラベル 2 | 合計 |
|---|---|---|---|---|
| ラベル 0 | 11 | 1 | 0 | 12 |
| ラベル 1 | 11 | 141 | 0 | 152 |
| ラベル 2 | 0 | 0 | 15 | 15 |
| 合計 | 22 | 142 | 15 | 179 |