大規模言語モデルにより生成した疑似データを用いた自由記述アンケートの自動集約

銭本友樹 [1] 長谷川遼 [2] 宇津呂武仁 [1]
[1] 筑波大学大学院 システム情報工学研究群 [2] 筑波大学 理工学群 工学システム学類
{s2220753, s2012166}_@u.tsukuba.ac.jp utsuro_@iit.tsukuba.ac.jp

概要

自由記述回答を用いたアンケート調査は，調査対象における新たな価値や意見の発見に貢献する重要な手法である．この自由記述回答の分析作業は，回答中の意見抽出や類似意見のクラスタリングなどの複数の人手作業が必要であり，一般に大規模な回答データを対象とした分析には大きなコストがかかる．そこで本研究では，大規模言語モデルを用いた自由記述回答中の意見抽出と類似意見のクラスタリングの自動化手法を提案する．「新型コロナ不満アンケートデータ」の自由記述回答を利用した実験により，提案手法が人手による分析に匹敵する精度の分析を低コストかつ短時間で行えることを明らかにした．

1 はじめに

自由記述回答を用いたアンケート調査は，調査対象における新たな価値や意見の発見に貢献する重要な手法である [1]．このような自由記述回答の分析は質的データ分析（QDA） [2, 3] と呼ばれ，教育 [4] や医療 [5] などの様々な分野で行われている．しかしながら QDA では多くの複雑な人手作業が必要であり，大規模データを対象とした分析には非常に大きなコストがかかるという問題がある．

そこで本研究では，大規模言語モデル（LLM）[6] を用いて，QDA の代表的な手法であるグラウンデッド・セオリー・アプローチ (GTA) [7] の手順の一部を自動化する手法を提案する (図 1)．この手法を用いることで，これまでは困難であった大量の自由記述回答のアンケートから意見を抽出し分析する作業が可能となる．また，LLM を用いた提案手法と手作業による分析にかかる時間と費用を比較し，提案手法の有用性を検証する．

2 GTA の分析手順

グラウンデッド・セオリー・アプローチ（GTA）は，調査対象の様々な価値観や側面を理解するために使用される質的研究方法である [7, 8, 9, 10]．GTA の目的は，データを単に要約することではなく，アンケート回答やインタビュー対話などのデータに現れる現象のメカニズムを明らかにする「理論」を発見することにある．付録の表 4 のように，GTA には「コーディング」という以下 4 つの人手作業，(1) 回答を異なる意見ごとに分割する切片化，(2) 切片ごとのプロパティ (切片に含まれる様々な属性や概念) とディメンション (プロパティの実際の状態) の抽出，(3) データの内容を要約したラベルの付与，(4) 類似した内容の切片をまとめ，それらを端的に表すより抽象的なカテゴリの付与が必要である [1)]．本研究ではこれらの作業のうち，(1) 切片化と (4) カテゴリの付与の自動化を目標とする．

3 データセット

実験には，株式会社 Insight Tech が運営する Web サービス「不満買取センター」[2)] 上で 2020 年 3 月から 2021 年 1 月にかけて行われた COVID-19 に関するアンケート調査データ [11] を使用する．アンケートには，自由記述回答式と選択回答式の様々な質問が含まれている．本研究では，2020 年 3 月と 2020 年 6 月にそれぞれ収集されたアンケート中の自由記述回答型の質問「「新型コロナウイルス」に関して懸念していることや不満をお知らせください。」の 5,993 回答 [3)] を対象とし，回答中に含まれる不満と時期ごとの不満の変化について分析する．

1） GTA では通常，得られたカテゴリ間の関係の分析と，最終目標として新しい「理論」の発見が必要であるが，本論文ではこれらの問題は扱わない．
2） https://fumankaitori.com/
3） 2020 年 3 月は 2,996 回答，2020 年 6 月は 2,997 回答が収集された．

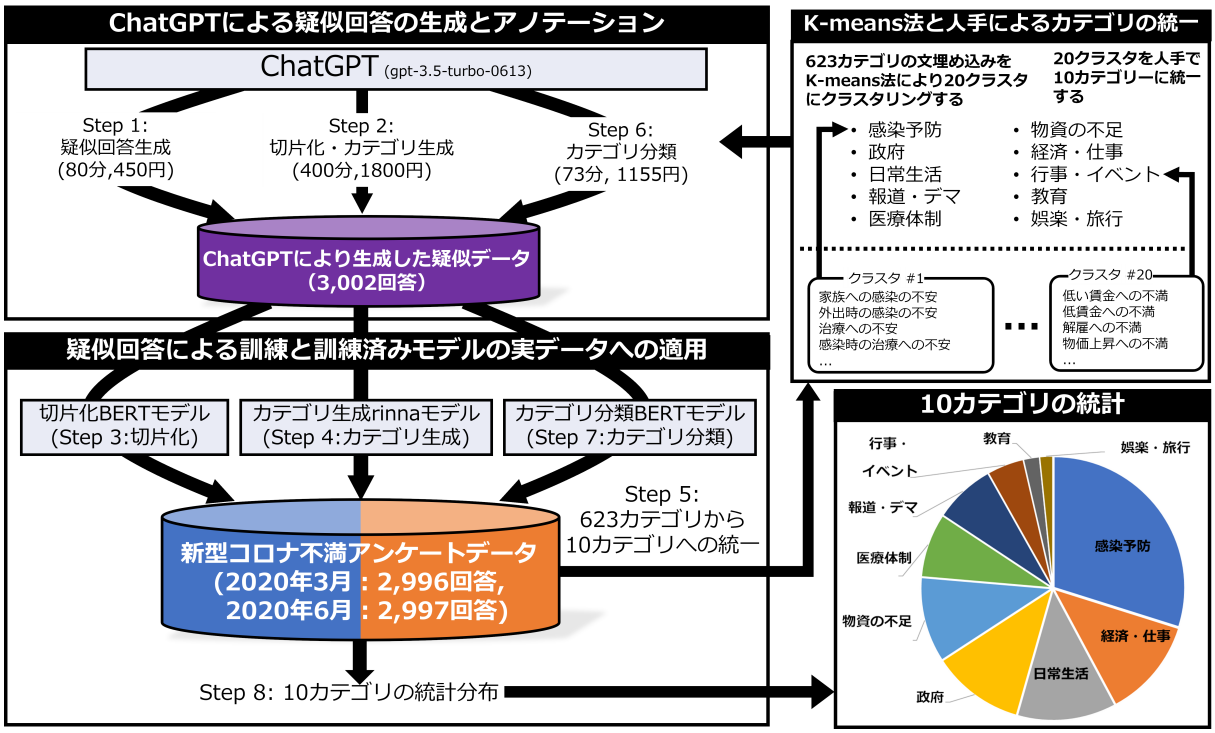

図 1　大規模言語モデルを用いた大規模な自由記述回答の分析の流れ

4　ChatGPT による擬似回答の生成

本研究で使用するアンケートデータは，利用規約により OpenAI 社が提供する ChatGPT [12] の入力に含めることができない．そのため本研究では，実回答を ChatGPT に直接入力するのではなく，ChatGPT を用いてローカル環境で動作可能な分析モデルを構築し，その分析モデルを用いて実回答を分析する．ChatGPT には gpt-3.5-turbo-0613 を用いる．

付録の表 5 に，疑似回答の生成に用いるプロンプトを示す．可能な限り実回答に類似した意見を生成するため，実回答で頻出したキーワードを含めた意見を生成するようにプロンプトを作成し，最終的に 3,002 件の疑似回答を生成した．

5　回答の切片化

実回答の切片化を自動で行うため，4 節で作成した疑似回答と ChatGPT を用いて，ローカル環境で動く切片化モデルを構築する．

5.1　アノテーション

切片化モデルの訓練と評価のため，以下 2 種類の切片化データを作成する [4]．

疑似回答自動切片化データ

4 節で作成した疑似回答を ChatGPT で切片化したデータを作成する．付録の表 6 に，ChatGPT で切片化を行うためのプロンプトを示す．切片化の結果 6,421 切片が得られた．

実回答人手切片化データ

実回答を人手で切片化したデータを作成する．この人手による切片化は，2020 年 3 月と 6 月それぞれの回答からランダムに 520 件ずつ抽出した回答を対象として行った．切片化の結果 1,716 切片が得られた．

5.2　切片化モデルの構築と評価

疑似回答自動切片化データを用いた切片化モデル (疑似回答切片化モデル) と実回答人手切片化データを用いた切片化モデル (実回答切片化モデル) を構築し，その性能を比較する．切片化モデルには事前学習済み言語モデルである東北大版 BERT [13] [5] を使用する．疑似回答切片化モデルは 8 割を訓練データ，2 割を検証データとして扱い，検証データでの損失が最小のモデルを実回答に対して適用する．実回答切片化モデルは，8 割を訓練データ，1 割を検証データ，残り 1 割をテストデータとして扱い 10

4)　ChatGPT と人手ともに，実際には切片化，プロパティとディメンション抽出，ラベル付与，カテゴリ付与のすべてのアノテーションを同時に行っている．人手アノテーションは第二著者によって行われ，費用は時給 1,500 円で計算した．

5)　https://huggingface.co/cl-tohoku/bert-base-japanese-v3

表 1　切片化モデルの性能比較

モデル	適合率	再現率	F1 値
疑似回答切片化モデル	55.7	77.7	64.9
実回答切片化モデル	67.5	69.2	68.4

分割交差検証を行う．

表 1 に，実回答人手切片化データを評価データとした 2 つの切片化モデルの評価結果を示す．表 1 から，疑似回答切片化モデルよりも，実回答切片化モデルの方が F1 値は高かった．疑似回答切片化モデルは再現率が高く，元の回答を過剰に分割する傾向があった．また，疑似回答切片化モデルを実回答全体の 5,997 回答に適用したところ，12,490 切片に分割された．

6　カテゴリの生成と統一

GTA では類似した内容の切片をまとめ，その内容を端的に表す抽象的なカテゴリを付与する．しかし現段階では，データ全体にどのような内容がどの程度存在するかが不明のため，カテゴリの名前や数を指定することができない．そこで，まず切片ごとにその内容を端的に表す抽象的なカテゴリを自由に生成し，データ全体にどのような内容がどの程度存在するかを分析する．

6.1　アノテーション

ローカル環境で動くカテゴリ生成モデルを構築するため，5.1 節で作成した疑似回答自動切片化データに対して，ChatGPT を用いてカテゴリ生成を行う．付録の表 6 に，ChatGPT で切片化を行うためのプロンプトを示す．

6.2　カテゴリ生成モデルの構築

カテゴリ生成モデルには，大規模言語モデルである rinna 社の bilingual-gpt-neox-4b-instruction-sft [6] を使用し，LoRA [14] を用いて訓練を行う [7]．6.1 節で作成したデータの 8 割を訓練に使用し，残りの 2 割を検証データとして扱い，検証データでの損失が最小のモデルを実回答に対して適用する．

6.3　実回答への適用とカテゴリの統一

5.2 節で作成した疑似回答切片化モデルにより切片化された実回答の 12,490 切片に対してカテゴリ生成モデルを適用する．その結果，計 623 種類のカテゴリが生成された．生成されたカテゴリの例を図 1 の右上の表中に示す．続いて，この 623 カテゴリ中の類似したカテゴリを統一する．類似したカテゴリを見つけるため，623 カテゴリの文字列を文ベクトルに変換し，その文ベクトルに対して k-means クラスタリングを適用する．各カテゴリの文埋め込みには，日本語 Sentence-BERT モデル [15][8] を利用する．まず k-means 法により 623 カテゴリを 20 種類のクラスタにクラスタリングした後，この 20 クラスタを第一著者が確認し，第一著者の判断で図 1 に示されている 10 種類のカテゴリに統一した．

表 2　カテゴリ分類モデルの性能比較

モデル	正解率
疑似回答カテゴリ分類モデル	64.9
実回答カテゴリ分類モデル	80.7

7　カテゴリの分類

最後に，5.2 節で作成した疑似回答切片化モデルにより切片化された実回答の 12,490 切片を，6.3 節で統一した 10 種類のカテゴリに分類する．

7.1　アノテーション

カテゴリ分類モデルの訓練と評価のため，以下 2 種類のカテゴリ分類データを作成する．

疑似回答自動カテゴリ分類データ

5.1 節で作成した 6,421 件の疑似回答自動切片化データを ChatGPT を用いて 10 カテゴリに分類する．付録の表 8 に ChatGPT でカテゴリ分類を行うためのプロンプトを示す．

実回答人手カテゴリ分類データ

5.1 節で作成した 1,716 件の実回答人手切片化データを人手で 10 カテゴリに分類する．

7.2　カテゴリ分類モデルの構築と評価

疑似回答自動カテゴリ分類データを用いたカテゴリ分類モデル (疑似回答カテゴリ分類モデル) と実回答人手カテゴリ分類データを用いたカテゴリ分類モデル (実回答カテゴリ分類モデル) を構築し，その性能を比較する．使用する言語モデルと訓練方法は，5.2 節の切片化モデルの構築と同様である．

6)　https://huggingface.co/rinna/bilingual-gpt-neox-4b-instruction-sft
7)　LoRA のハイパーパラメータは，$r = 16$，$\alpha = 16$ とした．
8)　https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

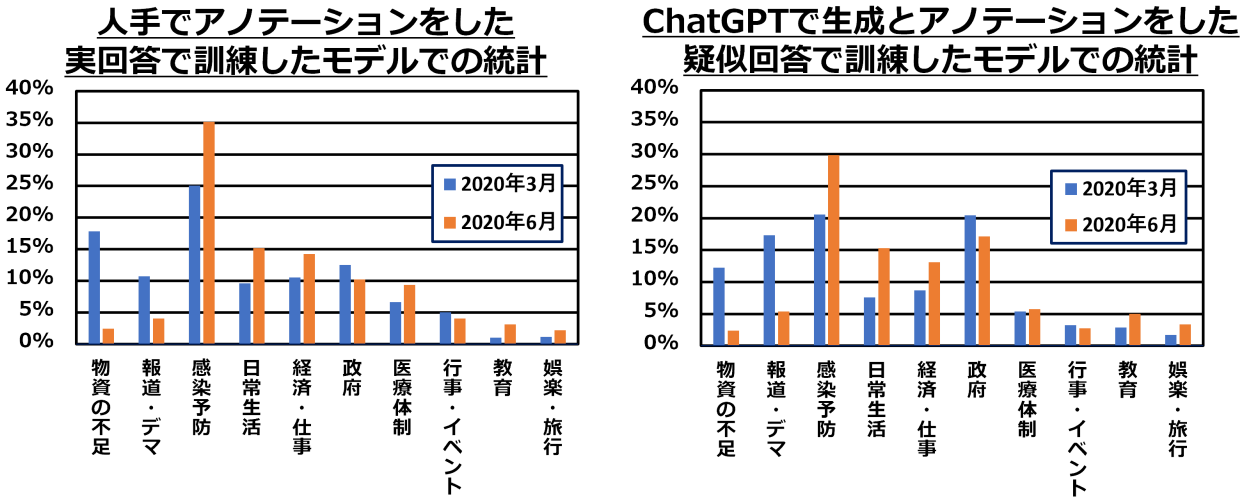

図 2　実回答の全 5,997 回答に対して切片化とカテゴリ分類を適用した結果の統計比較

表 3　データ作成に必要な時間と費用の比較

タスク	人手 (1,040 回答)		提案手法 (3,002 回答)	
	時間 (分)	費用 (円)	時間 (分)	費用 (円)
疑似回答の生成	—	—	80	450
切片化・カテゴリ生成	2,340	58,500	400	1,800
カテゴリ分類	87	2,163	73	1,155
合計	2,427	60,663	553	3,405

表 2 に，実回答人手カテゴリ分類データを評価データとした 2 種類のモデルの評価結果を示す．表 2 から，疑似回答カテゴリ分類モデルは実回答カテゴリ分類モデルよりも正解率が低いことがわかる．疑似回答カテゴリ分類モデルは，切片中の「政府」や「経済」などの単純な単語の有無でカテゴリを分類する傾向があり，切片全体の文脈を考慮できていないことが多かった．

8　実回答への適用と統計分析

最後に，実回答の全 5,997 回答に対して，実回答切片化モデルと実回答カテゴリ分類モデルを適用した結果 (図 2 左) と，疑似回答切片化モデルと疑似回答カテゴリ分類モデルを適用した結果 (図 2 右) を比較する．これらの図の比較から，ChatGPT を用いて自動で作成したモデルは，実際の回答を人手でアノテーションして作成したモデルと同等の統計結果を得られていることがわかる．特に，2020 年 3 月と 6 月のカテゴリ間の大小関係は全てのカテゴリで一致していた．加えて表 3 に人手と提案手法それぞれのデータ作成に必要な時間と費用の比較結果を示す．この表から，提案手法は人手の 4 分の 1 以下の時間と 17 分の 1 以下の費用での分析が可能であり，非常に効率的であることがわかる．しかしながら図 2 から，提案手法は政府カテゴリを誤って多く分類してしまう問題があることがわかる．これは，疑似回答中に政府に関する不満が多く含まれていたことが原因だと考えられる．

9　おわりに

本研究では，本来多くの複雑な人手作業が必要である質的データ分析の一部を大規模言語モデルを用いて自動化する手法を提案した．また提案手法は人手分析と比較して，時間と費用の面で非常に優れていることを明らかにした．提案手法を用いることで，大規模なアンケートの自由記述回答を対象として，データ全体にどのような意見がどの程度含まれているか不明な状態から，図 2 のようなカテゴリごとの統計を出すことが可能である．

今後は疑似回答の生成方法を改善することで，切片化モデルとカテゴリ分類モデルの性能の向上を図りたい．提案手法の切片化とカテゴリ分類の性能は，どちらも人手で作成したデータを用いたモデルよりも性能が低かった．提案手法の各モデルの性能は ChatGPT により生成した疑似回答の品質に大きく依存しているため，疑似回答全体の言語的特徴をより実回答に近づける工夫や，データ数の増加により各モデルの性能向上は可能だと考えられる．また本研究では，グラウンデッド・セオリー・アプローチにおける切片化とカテゴリ分類のみしか自動化できていないため，プロパティとディメンションの抽出，ラベル生成，カテゴリーの統一，すべての作業の自動化を目指したい．

謝辞

本研究では，NII の IDR データセット提供サービスにより株式会社 Insight Tech から提供を受けた「不満調査データセット」を利用した．

参考文献

[1] Ursa Reja, Katja Manfreda, Valentina Hlebec, and Vasja Vehovar. Open-ended vs. close-ended questions in web questionnaires. **Adv Methodol Stats**, Vol. 19, , 01 2003.

[2] Michael Quinn Patton. **Qualitative Research & Evaluation Methods: Integrating Theory and Practice**. SAGE Publications, Inc., 4th. edition, 2014.

[3] Jane Ritchie, Jane Lewis, Carol McNaughton Nicholls, and Rachel Ormston. **Qualitative Research Practice A Guide for Social Science Students and Researchers**. SAGE Publications, Inc., 2nd. edition, 2014.

[4] Gaurav Nanda, Aparajita Jaiswal, Hugo Castellanos, Yuzhe Zhou, Alex Choi, and Alejandra J. Magana. Evaluating the coverage and depth of latent dirichlet allocation topic model in comparison with human coding of qualitative data: The case of education research. **Machine Learning and Knowledge Extraction**, Vol. 5, No. 2, pp. 473–490, 2023.

[5] Chao Fang, Natasha Markuzon, Nikunj Patel, and Juan-David Rueda. Natural language processing for automated classification of qualitative data from interviews of patients with cancer. **Value in Health**, Vol. 25, No. 12, pp. 1995–2002, 2022.

[6] 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.

[7] Anselm L. Strauss. **Qualitative Analysis for Social Scientists**. Cambridge University Press, 1987.

[8] Juliet Corbin and Anselm L. Strauss. **Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory**. SAGE Publications, Inc., 3rd. edition, 2008.

[9] Shigeko Saiki-Craighill. Qualitative nursing research in Japan: A state of the science and indications for future directions. In **Routledge International Handbook of Qualitative Nursing Research**, pp. 597–609. Taylor and Francis, April 2013.

[10] Shigeko Saiki-Craighill. Overview of grounded theory approach. **Keio SFC Journal**, Vol. 14, No. 1, pp. 30–43, 2014. (in Japanese).

[11] Insight Tech Ltd. Discontent questionnaire data on COVID-19. informatics research data repository, national institute of informatics. (dataset)., mar 2021.

[12] OpenAI. GPT-4 technical report, 2023.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[14] Ning Ding, Xingtai Lv, Qiaosen Wang, Yulin Chen, Bowen Zhou, Zhiyuan Liu, and Maosong Sun. Sparse low-rank adaptation of pre-trained language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 4133–4145, Singapore, December 2023. Association for Computational Linguistics.

[15] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)**, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

A 付録

表 4 グラウンデッド・セオリー・アプローチに基づく自由記述回答の分析例

回答	切片	プロパティ：ディメンション	ラベル	カテゴリ
学校が休みになり，子供の勉強の事が心配 … 小学校の対応が悪い.	学校が休みになり，子供の勉強の事が心配 … 小学校の対応が悪い.	感情 : 心配 学校の状態 : 休み 心配の対象 : 子供の勉強 小学校の対応 : 悪い	小学校の対応への不満	教育
出かけるところがへった.	出かけるところがへった.	出かけるところ：減った	外出の難しさへの不満	日常生活

表 5 ChatGPT で疑似回答を生成するためのプロンプト (実際には具体例を 3 つ含めている)

「新型コロナウイルス」に関して懸念していることや不満をお知らせください.

というアンケートの回答とそのカテゴリを可能な限りたくさん生成してください.

回答には [keyword] を含めてください.

回答は可能な限り長い文章にし，多様な意見を含めてください.

回答:

. . .

表 6 ChatGPT で疑似回答に対して切片化，プロパティとディメンションの抽出，ラベル名の付与，カテゴリの生成を行うプロンプト

The subsequent statement constitutes a response to the query: 「新型コロナウイルス」に関して懸念していることや不満をお知らせください.

Firstly, split the answer into different opinions. Then, extract the properties, dimensions, label and category from the answer.
Sentence: 感染のリスクを甘く見ている人が多く見受けられます．特に，高齢者に感染した場合，致死率が高くなることを重視するべきだと感じています.

Opinion1: 感染のリスクを甘く見ている人が多く見受けられます.

Property=Dimension: 懸念の対象=感染リスクの軽視

Label: 感染リスクへの懸念

Category: 感染リスク

Opinion2: 特に，高齢者に感染した場合，致死率が高くなることを重視するべきだと感じています.

Property=Dimension: 懸念の対象=高齢者の致死率

Label: 高齢者が感染することへの懸念

Category: 高齢者の感染リスク

. . .

表 7 rinna モデルでカテゴリ生成を行うためのプロンプト

指示:

ユーザー: 以下の入力を文脈を考慮して要約してください。

システム: 分かりました。

ユーザー: 入力: [input]; 文脈: [context]

システム: [output]

表 8 カテゴリ分類を行うためのプロンプト (実際にはカテゴリごとに 1 つの具体例を含めている)

The subsequent statement constitutes a response to the query: 「新型コロナウイルス」に関して懸念していることや不満をお知らせください.

Clustering the target opinion into following 10 categories

- 政府
- 物資の不足
- 日常生活
- 経済・仕事
- 報道・デマ
- 感染予防
- 医療体制
- 行事・イベント
- 教育
- 娯楽・旅行

Full Sentence: マスクをせずに外出する人が多くて，これ以上の感染拡大が心配．特に電車やバスの中でマスクをしない人が多いのは困る．[SEP] マスクをするのは自己防衛のためだけでなく，他人への感染予防のためでもあることをもっと認識してほしい．Target Opinion: マスクをするのは自己防衛のためだけでなく，他人への感染予防のためでもあることをもっと認識してほしい.

Category: 感染予防

. . .

表 9 人手でカテゴリ分類した実回答の具体例

カテゴリ	2020 年 3 月	2020 年 6 月
物資の不足	マスクやトイレットペーパーなどいつも買えるものが買えない。	マスクの供給は元通りになったように感じるが、結局価格が上がったままコロナ前の値段には戻っていない。
報道・デマ	デマ情報が多すぎて、日常生活まで (食料品の品薄など) 今まで通りに過ごせなくなりそうで怖い。	テレビなどの煽りと言っても良い放送が異常すぎるので、もう少しまともな放送をしてほしい。
感染予防	いつか自分も罹患してしまうのではないかと怯えています。クルーズ船から降りてきた人に、自分勝手な行動が多いこと。	第 2 波がくるかもしれないので、怖がっています。感染の第 2 波第 3 波と夜の繁華街からの感染者が増えている事。
日常生活	満員電車に乗ることが恐怖。外出しづらい。	更に電車の混雑が戻ってるからテレワークでできるところは強制して欲しい。いつまでもダラダラと続く自粛が辛いがやっぱり感染したくない。