Coding Open-Ended Responses using Pseudo Response Generation by Large Language Models

Yuki Zenimoto[1], **Ryo Hasegawa**[2], **Takehito Utsuro**[2]
Masaharu Yoshioka[3], **Noriko Kando**[4]
[1]Nagoya University, [2]University of Tsukuba
[3]Hokkaido University, [4]National Institute of Informatics
zenimoto.yuki.u1_@_s.mail.nagoya-u.ac.jp, s2420791@u.tsukuba.ac.jp
utsuro_@_iit.tsukuba.ac.jp yoshioka_@_ist.hokudai.ac.jp, kando_@_nii.ac.jp

Abstract

Survey research using open-ended responses is an important method that contributes to the discovery of unknown issues and new needs. However, survey research generally requires time and cost-consuming manual data processing, indicating that it is difficult to analyze large dataset. To address this issue, we propose an LLM-based method to automate parts of the grounded theory approach(GTA), a representative approach of the qualitative data analysis. We generated and annotated pseudo open-ended responses, and used them as the training data for the coding procedures of GTA. Through evaluations, we showed that the models trained with pseudo open-ended responses are quite effective compared with those trained with manually annotated open-ended responses. We also demonstrate that the LLM-based approach is highly efficient and cost-saving compared to human-based approach.

1 Introduction

In the qualitative data analysis (QDA) (Patton, 2014; Ritchie et al., 2014), survey research based-on open-ended questionnaire responses is an essential method for the discovery of unknown issues and new needs. The grounded theory approach (GTA) (Strauss, 1987), a representative method of the QDA, requires several manual complex procedures, referred to as "coding." As a result, analyzing large qualitative data becomes exceptionally challenging. In addition, for most confidential information such as survey responses, the use of external APIs like ChatGPT (OpenAI, 2023) is prohibited by terms of service.

Therefore, we propose a pipeline that can automate parts of the grounded theory approach (Strauss, 1987; Corbin and Strauss, 2008). To avoid the obstacle of not being able to use external APIs such as ChatGPT, we firstly generated and annotated pseudo open-ended responses, and used them as the training data for the coding procedures of GTA as shown in Figure 1. Through evaluations, we showed that the Grounded Theory Approach Pipeline trained with pseudo open-ended responses are quite effective compared with those trained with manually annotated open-ended responses. We also demonstrate that the proposed pipeline is highly efficient and cost-saving compared to human-based approach. The code is available at `https://github.com/Zeni-Y/naacl2024-coding-open-ended`

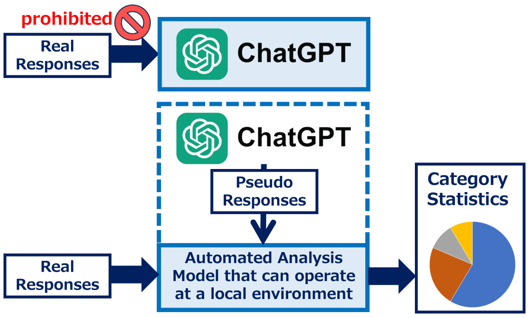


Figure 1: Overview of Coding an Open-Ended Responses utilizing Pseudo Responses generated by ChatGPT

2 Related Work

Initial studies on coding automation proposed symbolic approaches based on rules created by researchers or statistical approaches based on corpora (Inui et al., 2003; Crowston et al., 2012). However, these approaches required considerable effort to develop rules, and such rules were not adoptable to different domains. To address these issues, more versatile methods using supervised learning have been proposed to label and cluster qualitative data (Stenetorp et al., 2012; Klie et al., 2018; He and Schonlau, 2020). Additionally, approaches that combine a rule-based method with a machine learning method to assist human coding tasks have also been proposed (Rietz and Maedche,

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 242–254
June 18, 2024 ©2024 Association for Computational Linguistics

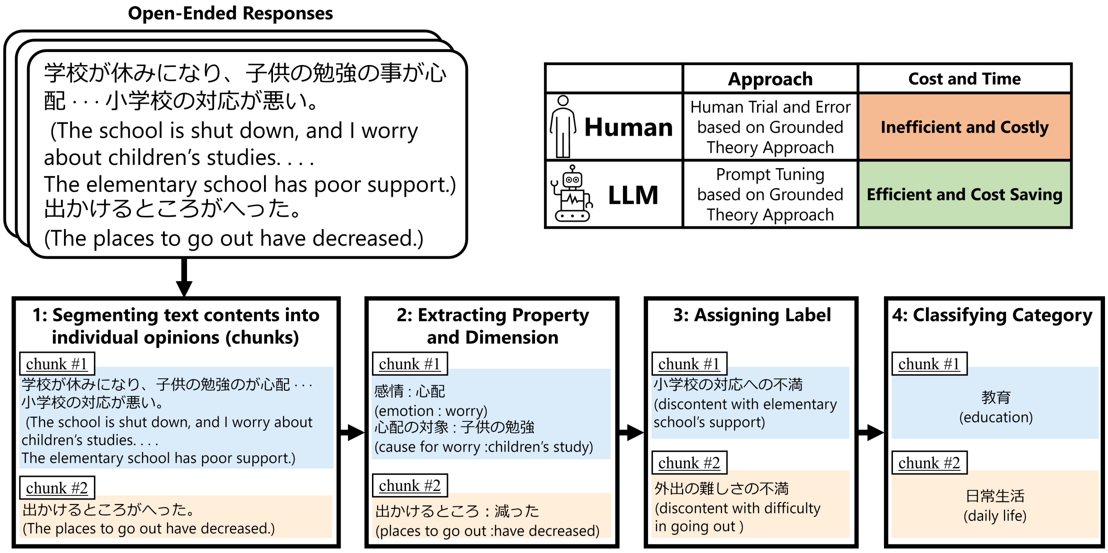

Open-Ended Responses
学校が休みになり、子供の勉強の事が心配・・・小学校の対応が悪い。
(The school is shut down, and I worry about children’s studies. . . .
The elementary school has poor support.)
出かけるところがへった。
(The places to go out have decreased.)
Approach
Cost and Time
Human
Human Trial and Error based on Grounded Theory Approach
Inefficient and Costly
LLM
Prompt Tuning based on Grounded Theory Approach
Efficient and Cost Saving
1: Segmenting text contents into individual opinions (chunks)
chunk #1
学校が休みになり、子供の勉強のが心配・・・小学校の対応が悪い。
(The school is shut down, and I worry about children’s studies. . . .
The elementary school has poor support.)
chunk #2
出かけるところがへった。
(The places to go out have decreased.)
2: Extracting Property and Dimension
chunk #1
感情 : 心配
(emotion : worry)
心配の対象 : 子供の勉強
(cause for worry :children’s study)
chunk #2
出かけるところ : 減った
(places to go out :have decreased)
3: Assigning Label
chunk #1
小学校の対応への不満
(discontent with elementary school’s support)
chunk #2
外出の難しさの不満
(discontent with difficulty in going out)
4: Classifying Category
chunk #1
教育
(education)
chunk #2
日常生活
(daily life)

Figure 2: Coding an Open-Ended Response on COVID-19: Comparison between by GTA and by an LLM (“Property”, “Dimension”, “Label”, and “Category” are terminology of GTA, where “Property” represents the various important aspects in the raw data, “Dimension” the variation of the property, “Label” the summary of the response, and “Category” a small number of classes that aggregate all the set of labels.)

2021; Gebreegziabher et al., 2023). Nevertheless, researchers still had to design the types and definitions of labels and clusters.

In terms of automatically determining labels and clusters, there exists research that uses the LDA topic model (Blei et al., 2003) to generate preliminary category suggestions and support human QDA procedures (Nanda et al., 2023). However, as is known in GTA, which involves properties, dimensions, labels, and categories, coding tasks that require aggregating information from bottom up demand an advanced linguistic interpretation ability to capture the various aspects of the data. Thus, simply using BERT or the LDA topic model is insufficient to conduct coding tasks.

Large language models (LLMs) have achieved high performance in tasks similar to various steps in qualitative analysis, such as text clustering (Viswanathan et al., 2023; Zhang et al., 2023), text summarization (Stiennon et al., 2020), aspect-based sentiment analysis (Hosseini-Asl et al., 2022). Additionally, numerous comparative experiments between manual annotations and LLM-generated annotations have been conducted, and it has been shown that the LLMs can annotate with an accuracy comparable to that of humans (Pan et al., 2023; Gilardi et al., 2023; Ding et al., 2023a). Therefore, it can be inferred that LLMs have the potential for automating qualitative analysis. However, there has not been any research on using an LLM to automate qualitative analysis approaches.

3 Coding Procedure of GTA

GTA aims not just to summarize data but to discover a “theory” that elucidates the mechanism by which the phenomena appears in the data such as questionnaire responses and interview dialogues. As shown in Figure 2, GTA requires the following complex manual procedures, referred to as “coding”: (1) segmenting text contents into individual opinions (referred to as “chunk”), (2) extracting attributes and concepts from the chunks, (3) assigning labels that summarize the content of the chunks, (4) classifying similar chunks into more abstract categories. In this study, we aim to automate (1) segmentation and (4) classification.

4 Dataset

4.1 Real Response

We use following two types of open-ended response data collected on the web service “Fuman Kaitori Center (FKC)” operated by Insight Tech Ltd[1].

COVID-19 Discontent Data[2] This dataset consists of open-ended responses related to discontent regarding COVID-19. In this study, we use the 1,040 responses (540 responses collected in March 2020 and 540 responses collected in June 2020.)

[1] https://fumankaitori.com/

[2] https://www.nii.ac.jp/dsc/idr/fuman/fuman_covid19.html

Dataset	Real Response			Pseudo Response	
	#Unannotated Response	#Annotated Response	#Annotated Chunk	#Annotated Response	#Annotated Chunk
COVID-19 Discontent Data	5,993	1,040	1,716	800	1,550
General Discontent Data	5,250,000	1,000	1,000	1,000	1,000

Table 1: Statistics on Annotated Responses and Chunks

BERT fine-tuned with	P	R	F1
(a) real responses	67.5	69.2	**68.4**
(b) pseudo responses	55.7	77.7	64.9

Table 2: Evaluation Results of Segmentation Models

General Discontent Data[3] This dataset consists of open-ended responses expressing various everyday discontent. Each response is categorized into one of 29 categories, such as "Living and Housing" and "Food and Beverages." Segmenting is not necessary as each response has only one opinion. In this study, we use 10,000 responses randomly extracted from the 10 most frequently occurring categories[4].

4.2 Pseudo Response Generation

It is prohibited to re-distribute the real responses due to the terms of use. This can be considered that the dataset can not be processed using the external LLMs services such as ChatGPT. Therefore, we generate pseudo open-ended responses by ChatGPT(*gpt-3.5-turbo-0613*) to construct local models that can process the real responses.

We design the prompt that generates pseudo open-ended responses that are similar to the real responses. Firstly, we extracted the sets of keywords contained in the real responses. Then, we created prompts designed to generate responses that include those sets of keywords. These sets consist of nouns contained within a single response[5]. Table 6 of Appendix A shows a prompt for generating pseudo open-ended responses.

4.3 Annotation

In this study, we compare the pseudo responses generated and annotated by ChatGPT with the real responses annotated manually through the tasks of segmenting and clustering. A summary of dataset statistics is shown in Table 1.

4.3.1 Manual Annotation

We manually annotate the real responses. Firstly, we manually segment the real responses into chunks. Next, to each of those chunks, we manually annotate a single most appropriate category as well as multiple categories each of which is considered to be appropriate. This set of category is defined through the k-means clustering and manual selection described in Section 5.2. The single most appropriate category is used as a strict criterion of evaluating accuracy, where the reference is a single category. The multiple categories, on the other hand, are used as a looser criterion of evaluating accuracy, where the reference consists of multiple categories and an estimated category is judged as correct when it is among those multiple reference categories[6].

4.3.2 Annotation by ChatGPT

We automatically annotate the pseudo responses using ChatGPT. Firstly, we segment the pseudo responses into chunks and generate a category from each chunks using ChatGPT. Table 7 of Appendix A shows a prompt for segmenting and generating a category. Next, to each of those chunks, we annotate a single category by ChatGPT. Table 8 of Appendix A shows a prompt for classification.

5 Experiments

In this section, we described the each step in the proposed pipeline. The flow of this pipeline is shown in Figure 3.

5.1 Segmenting a Response into Chunks

We construct two types of segmentation models: a segmentation model using pseudo responses seg-

[3] https://www.nii.ac.jp/dsc/idr/fuman/fuman.html

[4] These 10 categories are as follows: "Eating Out and Stores," "Living and Housing," "Hobbies and Entertainment," "Industry and Sector," "Food and Beverages," "Public and Environment," "Human Relations," "Digital and Electronics," "Fashion," "Beauty and Health"

[5] We limited the number of keywords included in one response to a maximum of five.

[6] We measure the inter-annotator agreement of the annotation between the first and second authors of the paper, achieving a Kappa score (Fleiss et al., 1969) of 0.936 for the single most appropriate category, suggesting that there is no significant disagreement between the two annotators for the single most appropriate category.

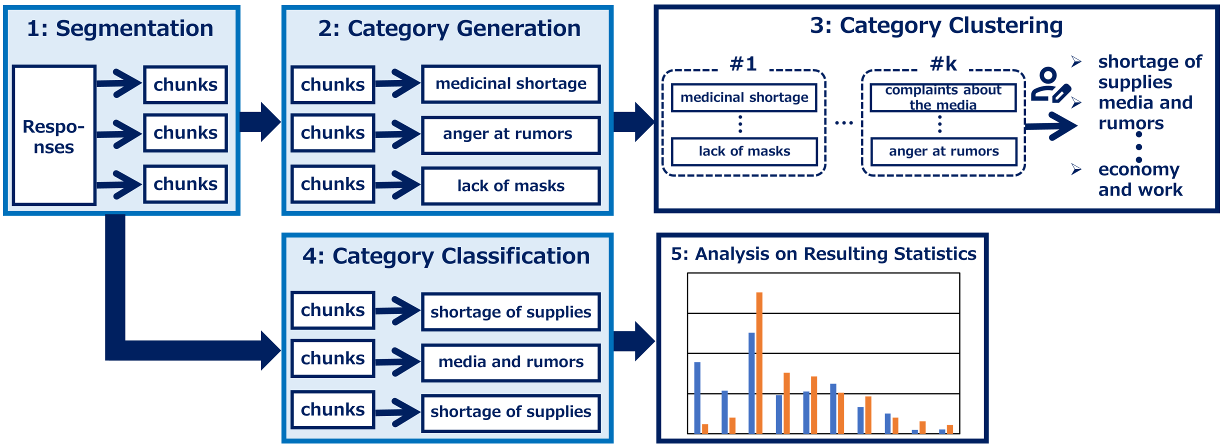


Figure 3: Flow of Grounded Theory Approach Pipeline

mented automatically, and a segmentation model using real responses segmented manually. We then compare the performance of these two models. As the model for segmenting a response into chunks, we use a pre-trained BERT (Devlin et al., 2019), i.e., Tohoku University's Japanese version of BERT-base[7][8][9]. In the training of these models, 80% of the responses are used for training, while the remaining 20% are used for validation and the model with the minimum loss on the validation data is selected.

In Table 2, we present the evaluation results of the two segmentation models. From Table 2, it is observed that the segmentation model using real responses achieved a higher F1 score than the segmentation model using pseudo responses. In addition, the segmentation model using pseudo responses has a high recall rate, indicating a tendency to overly segment the responses.

5.2 Category Generation and Clustering

In GTA, similar content chunks are grouped together, and an abstract category that succinctly represents their content is assigned. However, it is unclear what kind of content exists and to what extent throughout the data, making it impossible to specify the names and numbers of categories. Therefore, the initial step involves freely generating abstract categories that succinctly represent the content of each segment, and then applying unsupervised clustering to these categories to determine the appropriate categories.

As the model for generating a category from a chunk of real responses, we use a GPT (Brown et al., 2020; Black et al., 2022)[10] model, where we employ LoRA (Ding et al., 2023b)[11] (with the hyper-parameter $r = 16$ and $\alpha = 16$) for reducing the cost of fine-tuning and the number of parameters. In the training of those models, 80% of the pseudo responses above are used for training, while the remaining 20% are used for validation and the model with the minimum loss on the validation data is selected.

Then, we apply the k-means clustering method[12] to the sentence embeddings of the strings of those generated categories obtained in the previous section, which are then clustered into 20 clusters. For the construction of the sentence embedding of each category string, we utilize the Japanese Sentence-BERT model (Reimers and Gurevych, 2019)[13]. Finally, we manually select and merge those 20 clusters into 10 categories shown in Table 3.

5.3 Category Classification

We construct two types of category classifying models: one using pseudo-responses segmented automatically, and the other using real responses segmented manually. We then compare the performance of these two models. As the model for classifying chunks into categories, we use a pre-

[7] https://huggingface.co/cl-tohoku/bert-base-japanese-v3

[8] The batch size is set to 64 sentences, and the maximum input token length is set to 256 tokens.

[9] We also evaluated a base-sized Japanese RoBERTa model (Liu et al., 2019) of https://huggingface.co/rinna/japanese-roberta-base in the same task, where the performance was almost similar to that of BERT

[10] https://huggingface.co/rinna/bilingual-gpt-neox-4b-instruction-sft

[11] https://github.com/microsoft/LoRA

[12] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[13] https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

COVID-19 Discontent Data	General Discontent Data
物資の不足(Shortage of Supplies)	外食・店舗(Eating Out and Stores)
報道・デマ(Media and Rumors)	暮らし・住まい(Living and Housing)
感染予防(Infection Preverntion)	趣味・エンタメ(Hobbies and Entertainment)
日常生活(Daily Life)	業界・業種(Industry and Sector)
経済・仕事(Economy and Work)	食品・飲料(Food and Beverages)
政府(Government)	公共・環境(Public and Environment)
医療体制(Medical Infrastructure)	人間関係(Human Relations)
行事(Events)	デジタル・家電(Digital and Electronics)
教育(Education)	ファッション(Fashion)
娯楽・旅行(Entertainment and Travel)	美容・健康(Beauty and Health)

Table 3: Category Lists that are Determined Manually

BERT fine tuned with	General Discontent Data		COVID-19 Discontent Data	
	single category	multiple category	single category	multiple category
(a) real responses	65.0	80.0	80.7	91.4
(b) pseudo responses	52.0	68.0	64.1	80.0

Table 4: Evaluation Results of Category Clustering (“single category” represents a strict criterion of evaluating accuracy, where the reference is a single category, while “multiple categories” represents a looser criterion of evaluating accuracy, where the reference consists of multiple categories and an estimated category is judged as correct when it is among those multiple reference categories.)

trained BERT (Devlin et al., 2019), i.e., Tohoku University’s Japanese version of BERT-base. In the training of these models, 80% of the chunks are used for training, while the remaining 20% are used for validation and the model with the minimum loss on the validation data is selected. Finally, we apply the trained model to the chunks of real responses and obtain the statistics of the 10 categories.

In Table 4, we present the evaluation results of the two types of models. From Table 4, it is clear that the classifying model using pseudo responses have a lower accuracy compared to the classifying model using real responses in the both of General Discontent Data and COVID-19 Discontent Data. The classifying models using pseudo responses tend to classify categories based on the presence or absence of simple words within the chunks, such as “government” or “economy”, and often fails to consider the context of the entire text.

5.4 Analysis on Resulting Statistics

As a result, we construct following two types of GTA pipeline:

Real Response Pipeline A pipeline composed of a segmentation model and a category classifying model, both constructed using real responses annotated manually.

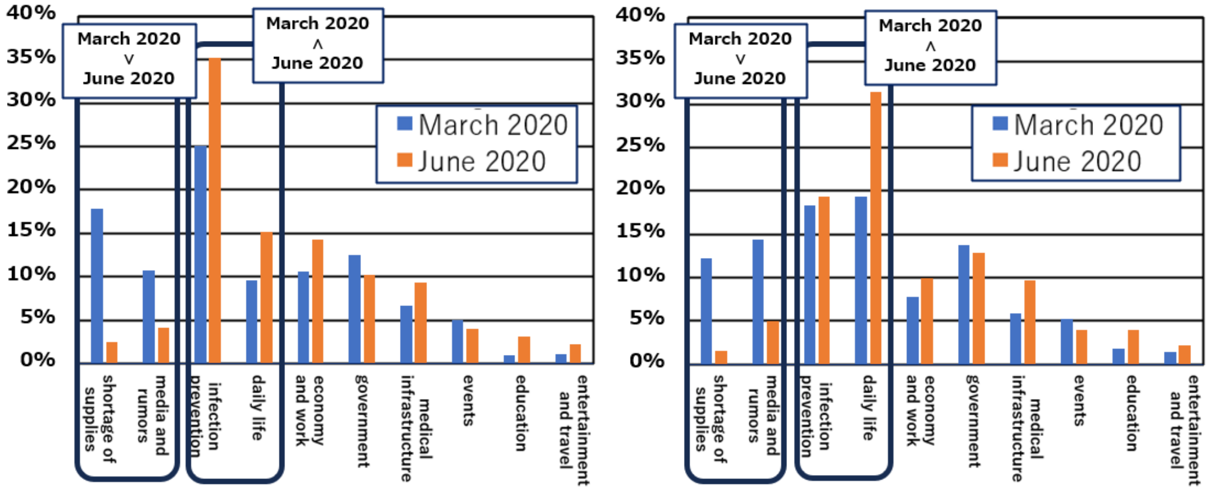
Pseudo Response Pipeline A pipeline composed of a segmentation model and a category classifying model, both constructed using pseudo responses generated and annotated by ChatGPT.

We apply these two pipelines to the entire set of real responses of COVID-19 Discontent Data. The left side of Figure 4 shows the statistics result of real response pipeline, while the right side shows the statistics result of pseudo response pipeline. Comparing these figures reveals that the pipeline created automatically using ChatGPT are capable of producing statistical results comparable to those obtained from manually annotated real responses. Notably, the increase/decrease relationships between categories in March and June 2020 were consistent across all categories. However, Figure 4 indicates that the pipeline using pseudo responses has a problem with over-classification in daily life. The cause of this problem is considered to be that our proposed pseudo response generation method inappropriately generated large number of responses related to daily life.

Additionally, Table 5 presents a comparison of the time and cost required to create data using manual methods and the proposed automated method. From this table, it is evident that the proposed method can perform in less than one-thirteenth of

Task	General Discontent Data				COVID-19 Discontent Data			
	Manual		ChatGPT		Manual		ChatGPT	
	Time (Mins)	Cost (USD)	Time (Mins)	Cost (USD)	Time (Mins)	Cost (USD)	Time (Mins)	Cost (USD)
Response Generation	—	—	3	0.1	—	—	3	0.1
Segmentation	—	—	—	—	225	37.5	13	0.4
Category Clustering	8.4	1.4	2	0.26	8.4	1.4	2.4	0.26
Total	8.4	1.4	**5**	**0.36**	233	38.9	**18.4**	**0.76**

Table 5: Comparison of Time and Cost Required for Data Creation (per 100 responses)



(a) by the Model Trained with 832 Real Responses Annotated Manually

(b) by the Model Trained with 800 Pseudo Responses Generated and Annotated by ChatGPT

Figure 4: Statistics Comparison Generated by Real Response Pipeline and Pseudo Response Pipeline for open-ended responses on COVID-19.

the time and at less than about one-fiftieth of the cost of manual methods, making it highly efficient.

6 Conclusion

This study proposed a pipeline for automating parts of Grounded Theory Approach, which typically requires many complex manual tasks. It also demonstrated that the proposed pipeline is significantly superior in terms of time and cost when compared to manual analysis. By employing the proposed method, it is possible to easily generate statistics, from a state where it is unclear what kind of content exists and to what extent throughout the data.

Future work will aim to improve the pseudo response generation method to enhance the performance of both the segmentation and category classification models. The performances of the segmentation model and classification model using pseudo responses were lower than that of models using real responses. Refining the linguistic features of the pseudo-responses to more closely resemble real responses could potentially improve model performance since the performance of each model heavily depends on the quality of the pseudo responses generated by ChatGPT. Furthermore, this study has only automated the segmentation and category classification within the Grounded Theory Approach. Future efforts should aim to automate the extraction of properties and dimensions, label generation, and the unification of categories to automate all tasks.

Acknowledgments

In this paper, we used “General Discontent Dataset” provided by Insight Tech Ltd. via IDR Dataset Service of National Institute of Informatics. The research was supported by ROIS NII Open Collaborative Research 2024 (24FS01), the Telecommunications Advancement Foundation, JSPS KAKENHI Grant Number 21H00901 and 23H00994.

References

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An opensource autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

Kevin Crowston, Eileen E. Allen, and Robert Heckman. 2012. Using natural language processing for qualitative data analysis. *International Journal of Social Research Methodology*, 15(6):523–543.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023a. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Ning Ding, Xingtai Lv, Qiaosen Wang, Yulin Chen, Bowen Zhou, Zhiyuan Liu, and Maosong Sun. 2023b. Sparse low-rank adaptation of pre-trained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4133–4145, Singapore. Association for Computational Linguistics.

Joseph L. Fleiss, Jacob Cohen, and B. S. Everitt. 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72:323–327.

Simret Araya Gebreegziabher, Zheng Zhang, Xiaohang Tang, Yihao Meng, Elena L. Glassman, and Toby Jia-Jun Li. 2023. PaTAT: Human-AI collaborative qualitative coding with explainable interactive rule synthesis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–19, New York, NY, USA. Association for Computing Machinery.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):1–3.

Zhoushanyue He and Matthias Schonlau. 2020. Automatic coding of open-ended questions into multiple classes: Whether and how to use double coded data. *Survey Research Methods*, 14(3):267–287.

Ehsan Hosseini-Asl, Wenhao Liu, and Caiming Xiong. 2022. A generative language model for few-shot aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 770–787, Seattle, United States. Association for Computational Linguistics.

Hiroko Inui, Masao Utiyama, and Hitoshi Isahara. 2003. Criterion for judging request intention in response texts of open-ended questionnaires. In *Proceedings of the Second International Workshop on Paraphrasing*, pages 49–56, Sapporo, Japan. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Gaurav Nanda, Aparajita Jaiswal, Hugo Castellanos, Yuzhe Zhou, Alex Choi, and Alejandra J. Magana. 2023. 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*, 5(2):473–490.

OpenAI. 2023. GPT-4 technical report.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? Measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 26837–26867. PMLR.

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

Nils Reimers and Iryna Gurevych. 2019. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Tim Rietz and Alexander Maedche. 2021. Cody: An ai-based system to semi-automate coding for qualitative research. In *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, New York, NY, USA. Association for Computing Machinery.

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

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Anselm L. Strauss. 1987. *Qualitative Analysis for Social Scientists*. Cambridge University Press.

Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2023. Large language models enable few-shot clustering.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. ClusterLLM: Large language models as a guide for text clustering.

A Prompts for Generating Pseudo Open-ended Responses and Segmenting into Chunks and Generating a Category from a Chunk /Classifying a Chunk into the 10 Categories

Table 6 shows an example of the prompt for generating pseudo open-ended responses. Table 7 shows an example of the prompt for segmenting a response into chunks and generating a category from a chunk. Table 8 shows an example of the prompt for classifying chunks of the pseudo responses into the 10 categories.

B Example Responses to the COVID-19 Discontent Data that are Representative for Each of 10 Categories

As illustrated in Table 9, for the category "shortage of supplies", majority of responses are about discontent with the shortage of supplies such as masks both in March and June 2020, while their rates are extremely higher in March 2020 than in June 2020. For the category "media and rumors", discontent with fake and misleading information is mostly observed in March 2020, while that with provocative media broadcasts is observed both in March and June. For the categories "infection prevention", on the other hand, more and more responses are closely related to infection risk and spread regarding the second and third eddies in June 2020 than in March 2020. In the category "daily life", majority of responses are closely related to discontent with overcrowded trains as well as self-restraint issues, where their numbers are increasing more in June 2020.

instruction	「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。 (Please express your concerns or dissatisfaction regarding the “COVID-19”.) というアンケートの回答とそのカテゴリーを可能な限りたくさん生成してください。 Create as many responses and associated categories as possible for the survey. 以下の例を参考にしてキーワードとなる「マスク;感染者;通勤電車、生活」を含めた回答を3つ生成してください。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。 Based on the following examples, generate three responses that include the keywords “mask, infected individuals, commuter trains, daily life.” Ensure the responses are as lengthy as possible and incorporate a diverse range of opinions.
# 1 of 2 gold standard samples	Keyword:症状(symptoms), 高齢者(elderly), 感染(infection), 致死率(mortality rate), テレビ(TV), デマ(rumor), トイレットペーパー(toilet paper), マスク(mask), 報道(media), 政府(government), 医療体制(medical infrastructure) Response1: 若い人は症状が軽いからといっても高齢者に感染させてしまうので、高齢者に感染すると致死率が高いということを、テレビはもっと報道して若年層に周知させるべきだと思う。それと、デマのせいでトイレットペーパーやマスクが買えなくなってるので、むやみに市民を煽るような報道は控えてほしい。政府も適切な報道と対応をしてほしい。政府は現場の意見を適切に聞いて、医療体制や物資の不足などの現状を正確に把握してほしい。 (There is a pressing need for television broadcasters to underscore the grim reality that, although younger individuals may only exhibit mild symptoms, the transmission of the virus to the elderly results in a significantly higher mortality rate. Thus, raising awareness among the younger demographic is imperative. Moreover, the proliferation of misinformation has led to a scarcity of essentials such as toilet paper and masks, which has unnecessarily incited public hysteria - a trend in media reporting that must be curtailed. It is incumbent upon the government to not only disseminate accurate information but also to grasp the actual state of the healthcare system and address the shortage of medical supplies.) Category:報道への要望(requests for responsible journalism)，物資の不足への不満(grievances about the scarcity of necessities)，政府の対応への不満(dissatisfaction with governmental response)，政府の対応への不満(criticism of government handling of the crisis)
# 2 of 2 gold standard samples	Response2: · · · · · · · · · · · · · · ·

Table 6: A 2-Shot Prompt for Generating Pseudo Open-ended Responses by ChatGPT (English translation of Japanese sentences is given only for explanation but not included in the actual prompts.)

instruction	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。 (Please express your concerns or dissatisfaction regarding the “COVID-19”.) Firstly, split the answer into different opinions. Then, extract the properties, dimensions, label and category from the answer.
# 1 of 3 gold standard samples	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。テレビやその他メディアの報道は、しばしばデマや不必要な情報を流してしまい、社会に混乱を招いています。トイレットペーパーやマスクの品不足はその典型例で、人々の無用なパニックを引き起こしているのです。 (There appears to be a prevalent underestimation of the risk associated with infection. It is particularly imperative to emphasize the increased mortality rate in cases where the elderly are infected. The reporting by television and other media outlets often disseminates misinformation and superfluous details, precipitating chaos within society. The quintessential examples of this phenomenon are the shortages of toilet paper and masks, which have triggered unnecessary panic among the populace.) 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。 (There appears to be a prevalent underestimation of the risk associated with infection.) Property=Dimension: 懸念の対象(Subjects of Concern)=感染リスクの軽視(The Underestimation of Infection Risk) Label:感染リスクへの懸念(Concerns pertaining to the risk of infection) Category:感染リスク(Infection risk) 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。 (It is particularly imperative to emphasize the increased mortality rate in cases where the elderly are infected.) Property=Dimension:懸念の対象(Subjects of Concern)=高齢者の致死率(The Mortality Rate Among the Elderly) Label:高齢者が感染することへの懸念(Apprehensions regarding the infection of the elderly) Category:高齢者の感染リスク(The risk of contagion in senior populations) Opinion3: テレビやその他メディアの報道は、しばしばデマや不必要な情報を流してしまい、社会に混乱を招いています。 (The reporting by television and other media outlets often disseminates misinformation and superfluous details, precipitating chaos within society.) Property=Dimension:情報(Information)=デマ(The Dissemination of Misinformation); 結果(Consequence)=社会の混乱(Social Unrest) Label:メディア報道への不満(Dissatisfaction with media coverage) Category:報道(Reportage) Opinion4: トイレットペーパーやマスクの品不足はその典型例で、人々の無用なパニックを引き起こしているのです。 (The quintessential examples of this phenomenon are the shortages of toilet paper and masks, which have triggered unnecessary panic among the populace.) Property=Dimension:問題(Issue)=物資の不足(Scarcity of Supplies);結果(Consequence)=パニック(Public Panic) Label:品不足によるパニック(Panic induced by product shortages) Category:物資の不足(Deficiency of commodities)
# 2 of 3 gold standard samples	Sentence: ··· Opinion1: ··· ....
# 3 of 3 gold standard samples	Sentence: ··· Opinion1: ··· ....

Table 7: A 3-Shot Prompt for Segmenting a Response into Chunks and Generating a Category from a Chunk by ChatGPT (English translation of Japanese sentences is given only for explanation but not included in the actual prompts.)

instruction	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。 (Please express your concerns or dissatisfaction regarding the “COVID-19”.) Clustering the target opinion into following 10 categories - 政府(Government) - 物資の不足(Shortage of Supplies) - 日常生活(Daily Life) - 経済・仕事(Economy and Work) - 報道・デマ(Media and Rumors) - 感染予防(Infection Prevention) - 医療体制(Medical Infrastructure) - 行事・イベント(Events) - 教育(Education) - 娯楽・旅行(Entertainment and Travel)
# 1 of 1 gold standard sample	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> (The prevalence of individuals venturing outdoors without masks is alarming, with a potential escalation in transmission rates as a cause for concern. This issue is particularly pronounced in confined spaces such as trains and buses, where the incidence of non-mask wearers is notably high. <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。 <SEP> (It is imperative for the public to develop a heightened awareness that mask usage serves not solely as a means of self-protection but also as a vital mechanism to prevent the transmission of infection to others. <SEP>) それから、生活必需品の買い占めが起こっているが、これも生活に支障が出て困っている。トイレットペーパーや食料品が品切れ状態で手に入らないことがある。こういった買い占めは本当に必要な人に対しての配慮が欠けていると思う。<SEP> (Additionally, the phenomenon of stockpiling essential goods has led to significant inconveniences in day-to-day life. There are instances where necessities such as toilet paper and groceries are unavailable due to stockout conditions. Such hoarding behavior reflects a lack of consideration for those in genuine need. <SEP>) 政府は、適切に物流を管理し、供給を安定させる努力をしてほしい。 (It is incumbent upon the government to exercise appropriate control over logistics and endeavor to stabilize the supply chain.) Target Opinion: それから、生活必需品の買い占めが起こっているが、これも生活に支障が出て困っている。トイレットペーパーや食料品が品切れ状態で手に入らないことがある。こういった買い占めは本当に必要な人に対しての配慮が欠けていると思う。 (Additionally, the phenomenon of stockpiling essential goods has led to significant inconveniences in day-to-day life. There are instances where necessities such as toilet paper and groceries are unavailable due to stockout conditions. Such hoarding behavior reflects a lack of consideration for those in genuine need.) Category: 物資の不足(Resource scarcity)

Table 8: A 1-Shot Prompt for Classifying Chunks of Pseudo Responses into the 10 Categories by ChatGPT (English translation of Japanese sentences is given only for explanation but not included in the actual prompts.)

(a-1) category: shortage of supplies

Marjory of responses are about discontent with the shortage of supplies such as masks both in March and June 2020, while the rate of the category "shortage of supplies" is much higher in March 2020 than in June 2020.

An example response of March 2020

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。
(Ordinary purchasable items such as masks and toilet paper have become unattainable.)

An example response of June 2020

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。
(Supply chains for masks seem to have revived, however, despite this, prices remain elevated and have not returned to pre-pandemic levels.)

(a-2) category: media and rumors

Discontent with fake and misleading information is mostly observed in March 2020, while that with provocative media broadcasts is observed both in March and June.

An example response of March 2020

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。 (The prevalence of false information, extending even to everyday life elements such as food shortages, incites fear as it appears to endanger the continuity of customary lifestyle.)

An example response of June 2020

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。 (Given the excesses of sensationalism spearheaded by outlets such as television, it would be appeasing to see more responsible broadcasting.)

(a) Comparison of Rates in Figure 4: March 2020 are Higher than June 2020

(b-1) category: infection prevention

More and more responses are closely related to infection risk regarding the second and third eddies in June 2020 than in March 2020.

An Example response of March 2020

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 (I harbor a profound trepidation that I might someday contract the disease.)
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。
(There is an abundance of selfish actions observed among individuals who have descended from the cruise ship.)

An example response of June 2020

第2波がくるかもしれないので、怖がっています。 (Amidst the potential advent of a second wave, I find myself immersed in trepidation.)
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。 (An increasing incidence of infection from nocturnal entertainment districts accompanying the second and third waves of the pandemic is observed.)

(b-2) category: daily life

Majority of responses are closely related to discontent with overcrowded trains as well as self-restraint issues, where their numbers are increasing more in June 2020.

Example responses of March 2020

満員電車に乗ることが恐怖。 (Boarding packed trains incites fear and anxiety.)
外出しづらい。 (Venturing outdoors has become increasingly difficult.)

Example responses of June 2020

更に電車の混雑が戻ってるからテレワークできるところは強制して欲しい。 (Furthermore, the resurgence in train congestion necessitates the adoption of teleworking measures, wherever proven feasible.)
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。 (The seemingly interminable period of self-restraint has engendered a level of discomfort, indeed, yet the fear of contracting the infection persists.)

(b) Comparison of Rates in Figure 4: June 2020 are Higher than March 2020

Table 9: Example Responses to the COVID-19 Discontent Data that are Representative for Each of 13 Categories