質的研究の自動化:
患者自由記述テキストからの潜在的トピックの発見

橋本清斗 [1]　清水聖司 [1]　工藤紀子 [1]　矢田竣太郎 [2]
若宮翔子 [1]　江本駿 [3]　西村由希子 [3]　荒牧英治 [1]
[1] 奈良先端科学技術大学院大学　[2] 筑波大学　[3] 特定非営利活動法人 ASrid
{hashimoto.kiyoto.hi3}@naist.ac.jp　yada@slis.tsukuba.ac.jp
{shimizu.seiji.so8, noriko.kudo, wakamiya, aramaki}@is.naist.jp
{emoto, yucko}@asrid.org

概要

医学研究における自由記述テキストデータは，患者・家族や医療従事者の声を直接反映し，新たな知見の発見や意思決定，施策立案において重要な役割を果たす．しかし，これらのデータを対象とした質的データ分析は，膨大な人的労力を要する．本研究では，希少・難治性疾患患者が新型コロナウイルス感染症の流行期間中に経験した困難を分析対象とし，大規模言語モデル（Large Language Model; LLM）を活用して質的データ分析の自動化を試みた．具体的には，患者・家族の自由記述テキストの各文に対してタグを生成し，タグ間の類似性に基づいて段階的に統合を行う手法を提案した．結果として，提案手法は分析時間を大幅に削減しつつ，人手による分析結果との一定の一致を示した．本研究の成果は，LLM を活用することで，自由記述テキストの効率的な分析手法の実現に向けた基盤を構築するものである．

1　はじめに

医学研究は，ランダム化試験 (RCT) を始めとした厳密な定量的な研究と，患者の感情や発話を読み解く定性的な研究の 2 つに大別される．前者は系統的な方法が定められており，自動化の試みが注目されている [1]．一方，定性的な研究では，質的データ分析を用いることで，患者や医療従事者の自由記述テキストから抽出された新たな知見が，意思決定や施策設計に活用されている [2, 3]．

既存の質的データ分析手法は，類似データのまとめ上げ，統合といったプロセスを繰り返すことにより，データの構造化を行うものであり，グラウンデッド・セオリー・アプローチ（Grounded Theory Approach; GTA）[4, 5] をはじめとした手続きが定義されている．しかし，人手で行われるため，高度な経験と膨大な時間を要する．

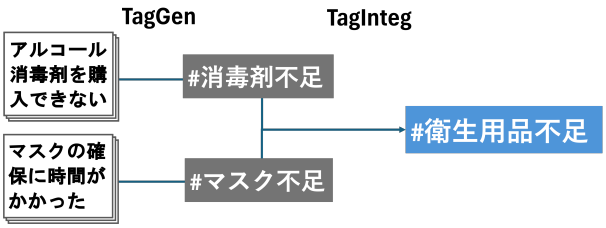

図 1　自由記述テキストの（**TagGen**: タグの生成）と（**TagInteg**: タグの統合）のプロセス：患者自由記述テキストから抽出したタグを段階的に統合し，上位の抽象的なカテゴリを形成するプロセスを示す．

そこで本研究では，「希少・難治性疾患患者のコロナ禍の困難のアンケート」を対象とし，医学分野における質的研究の自動化を試みる．具体的には，LLM を用いた，自由記述テキストに対するタグ付け・意見の統合手法 [6] を応用し，アンケート結果に対するタグの生成 (**TagGen**)・統合（**TagInteg**）を行う（図 1）．これにより，「希少・難治性疾患患者が経験した困難」を構造化する．実験では，提案手法で構造化された結果が，熟達した人間によるものと一致するかを検証する．本研究の成果は，LLM を活用した質的データ分析の可能性を示すものである．

2　手法

本研究では，LLM を用いることで，患者・家族の自由記述の回答データを質的分析する手法を提案する．提案手法は，タグの生成（2.1 節）とタグの統合（2.2 節）から構成される．本章では，この 2 つのプロセスについて詳細に説明する．

Algorithm 1 Tag Generation （TagGen）

Require: A set of free-text descriptions D
Ensure: A set of tagged text pairs T

$T \leftarrow \emptyset$
$L \leftarrow \emptyset$
for each text $d \in D$ **do**
  if Classifiable(d) = False **then**
    $tag_d \leftarrow$ Generate(d)
    $L \leftarrow L \cup tag_d$
  else
    $tag_d \leftarrow$ Classify(d, L)
    $T \leftarrow T \cup \{(d, tag_d)\}$
  end if
end for
return T

2.1 タグの生成: TagGen

タグの生成プロセスでは，自由記述テキストに対しおおまかな要約結果をタグとして付与することでデータの概観を把握する．具体的には，Algorithm 1 に示すように，自由記述テキスト集合 D に対して，LLM を用いて以下の操作を繰り返し適用する．

分類可能性の判定 - Classifiable(d): テキスト d に対して，既に生成されたタグ集合のいずれかに分類できるかを判定する. 分類できる場合は，そのタグを付与する (Classify(d, L))．分類できない場合は，次の「タグの生成」を行う．

タグの生成 - Generate(d): いずれのタグにも分類できないテキストに対して，新たなタグを生成する．具体的には，テキストを入力として，テキスト中に表現されている「患者・家族が経験した困難」という観点から， 20 文字以下のタグを生成する．

このように，常に既存のタグ集合を参照し，新たな生成を最小限に抑えることで，タグの冗長性を削減できる．

2.2 タグの統合: TagInteg

生成されたタグを基に，類似タグを統合する．Algorithm 2 に示すように，この統合プロセス (**TagInteg**) を段階的に繰り返す：

統合候補の選定 - Candidates(L): 生成されたタグリストを用い，意味的に近いタグペアを選定する．タグ間の類似性評価は，LLM を用いてタグの意味的な類似性をもとに 0 から 100 の範囲でスコア化する (Similarity(l_i, l_j)) ．類似度スコアが閾値 θ 以上のペアを統合を行うペアとして決定する．

含意関係の判定による統合 - Integration(l_i, l_j): 統合対象のペアを統合する．片方のタグが含意関係でより大きな集合を表す場合，そのタグを選択する．含意関係がない場合は，両タグを含意する新たなタグを生成する．

例えば，図 1 では，消毒剤不足とマスク不足の 2 つのタグを統合する際に，消毒剤とマスクは含意関係にないので，両者を含意する「衛生用品不足」というタグが生成される．

再分類 - ReClassify(d, L): 作成された新たなタグリストに対し再分類を行う．出力揺れに対応し，一貫性を保った分類を行うため，以下の操作を実施する：

- 自由記述テキストとタグリストをプロンプトとして LLM に入力し，最適なタグを選択させる．出力は新たに作成したタグリストから選択した単一のタグのみとし，余分な文字や説明文を排除する．
- LLM が出力したタグが既存タグリストに存在するか，または 20 文字未満であるかを確認し，これに合致しない場合は「未分類」として扱う．

統合されない（あるいは統合候補にならない）タグは元のままリストに残り，最終的に再分類の対象となる．このように，処理を複数の段階に分割することで，安定性が向上し，データの整合性および分類結果の一貫性を維持することが可能となる．

本実験は，ASrid の倫理審査委員会にて承認を受けたものである (承認番号: ap240901301).

Algorithm 2 Tag Integration and Reclassification (TagInteg)

Require: A set of free-text descriptions D, an existing tag list L, a threshold θ
Ensure: A set of tagged text pairs T with the integrated tag set L'

$C \leftarrow$ Candidates(L)
for each pair $(l_i, l_j) \in C$ **do**
  simScore $\leftarrow$ Similarity(l_i, l_j)
  if simScore $\geq \theta$ **then**
    $L \leftarrow$ Integration(l_i, l_j, L)
  end if
end for
$T \leftarrow \{ \big(d,$ ReClassify$(d, L)\big) \mid d \in D \}$
return T

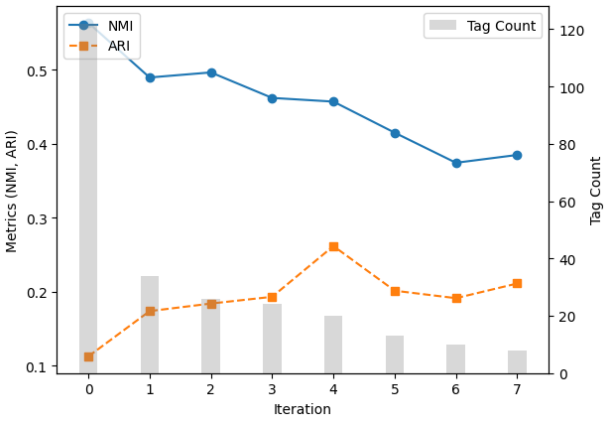

図 2　タグ数とメトリクスの推移: NMI と ARI は統合操作の試行回数によって変動し，最適な試行回数を見つけることが重要であることが示唆される．

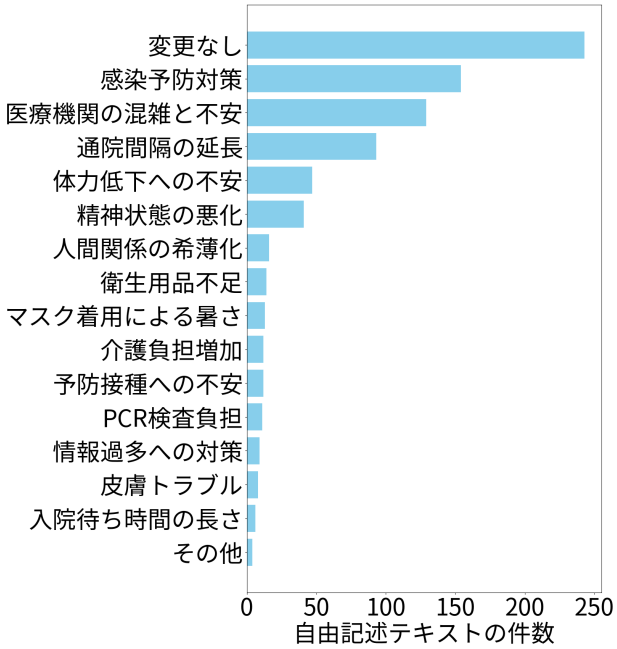

図 3　LLM によって生成されたタグとそのタグに分類されたテキストの統計．適切なタグが生成されテキストもバランスされている．

3　実験

本章では実験に用いた自由記述テキストの詳細（3.1 節）と 2 章で述べた提案手法の有効性を定量的に評価する方法（3.2 節）について説明する．本実験では ‘gemma2:27b-instruct-fp16’ モデル [7] を使用した．

3.1　実験材料

本研究で用いたデータは，特定非営利活動法人 ASrid（以下，ASrid）が実施した調査研究に基づく．この調査は，希少・難治性疾患領域の患者およびその家族が新型コロナウイルス感染症（COVID-19）の影響をどのように受けたかを明らかにすることを目的として実施された．本研究では，2020 年 5 月から 2021 年 1 月までの 9 か月間にわたり，希少・難治性疾患の患者およびその家族 110 名から取得した 813 件の自由記述テキストを用いた．個人情報への配慮として収集者である ASrid のほうで個人情報を排した形にした．自由記述テキストには，全般的な不安感や診療に関する不安に関する記述が含まれる．

実験では，これらの自由記述テキストに対し 2 章で述べた 2 つのプロセスを実行し構造化を行う．比較評価のために，813 件のデータのうち 251 件に人手でタグを付与した．タグ付けした 251 件の自由記述テキストについて，提案手法を用いて構造化し，3.2 節に示す定量的な評価を行った．

3.2　評価方法

評価では，提案手法によって生成されたタグ集合と，人手で作られたタグ集合間で，類似した意味合いを持つタグ同士をマッピング（例：「通院困難」≒「通院が困難になった」）し，各評価データに付けられたタグを正解ラベルとした．ここで行われたマッピングの結果の詳細は付録に記載した（表 2）．この正解ラベルを用いて，提案手法のタグ付けを，分類問題として Cohen’s Kappa と F1 スコアを用いて評価した．各要素のマッピングは LLM（GPT-4o）を用いて行い，明確な対応がないタグは「その他」とした．

さらに，タグ同士のマッピングを必要としない，クラスタリング性能の評価指標である NMI および ARI による評価も行った．

NMI (Normalized Mutual Information): クラスタ間の分布の一致度を測定する．

ARI (Adjusted Rand Index): クラスタリング結果と正解ラベルの一致度を，ランダムな一致を考慮して調整する．

4　結果

図 2 に，251 件の自由記述テキストに対して人手で付与されたタグと自動で付与されたタグとの一致度（NMI と ARI）を示す．実験にあたり最適な試行回数を探索するため，**TagInteg** の試行回数を 7 回に設定した．

統合操作の初回試行では NMI と ARI が比較的高い値を示したが，試行回数を重ねるごとに NMI は

表 1 $i = 4$ 時点でのマッピング後の各評価指標の値：人手で作成した正解ラベル付きデータと本研究で提案したプロセスの実行結果が中程度の一致を示している．

NMI	ARI	Cohen's Kappa	F1 Score
0.4165	0.2700	0.4592	0.5151

徐々に低下し，ARI は 4 回目の試行で最高値を記録した後，やや低下する傾向が見られた．GPT-4o を用いたタグ同士のマッピングによる評価では，最も ARI の値が高かった $i = 4$ 時点での結果を採用した．

人手で作成した正解ラベル付きデータとマッピングを行なった評価結果を表 1 に示す．F1 スコアは 0.5151 であり，Cohen's Kappa は 0.4592 であった．これらの結果から，提案手法により生成されたタグと正解ラベル付きデータの間には一定の一致度があることが示された．

次に，全データ 813 件に対して提案手法を適用し，集計した結果を図 3 に示す．集計結果の最も上位にある「変更なし」というタグは，COVID-19 による影響が無かったとする意見が分類されたものである．その一方で「感染予防対策」や「通院期間の延長」など通院が必須となる希少・難治性疾患領域の患者に特有の不安感も見られた．他にも「情報過多への対策」や「衛生用品不足」など，同時期に社会問題として挙げられていた不安感も見られた．

5 関連研究

本研究では，質的研究の自動化を目指し，「タグの生成」と「タグの統合」からなる手法を提案した．前者は短文要約 (Short Text Summarization) [8–10]，後者は含意関係認識 (Entailment) [11] や意味的テキスト類似度 (Semantic Textual Similarity, STS) [12, 13] としてそれぞれ研究されている．

近年，これらの個別に扱われてきたタスク（短文要約，含意関係認識，STS など）を LLM を活用して統合的に処理するアプローチが注目を集めている [6, 14, 15]．本研究は，これらの手法を医学分野における質的研究に応用することを主たる目的としている．「タグの生成」と「タグの統合」それぞれに最適化された手法を用いることで，個別プロセスの精度向上も期待されるが，本研究では質的研究の新たな枠組みの提案に主眼を置いている．このため，すべてのプロセスを同一の LLM によって一貫して実装する方針を採用した．

6 考察

実験結果から，生成されたタグと正解データは一定の一致を示した．特に，統合操作の試行回数が増えると NMI と ARI の値が変動し，最適な回数を探索する重要性が示唆された．

本研究の課題として，まず，データセットが特定の調査に基づくため，異なるドメインや自由記述テキストで同様の結果が得られるかは不明である．また，タグの生成と統合に使用した LLM の性能はモデルの種類や設定に依存するため，他のモデルや設定での再現性も検証が必要である．

さらに，タグの生成・統合のプロセスも，5 節で述べた短文要約，含意関係認識，STS などの最新の研究を応用することにより改善できる可能性がある．本研究では，既存の質的データ分析手法に基づきプロンプトを設計し，最終出力のみを評価した．今後，生成されたタグと自由記述テキストの合致度や含意関係の正確性などの詳細評価を行い，使用モデルの選定やアルゴリズムの改善が必要である．

その他にも，質的研究には分析者の主観によるバイアスが課題とされており，質的研究を LLM により代替した本研究にも同様のバイアスの可能性が考えられる．このような人と LLM のバイアス差異が施策設計などのタスクに与える影響については，慎重な検証が必要であると考えられる．

本研究の成果は，LLM を活用した質的データ分析の可能性を示すものであり，今後の質的研究の効率化と精度向上に寄与すると期待される．

7 おわりに

本研究では，患者・家族の自由記述テキストデータを用いた質的研究の自動化手法を提案し，その有効性を検証した．具体的には，LLM を用いた自由記述テキストに対するタグの生成と統合により，「希少・難治性疾患患者が経験した困難」を自動で構造化可能であることを示した．今後の研究では，新たなアルゴリズムの導入を通じて，タグの生成と統合プロセスの精度向上を図る．また，異なるドメインやデータセットに対する適用可能性を検証し，より汎用性の高い手法の開発を目指す．

本研究の成果は，LLM を活用した質的データ分析の効率化に寄与し，従来は膨大な労力と高度な経験に頼っていた質的研究を，今後大きく発展させる可能性があると信じている．

謝辞

本研究は，特定非営利活動法人 ASrid が実施した調査研究のデータを使用して行われた．ASrid にはデータ提供にあたり多大なるご協力をいただいたことを深く感謝いたします．本研究は，「戦略的イノベーション創造プログラム（SIP)」「統合型ヘルスケアシステムの構築」JPJ012425，JST CREST「リアルワールドテキスト処理の深化によるデータ駆動型探薬」JPMJCR22N1 の支援を受けたものである．

参考文献

[1] Faith Mutinda, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. **AUTOMETA: Automatic Meta-Analysis System Employing Natural Language Processing**, Vol. 290. 06 2022.

[2] 由美礒村, 雅恵堤, 千鶴永田. 意思伝達能力の低下した高齢者の意思を看護師がくみ取り援助を展開するプロセス, June 2020.

[3] 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, December 2022.

[4] Barney G. Glaser and Anselm L. Strauss. **The Discovery of Grounded Theory: Strategies for Qualitative Research**. Aldine Publishing Company, Chicago, 1967.

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

[6] Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. TopicGPT: A Prompt-based Topic Modeling Framework, 2023. Version Number: 2.

[7] Google DeepMind Gemma Team. Gemma 2: Improving open language models at a practical size. 2024. Available at https://storage.googleapis.com/deepmind-media/gemma/gemma-2-report.pdf.

[8] Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL Recognising Textual Entailment Challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, Florence d'Alché Buc（編）, **Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment**, 第 3944 巻, pp. 177–190. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. Series Title: Lecture Notes in Computer Science.

[9] Alexander M. Rush, Sumit Chopra, and Jason Weston. A Neural Attention Model for Abstractive Sentence Summarization. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 379–389, Lisbon, Portugal, 2015. Association for Computational Linguistics.

[10] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7871–7880, Online, 2020. Association for Computational Linguistics.

[11] Adina Williams, Nikita Nangia, and Samuel Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 1112–1122, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[12] 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. 3980–3990, Hong Kong, China, 2019. Association for Computational Linguistics.

[13] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6894–6910, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

[14] Yida Mu, Chun Dong, Kalina Bontcheva, and Xingyi Song. Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling, 2024. Version Number: 2.

[15] 銭本友樹, 長谷川遼, 宇津呂武仁. 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. 言語処理学会 第 30 回年次大会 発表論文集, pp. 2499–2504, 日本, 2024. 言語処理学会. This work is licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

A Appendix

表 2 マッピング結果：提案手法で生成されたタグと人手で作成されたタグの対応関係を示す．多くのタグが意味的に一致しており，提案手法の有効性が確認された．

提案手法で作成したタグ	人手で作成したタグ
情報入手困難さ	情報不足
マスク着用困難さ	リハビリやマスクの着用などの身体的影響
通院延期	通院や治療の延期・中止
送迎の必要性	通院が困難になった
入院時期の不透明さ	入院延期・目途が立たない
呼吸困難	リハビリやマスクの着用などの身体的影響
通院困難	通院が困難になった
体力低下	運動量の低下による身体の影響
付き添い負担	通院が困難になった
運動不足	運動量の低下による身体の影響
医療用品不足	医薬品（消毒薬，マスク含む）の不足
特になし	その他
外出時の感染不安	通院による感染への不安
外出制限	入院中の面会や行動の制限
診断基準の不明確さ	情報不足
医療機関での感染リスクへの不安	通院による感染への不安
リハビリテーション中止	リハビリやマスクの着用などの身体的影響
その他	その他
感染リスクへの不安	感染や重症化への不安
メンタルヘルスの悪化	メンタル面の問題やストレス

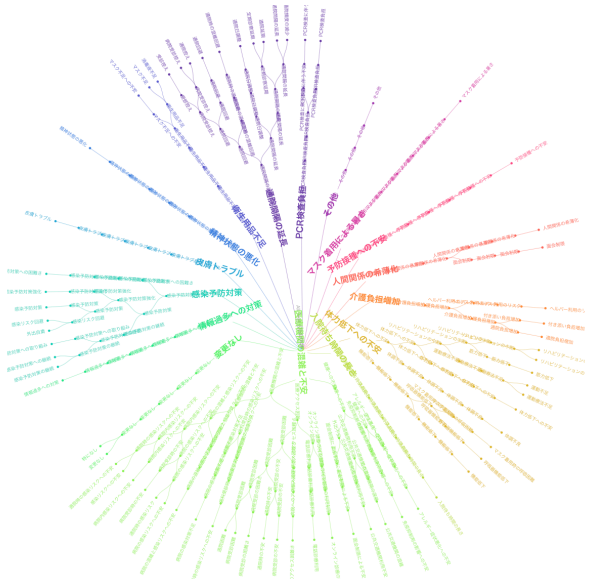

図 4 タグの統合過程：複数段階で行われたタグの統合プロセスの結果を示した放射状ツリー構造である．各階層は，生成されたタグが統合・プロセスを表しており，色分けされた分岐は意味的に類似したタグの統合結果を表現している．

タグ生成：TagGenに用いたプロンプト

以下の自由記述テキストに対して、困っていることに基づいて適切なタグを付与してください。
出力は以下の条件を満たしてください：

1. タグは既存のタグリストの中から最も適切なものを選択してください。
2. 該当する既存のタグがない場合、新しいタグを生成してください。ただし、タグは20文字以下にしてください。
3. タグには改行記号や括弧、解説文を含めないでください。
4. 出力には単一のタグのみを含め、複数のタグをつけないでください。

例

- **既存のタグを新しく生成したパターン**
  意見: 新たな治療薬がまだ開発されていないこと
  出力: 治療薬が未開発なこと
- **既存のタグを再利用したパターン**
  意見: もし感染した場合に有効な薬があるのかどうか不安
  出力: 治療薬が未開発なこと
- **新しいタグを生成したパターン**
  意見: 電話やLINEでコミュニケーションを継続した
  出力: SNSによるコミュニケーションの継続

タスク開始

既存のタグリスト: {existing_tags}
対象の意見: {opinion_escaped}
出力形式: タグ名のみを出力してください。

図 5 自由記述テキストに対して適切なタグを生成する際に用いたプロンプトの詳細を示している．タグの生成：TagGen では，LLM を活用して既存のタグリストに基づく分類および新規タグの生成を行い，データの冗長性を抑制しつつ，困りごとの概要を的確に捉えるタグを付与する．プロンプト設計では，タグの一貫性を確保するために文字数や形式に関する具体的な指示を与えている．