カスタマーサポート問い合わせに対する LLM 自動タグ付与の実現可能性の検討

岩田伸治 [1*] 佐藤志貴 [1*] 小比田涼介 [1] 石井陽子 [1] 岡本あずさ [1]
[1] サイバーエージェント
{iwata_shinji, sato_shiki}@cyberagent.co.jp

概要

カスタマーサポートでは，問い合わせに対して内容に応じたタグを付与し分類することで，タグごとに確立されたフローを用いた効率的な対応を実現するという運用がしばしば採用されるが，タグの付与は解釈と判断を伴う手作業であり負荷が高い．本稿では，こうした運用の効率化に向け，大規模言語モデルによる自動タグ付与の実現可能性を検討する．実験を通じて，タグ観点の多様性（たとえばトラブル内容を表すタグとその原因を表すタグの混在）がタグ付けを単純な多クラス分類タスクへ当てはめることを困難にすることがわかった．最後に，この性質をむしろ強みとして捉え，タグを観点別に整理し各観点からタグ付与を行うことで，問い合わせの意図を多角的に捉える枠組みを議論する．

1 はじめに

オペレーターの負担の軽減や迅速な返信による顧客体験の改善への期待から，カスタマーサポートにおける人工知能の活用は長年にわたり検討されてきた．近年では，顧客からの問い合わせに迅速に対応するために，人手による返信対応の代替として大規模言語モデル（LLM）を用いた自動での返信文の生成なども検討されている [1, 2]．

しかし，LLM の生成文には事実と異なる情報や不適切表現が混入するリスクが残るため，そのまま実運用に適用することは現状では難しい．これに対し，生成された文を人手で最終確認するなどの方法も考えられるが，オペレーション上の抜本的な変更が必要となり，導入の障壁となる．そのため，当初から完全自動生成を目指すのではなく，まず既存の対応フローを活かしつつ，LLM を用いた業務の効率化を検討することが現実的となる．

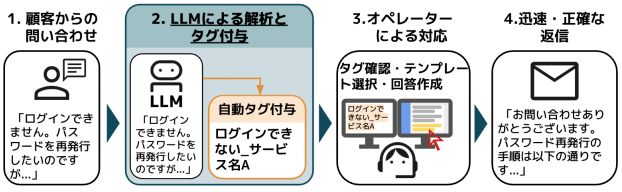

図 1 本研究において実現を目指すカスタマーサポートの運用．負荷が高いタグ付与を自動化することにより迅速な問い合わせへの対応を可能とすることを検討する．

カスタマーサポートにおける既存の対応フローの例として，問い合わせに対して内容に応じたタグ（たとえば『ログインができない』や『パスワードを変更したい』など）を付与し分類することで，タグごとに確立されたフローを用いた効率的な対応を実現するというものが挙げられる．一方で現場では，問い合わせ内容が何を指し示しているかを解釈して適切なタグを付与する工程に多くの時間を要する．そこで我々は，問い合わせ内容を示すタグを事前定義されたタグ集合から LLM で自動付与する支援ツールに着目する（図 1）．

本稿では，実際のカスタマーサポートの現場における過去の問い合わせと，それに対して人手で付与されたタグからなるデータセットを用いて，LLM による自動タグ付与の実現可能性を検討する．実験の結果，近年の高性能な LLM を用いたにもかかわらず人手で付与されたタグとの一致率は 52.9%にとどまり，今回の設定のままでは人手のタグ付与をそのまま自動化することは難しいことが示された．不一致の要因を分析したところ，複数観点の併存（たとえばトラブル内容を表すタグとその原因を表すタグの混在）を含むタグ体系に対して「1 件につき 1 タグ」という従来の多クラス分類の問題設定を適用すると，1 件の問い合わせに対して妥当なタグが複数成立し得るため，人手タグとの完全一致は難しくなるという構造的な要因があることがわかった．最後に，この性質をむしろ強みとして捉え，タグを観点別に整理し各観点に沿ってタグ付与を行うこと

＊ 本論文において両著者は同等に貢献した．

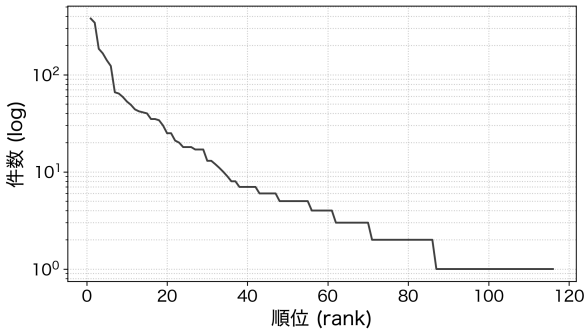

図 2　本データセットにおけるタグの出現頻度分布.

で，問い合わせの意図を多角的に捉える枠組みを議論する．

2　関連研究

カスタマーサポートの顧客体験向上に向けて，LLM が返信を自動生成する試みは多く行われてきた．例えば，RAG を用いた過去の問い合わせ内容の参照及び関連知識の活用 [2, 3, 4, 5, 6] や対話システムの設計を非専門家が行うアプリケーションの構築 [7]，知識グラフを用いたオペレータに近い柔軟な対応を行うシステムの構築 [8, 9, 10] 等がある．一方で，顧客体験を著しく低下させる意図しない自動返信が許容されないかつオペレーション上の大きな変更も難しい状況においては，これらの技術に加えて，人間が返信の主導権を保持したまま利用できる支援技術の実現が求められる．

3　データセット

本研究では，著者所属組織内で運用されるカスタマーサポート窓口において蓄積された過去の問い合わせログを用いる．データガバナンスの観点から，特定の部署名・サービス名・取扱領域（ドメイン）を推定可能な情報は本稿では開示せず，概要や特徴を抽象化して記述する．

概要．　データセットは，合計 4,855 件の問い合わせからなる．各事例は，顧客からの問い合わせメールの件名および本文（日本語テキスト）と，担当者が各メールにつき 1 件手作業で付与した問い合わせ内容タグからなる．各タグは，『ログインができない』や『パスワードを変更したい』のように，問い合わせ内容を端的に表す日本語の短いフレーズである．なお，問い合わせメールの件名および本文に含まれうる個人情報や機微情報は事前にマスキングを施した．

特徴 1：タグ種類数の多さ．　本データセットは，タグの種類数が 116 種類にのぼるロングテール性を有するデータセットで（図 2），分類問題としてはクラス数が非常に多い．一般にカスタマーサポートの運用において，タグ体系は運用開始時に固定的に設計できるものではなく，サービスを取り巻く状況の変化による新しい問い合わせの発生に伴いタグが追加されていくため，タグ種類数も大きくなる．

特徴 2：タグ観点の多様性．　本データセットのタグは，必ずしも画一的な観点のもと作成されているわけではない．ここで観点とは，各タグがどのような切り口で問い合わせ内容を捉えるかという設計意図を指す．たとえば，『ログインができない』はユーザが抱える問題（症状）という観点のタグであり，『パスワードを変更したい』はユーザが期待する対応（要望）という観点のタグである．先述のように問い合わせ内容の傾向は状況変化に応じて変動するため，必要なタグを動的に増やしていく必要がある．この過程のなかで最適なタグを追加する場合，必ずしも画一的な観点のもと新規タグを設計することが運用上最適となるわけではない．これにより，現場で用いられるタグ体系は複数の観点が併存するものとなりやすい．

4　実験

本稿の実験では，問い合わせメールの件名および本文を入力として，問い合わせ内容タグを LLM が推測可能であるかを確認する．

4.1　実験設定

4.1.1　自動推論の手法

本実験では，Honda らの Cheet Sheet ICL [11] を元に LLM へのインストラクションを自動生成したうえで，生成したインストラクション，出力候補となる 116 種類のタグの一覧，および推論対象となる問い合わせメールの件名・本文を LLM に与え，タグ一覧のうち適切なものを出力させた．

Cheat Sheet ICL は，多数のデモ例（many-shot ICL）を推論時にそのまま LLM への入力に用いる代わりに，デモ例からタスクの判断基準や解法を短いテキスト（cheat sheet）として要約し，推論時にはこの cheat sheet のみを文脈として与える二段階の枠組みである．本研究では，学習データ中の問い合わせとタグ付与の結果の組み合わせからタグ付与の判断基

準を cheat sheet として自動生成した．なお，推論時にはデモ例を含めなかった（0-shot）．

LLM のモデルパラメータの変動を伴う学習を実施せず，Cheat Sheet ICL を採用した主な理由として以下が挙げられる．

解釈性と柔軟性の担保． 生成される cheat sheet は，どのような観点・根拠（推論方針）でタグを選択すべきかを解釈可能な自然言語で明示するため，担当者が非エンジニアでもシステムの誤分類の原因分析が可能となる．さらに，仕様変更や運用ルールの更新が生じた場合でも，非エンジニアの担当者が cheat sheet を直接修正して判断基準の変更を反映できる．

特定のモデルへの依存の回避． 事前学習済みモデルへの追加学習は，特定のモデル仕様やバージョンに強く依存し，モデル更新のたびに再学習が必要になる可能性がある．一方で Cheat Sheet ICL は，モデルに与える基準を設計する手法であり，同様の指示追従能力を持つ他の LLM にも移植しやすい．

4.1.2 評価対象とする LLM

高い自然言語処理能力が報告されている Google Gemini API[1] の `gemini-3-pro-preview` を cheat sheet の作成およびタグ推論に用いる LLM として採用した．API 呼び出し時のパラメータはすべて初期値とした．

4.1.3 データ分割

データセット（全 4,855 件）は，学習データと評価データでメール件数が概ね 1:1 となるように分割した（学習 2,427 件，評価 2,428 件）．分割にあたっては，116 種類のタグそれぞれについて，学習データと評価データの双方で件数が可能な限り同数となるように層化分割を行った．

学習データは，cheat sheet 作成時に LLM に与えるデモ例として用いた．具体的には，学習データ中の問い合わせとそのタグを無作為に並び替えてデモ例として与えた．ただし，推論コストと入力トークン数を鑑みて，10 件以上のサンプルが存在するタグのデモ例については無作為抽出した 10 件のみに絞り（各タグ最大 10 件），最終的にデモ例として用いた問い合わせ件数は 404 となった．

評価データも同様に，推論コストを鑑みて，10 件以上のサンプルが存在するタグについては無作為抽出を施し各タグ最大 10 件とした．最終的に推論対象とした問い合わせ数は 463 件となった．評価データの件数が学習データの 404 件に比べて多い理由は，データセット全体で該当事例が 1 件しか存在しないタグについては評価データに優先的に含めるよう調整したためである．

4.2 結果

実験の結果，人手で付与されたタグと一致した事例の割合は 52.9%（245/463）にとどまった．誤ったタグ付与が不適切な担当者の割当や対応遅延につながりうることを踏まえると，今回用意した手法によって人手によるタグの付与をそのまま自動化することは困難であることがわかった．

5 分析

前節の実験において一致率が低かった要因を分析したうえで，タグ付与の自動化に向けた今後の方針などを議論する．ここでは，3 節で述べた本データセットの特徴であるタグ種類数の多さとタグ観点の多様性に着目した分析を行う．

5.1 タグ種類数の多さの影響

前節の実験では，候補タグが 116 種類と多い状況で，LLM にタグ一覧から 1 つを選択させる多クラス分類を行った．一般に，クラス数が増えるほど識別境界の設計が難しくなると考えられることから，今回の一致率の低さ（52.9%）は候補となるタグの種類数の多さに支配的な影響を受けた可能性がある．

この点を検証するために，現場からのヒアリングに基づき，根本的な内容が同一ではあるもののサービスごとに分割されているタグ（たとえば『ログインできない_サービス名 A』と『ログインできない_サービス名 B』）などについては，業務上の大きな影響なく統合できる可能性があると判断し，これらを同一タグへ集約した．その上で，学習データ・評価データは前節と同一の分割・同一事例を用いたまま，出力候補タグを 67 種類に集約した分類問題として同様の手法で再度実験を行った．結果として，一致率は 50.3%（233/463）となり，一致率の向上は見られなかった．このことは，少なくとも本研究で用いた Cheat Sheet ICL に基づく推論手法においては，クラス数の増加に対してある程度頑健であり，候補となるタグの種類数の多さ自体が正解率の低さを支配的に規定していない可能性を示唆する．

1 https://ai.google.dev/gemini-api/docs

表 1　タググループごとのタグ数・事例数・人手タグとの一致率.

タググループ	タグ数	事例数	一致率
発生事象	18	229	73.8%
原因	14	85	78.8%
要望	12	68	89.7%
案内方針	20	41	75.6%
その他	12	40	65.0%

5.2　タグ観点の多様性の影響

3 節で述べたように，本データセットでは，問い合わせの内容に対して付与されるタグが単一の観点に基づき体系化されているわけではない．たとえば，タグ候補に『ユーザが抱える問題』を表すものと『ユーザが期待する対応』を表すものが混在している場合，ひとつのメールに対して観点 A（問題）ではタグ a を付与することが適切であり，観点 B（期待する対応）ではタグ b を付与することが適切である，というように複数のタグが同程度に妥当となりうる．一方で，本実験ではどの観点でタグ付与を実施するかを指定しないまま各問い合わせに対して正解タグを 1 つに定める従来の多クラス分類の設定を採用した．このようなデータの性質と実験設定の間の齟齬が，結果における低い一致率の一因となっている可能性がある．

この仮説を確認するため，観点の異なりごとにタグをグルーピングし，タググループ内（すなわち観点が比較的揃った範囲）であれば同一手法で適切な分類が可能かを検証した．具体的には，まず前節の検証で用いた 67 種類のタグを，観点ごとのタググループに分割した．方法として，タグ候補一覧と各タグに該当する問い合わせ最大 10 件をプロンプトとして LLM に与え，観点ごとにタグを分類させたうえで，その結果を土台として著者が統合・分割の修正を行い，最終的に 5 つのタググループを作成した．タググループは，『発生事象（ユーザが抱える問題）』，『原因（ユーザが抱える問題の原因）』，『要望（ユーザが期待する対応）』，『案内方針（運営側が取るべき対応）』，『その他』の 5 種類になった．各グループに属するタグ数および評価データ中の事例数を表 1 に示す．

次に，それぞれのグループに属するタグが付与された問い合わせのみを学習データおよび評価データから抽出し，抽出後のデータに対して 4 節と同一の設定で多クラス分類を実施した．

各タググループにおける人手のタグとの一致率を表 1 に示す．グループに分割したことで分類問題における候補タグ数が小さくなったため 4 節や前節との単純な比較は難しいものの，全グループにおいて高い一致率となった．以上より，複数のタグ観点が含まれやすい実データに対してタグ付与を実施する観点を指定しないまま「1 件につき 1 タグ」という従来の多クラス分類の設定を当てはめたことで，多クラス分類問題として見ると人手と完全一致しないが生じた可能性が示唆される．

6　議論：多観点でのタグ付与の応用

前節では，多様な観点を含むタグ体系の下で各問い合わせにつき 1 タグという設定を採用した場合は人手タグを一意に再現することが難しくなる一方，少なくともタグ候補を同一観点に絞ったタグの付与では高い一致率が実現できることを確認した．

この結果は，問い合わせ対応の実運用における多角的なタグの自動付与の実現可能性を示唆する．この知見を活かした支援システムの例として，問い合わせに対して単一のタグを付与するのではなく，観点ごとに多角的なタグ付けを行い，各観点に紐づく返信テンプレート群から最適なものを選択するシステムが挙げられる．あるいは，各観点に紐づく返信テンプレート群のそれぞれから必要な部分だけを抜き出し組み合わせていくことで，より個別の問い合わせに合った返信テンプレートを作成するといった方向性も考えることができる．

7　おわりに

本稿では，カスタマーサポートにおける人手返信を前提とした支援技術として，実際の問い合わせログを対象に，LLM による自動タグ付与の実現可能性を検討した．Cheat Sheet ICL に基づき，学習データからタグ付与の判断基準を自然言語の cheat sheet として自動生成し，推論時にはタグ一覧と問い合わせ文を入力としてタグを選択させる枠組みを用いた．検証の結果，今回の設定のままでは人手のタグ付与をそのまま自動化することが難しいことが示された．一方で，タグ観点がデータセット内で多様である場合，1 件の問い合わせに対して妥当なタグが複数生じうることが本実験設定上での誤りとなってしまう点を分析し，観点ごとにタグを整理することで分類の見通しが改善する可能性を確認した．

謝辞

本研究の実施にあたってヒアリングに対応いただいた著者所属組織内のカスタマーサポート窓口の担当者の方々に感謝いたします.

参考文献

[1] Jingzhe Shi, Jialuo Li, Qinwei Ma, Zaiwen Yang, Huan Ma, and Lei Li. CHOPS: CHat with customer profile systems for customer service with LLMs. In **First Conference on Language Modeling**, 2024.

[2] Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. Retrieval-augmented generation with knowledge graphs for customer service question answering. In **Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval**, pp. 2905–2909, 2024.

[3] 小島淳嗣. Retrieval-augmented generation に基づくカスタマーサポートにおける返信メール自動生成の検討. 言語処理学会 第 30 回年次大会. 言語処理学会.

[4] 二宮大空, 戸田隆道. RAG における自己認識的不確実性の評価. 言語処理学会 第 30 回年次大会. 言語処理学会.

[5] Priyaranjan Pattnayak, Amit Agarwal, Hansa Meghwani, Hitesh Laxmichand Patel, and Srikant Panda. Hybrid AI for responsive multi-turn online conversations with novel dynamic routing and feedback adaptation. In **Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing**, pp. 215–229, 2025.

[6] Syed Shariyar Murtaza, Yifan Nie, Elias Avan, Utkarsh Soni, Wanyu Liao, Adam Carnegie, Cyril John Mathias, Junlin Jiang, and Eugene Wen. Implementing retrieval augmented generation technique on unstructured and structured data sources in a call center of a large financial institution. In Weizhu Chen, Yi Yang, Mohammad Kachuee, and Xue-Yong Fu, editors, **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)**, pp. 598–606, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[7] 二宮大空. カスタマーサポートにおける LLM を用いた RAG ベース対話システムの評価と事業活用に向けた取り組み. 人工知能学会研究会資料 言語・音声理解と対話処理研究会 99 回 (2023/12), pp. 191–192. 一般社団法人 人工知能学会, 2023.

[8] Izaskun Fernandez, Cristina Aceta, Cristina Fernandez, Maria Ines Torres, Aitor Etxalar, Ariane Mendez, Maia Agirre, Manuel Torralbo, Arantza Del Pozo, Joseba Agirre, Egoitz Artetxe, and Iker Altuna. Incremental learning for knowledge-grounded dialogue systems in industrial scenarios. In Tatsuya Kawahara, Vera Demberg, Stefan Ultes, Koji Inoue, Shikib Mehri, David Howcroft, and Kazunori Komatani, editors, **Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 92–102, Kyoto, Japan, September 2024. Association for Computational Linguistics.

[9] Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. Retrieval-augmented generation with knowledge graphs for customer service question answering. In **Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '24, p. 2905–2909, New York, NY, USA, 2024. Association for Computing Machinery.

[10] 山下雄大. 多様なユーザー問い合わせの意図理解を支援するナレッジグラフの構築. 人工知能学会全国大会論文集 第 39 回 (2025). 一般社団法人 人工知能学会, 2025.

[11] Ukyo Honda, Soichiro Murakami, and Peinan Zhang. Distilling many-shot in-context learning into a cheat sheet. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Findings of the Association for Computational Linguistics: EMNLP 2025**, pp. 17158–17178, Suzhou, China, November 2025. Association for Computational Linguistics.