LLM を用いたカスタマーハラスメント会話データ自動生成

An automatic method for generating customer harassment conversation data using large language models

木川 真実[1] 大室 昌也[1] 石橋 武史[1] 本間 塁斗[1] 繁永 直希[1] 浪岡 保男[1*]

Mami Kigawa[1] Masaya Omuro[1] Takeshi Ishibashi[1] Ruito Homma[1] Naoki Shigenaga[1] Yasuo Namioka[1*]

[1]東京都立産業技術大学院大学 Advanced Institute of Industrial Technology
*Corresponding author: Yasuo Namioka, namioka-yasuo@aiit.ac.jp

Abstract Customer harassment is a serious social issue requiring urgent public-private countermeasures. Since standardized criteria are difficult due to case-specific factors, automatic detection methods are needed. This study proposes using LLMs to generate synthetic conversation data. Training and evaluation using the generated data confirmed the effectiveness of the proposed approach.

Keywords large language model; conversational data; automatic generation; customer harassment; automatic detection

1 はじめに

カスタマーハラスメント[1]は近年深刻な社会問題とされ，官民連携による対策が急がれている．たとえば，「東京都カスタマー・ハラスメント防止条例」の制定[2]や，ANA グループと JAL グループ共同で行われた航空業界における対応方針の明文化[3]など，対応の制度化が進んでいる．

一方，現場では，過剰な要求や威圧的言動が従業員に深刻な影響を及ぼしている．池内[4]は，2014 年のコンビニエンスストアの恐喝事件などを例に，刑事事件に至らない「不当以上，違法未満」の悪質クレームが増加し，現場対応を困難にしていると指摘している．

カスタマーハラスメントの判断には，要求の妥当性や就業環境への影響など，個別事情を踏まえた評価が必要であり，汎用的・客観的な基準の策定は容易ではない．厚生労働省の「カスタマーハラスメント対策企業マニュアル」[1]においても，複数の観点からの総合的な評価の重要性を示している．

こうした背景から，小川等の研究[5]では，大規模言語モデル（Large Language Model: LLM）を用いてカスタマーハラスメントに関する会話データを自動生成し，更に，カスタマーハラスメント，顧客ストレス，就業者ストレスをスコア化するとともに，発生の有無を判定する手法が提案された．

本研究は，小川等が作成した自動生成データを用い，プロンプト設計の違いがアノテーション結果に与える影響を分析する．LLM による会話データの品質とその限界を明らかにし，LLM の社会実装に向けた設計指針と信頼性向上に資することを目的とする．

2 関連研究

LLM を活用した会話データの生成

カスタマーハラスメントに関する実態把握や判定基準の整備には，センシティブな会話データの収集と分析が不可欠である．しかしながら，現場の会話には個人情報や機密性の高い内容が含まれることが多く，大規模かつ多様な実データを収集・公開することは困難である．このような背景から，近年では LLM を活用して会話データを人工的に生成するアプローチが注目されている．

Bonaldi 等[6]は，ヘイトスピーチへの反論をテーマとした会話データセット「DIALOCONAN」を構築するため，GPT-3 による会話生成と人間による編集・アノテーションを組み合わせたハイブリッドな手法を提案した．センシティブな内容を含む会話の構築において，LLM の生成能力と人間の判断を組み合わせることで，多様性と品質を両立できることを示している．

また，Pujari および Goldwasser[7]は，文化的文脈に配慮した会話生成を目的として，LLM による生成出力に対し人間が文化規範に関する情報を付与するパイプラインを構築した．プロンプト設計や人手による補足が生成結果に与える影響を分析することで，LLM を活用した会話生成の妥当性と応用可能性を検証している．

これらの先行研究はいずれも，現実の会話が持つ倫理的・文化的側面や文脈の複雑さを踏まえ，LLM と人間の協働によって高品質なデータセットを構築する手法を提示している．

本研究は，カスタマーハラスメントという社会的に敏感な領域において，LLM によって生成された会話データの評価に着目し，プロンプト設計の違いがアノテーション結果に与える影響を分析するものである．これは，先行研究の成果を応用しつつ，生成データの整合性や信頼性を検証する点で新たな知見を提供するものである．

LLM を活用したデータ分析基盤

中塚等[8]はデータ分析基盤として，AIIT（Advanced Interactive Insight Tracer）を提案し，IndexedDB と LLM を組み合わせてブラウザ上で対話的にデータを処理・可視化する分析基盤を開発している．この基盤は，バックエンドの制約を受けずにフロントエンド環境のみでベクトル化・クラスタリング・プロンプト調整・可視化まで一連の分析を行える点が特徴である．さらに，本システムではプロンプトを System Prompt と User Prompt に分けて記述し，前者では処理条件・出力プロパティ名・JSON 形式指定などを記載し，後者では対象データのカラム指定や入力構成を指示する設計となっている．

本研究では，この基盤を活用し，会話データの自動生成及び分析処理を実現している．

3 提案手法

概要

本研究では，カスタマーハラスメントという社会的にセンシティブな問題において，プライバシー問題や収集コストなどの制約により実データの収集が困難な現状に着目し，LLM を用

Received: 2025-10-31; Published online: 2026-02-28

いた会話データの自動生成手法を提案する．本手法は，図 1 に示すように，プロンプト設計，LLM による会話データの自動生成，人間による多面的評価（アノテーション），そして生成データの品質検証までの一連のフローで構成される．

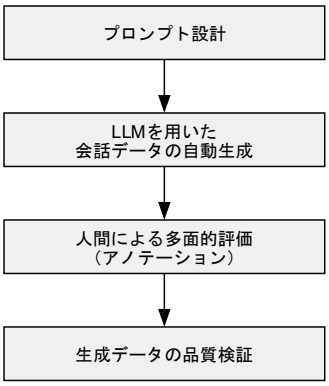

図 1　提案手法の流れ

生成手法

カスタマーハラスメントに関する実際の会話データの公開には，プライバシー侵害のリスクや社会的配慮，個人情報保護法や企業機密の観点から，法的・倫理的な制約が伴う．このため，小川等の研究[5]では，実データの収集は困難と判断し，LLM による人工的な会話データの生成を採用した．

同研究では，4 種類のプロンプトを設計した．各プロンプトは異なる設計意図に基づき，入力パラメータなどの生成条件を変えることで，通常の接客応対からカスタマーハラスメントに該当する可能性のある強い表現まで，幅広い会話を網羅できるよう工夫されている．会話データの生成には，中塚等のデータ分析基盤（AIIT）[8]から API を通じて OpenAI,Inc.が提供する GPT-4o（gpt-4o-2024-11-20）を使用した．

各プロンプトに対応する詳細表の一覧を表 1 に示し，プロンプトの構成に関する具体的な内容は付録表 A～D に記載している．prompt1～4 に共通して，顧客と就業者が日本語で交互に 5 回ずつ発話し，合計 10 の発話から構成される会話が生成される．出力形式は JSON である．

prompt1 では，situation（状況），customer（顧客の年代・性別），worker（就業者の年代・性別），score（カスタマーハラスメントの度合い）の 4 項目を入力パラメータとして指定する．score の値に応じて，0.0 では通常の会話，1.0 では悪質なカスタマーハラスメントとなるよう，生成される会話の内容が調整される（付録表 A 参照）．

prompt1 の出力結果では，同一の score に対して類似した会話が生成されやすく，特定の会話表現に収束する傾向が見られた．

表 A　各プロンプトと主な生成条件ならびにプロンプトの詳細との対応

プロンプト	主な生成条件	実際のプロンプト
prompt1	situation： 状況 customer： 顧客の年代・性別 worker： 就業者の年代・性別 score：0.0～1.0 で指定し，0.0 は通常の会話，値が大きいほど悪質なカスタマーハラスメントとなるよう出力を制御	付録：表 A
prompt2	situation： 状況 customer_sex：顧客の性別 customer_age：顧客の年齢層 request：顧客の要求内容 reason：要求の理由 tone：顧客の話し方 worker_response：就業者の対応 score：0.0～1.0 で指定し，0.0 は通常の会話，値が大きいほど悪質なセクシャルハラスメントとなるよう出力を制御	付録：表 B
prompt3	状況 会話のきっかけ セクハラ度：値が大きいほどセクシャルハラスメントの度合いが高まる	付録：表 C
prompt4	シチュエーション： 状況 要求内容：顧客の要求内容 話し方：顧客の話し方	付録：表 D

そこで prompt2 では，会話パターンの多様性を促すため，入力パラメータを拡張・再構成し，customer_sex（顧客の性別），customer_age（顧客の年齢層），request（顧客の要求内容），reason（要求の理由），tone（顧客の話し方），worker_response（就業者の対応），score（カスタマーハラスメントの度合い）を指定する形式とした．score は prompt1 と同様に値に応じて会話内容が調整され，さらに 0.4 以上の場合は request を，0.7 以上では reason および tone を考慮するよう指示している（付録表 B 参照）．

prompt1 および prompt2 では，会話の多様性を一定程度確保できたものの，「就業者・顧客の両者にストレスがある」または「両者にストレスがない」と評価されるケースが大半を占め，会話パターンに偏りが見られた．

prompt3 では，前述の偏りを制御し，さらなるバリエーションの拡張を図ることを目的として，セクシャルハラスメントに関する会話の生成を試みた．入力パラメータは，状況，会話のきっかけ，セクハラ度に変更している．

セクハラ度は，従来の score の使い方を拡張したものであり，score が 0.0 の場合は通常の会話，1.0 の場合は悪質なセクシャルハラスメントとなるように調整している．

また，データの偏りへの対策として，就業者の対応スタイルを「気弱な対応」「通常の対応」「毅然とした対応」の 3 種類に設定し，ランダムに選択されるようにした．これにより，同じセクハラ度でも異なる応答パターンが生成されるよう工夫されている（付録表 C 参照）．

prompt4 は，試行の一環として score による制御を行わず，シチュエーション，顧客の要求内容，顧客の話し方の 3 つの入力パラメータを組み合わせ，顧客の発話傾向にバリエーションをもたせることを試みた．これは，より多様なカスタマーハラスメント表現の生成可能性を探ることを目的としている（付録表 D 参照）．

これら prompt1～4 の設計に基づき，合計 4,034 件の会話データを生成した．内訳は，prompt1 が 306 件，prompt2 が 3,000 件，prompt3 が 386 件，prompt4 が 342 件である．

4 実験・評価

評価手法

本評価の目的は，3 章で述べた生成手法で得られた会話データが，人間の直感的な「カスタマーハラスメントらしさ」の印象とどの程度一致しているかを検証することである．特に，プロンプト設計やカスタマーハラスメントの度合い（score）などの生成条件がアノテーション結果に与える影響を分析し，生成データの妥当性と傾向を明らかにする．

評価は，学内でアノテーション協力者を募り，以下の 4 項目について実施した．評価形式は複数選択可とした．

- カスタマーハラスメントの有無
- 顧客のストレス有無
- 就業者のストレス有無
- 会話の意味が理解できない（該当する場合のみ）

評価対象の母集団は，合計 4,034 件の会話データのうち，小川等の研究[5]で 1 名のアノテーターによりアノテーションされた 1,532 件で構成される．内訳は，prompt1：306 件，prompt2：500 件，prompt3：384 件，prompt4：342 件であり，回答ラベルのバランスを考慮して選定されており，会話内容による除外は行っていない．表 B には，2024 年度に各プロンプトから生成されたデータのうち，回答ラベルに基づいて選定された件数と，2025 年度に抽出された件数を示す．

表 B　評価対象として抽出された 50 件の内訳

prompt	生成数（件）	抽出数（件）
prompt1	306	12
prompt2	500	10
prompt3	384	12
prompt4	342	16
合計/Total	1,532	50

本研究では，アノテーターの負荷を考慮し，この 1,532 件から評価対象を 50 件に絞った．各プロンプトのパラメータをもとに類似要素でグルーピングし，それらの組み合せに基づいて階層的に抽出を行った．全パラメータを階層に含めると過度な細分化を招くため，プロンプトごとの構成に応じて階層設計を調整した．網羅性には限界があるが，分布の偏りを抑えるよう配慮している．アノテーション作業には，OSS ツールである doccano[9]を使用した．

実験実施（アノテーション）

本研究では，筆者が所属する大学院の学生 11 名に依頼して実施した．加えて，小川等の研究[5]において研究室の学生 1 名が行ったアノテーション結果も評価対象に含めた．

アノテーターには，プロンプト設計の意図や生成時のパラメータは開示せず，生成会話のみを提示した．評価に先立ち，評価基準と手順を記載したインストラクションを配布し，説明後にアノテーション作業を開始した．作業は 2025 年 6 月中旬に約 1 週間かけて実施し，同一の 50 件の会話データを割り当てた．作業は中断・再開が可能な形式とし，各自の任意のタイミングで対応できるよう配慮した．

実験結果の分析

アノテーション結果に基づき，各 prompt で生成された会話データの評価傾向を分析した．

prompt 別に見ると，prompt4 におけるカスタマーハラスメント判定率は 90.1％と突出して高く，prompt1～prompt3（34.7～48.3％）との差は有意であった（図 2 参照）．

以下の会話例は，prompt4 によって生成された会話の一部抜粋である．本会話は，シチュエーションを「コンビニ」，要求内容を「不手際に対する謝罪」，話し方を「人格を否定する」として設定したプロンプトに基づき生成されたもので，12 名全員がカスタマーハラスメントと判断した．

顧客：いや，もういい加減にしろよ．お前のせいで時間が無

駄になった！
就業者：申し訳ありません．お待たせしてしまって…
顧客：謝って済むと思ってるの？お前，ほんと無能だな．
就業者：そんなことはないと思いますが…本当に申し訳ありません．
顧客：いや，もういい加減にしろよ．お前のせいで時間が無駄になった！

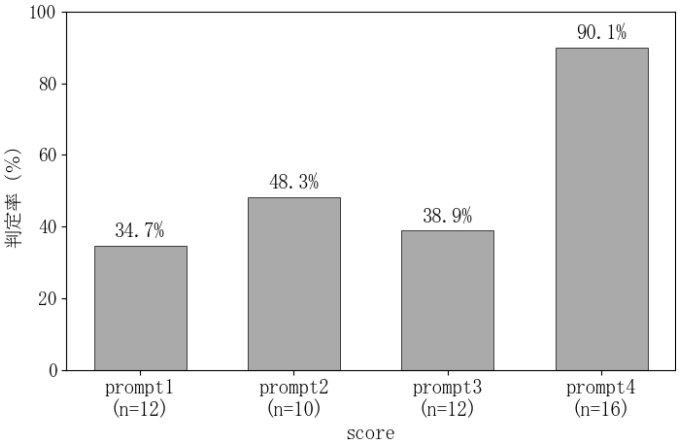

図 2　prompt 別：カスタマーハラスメント判断割合

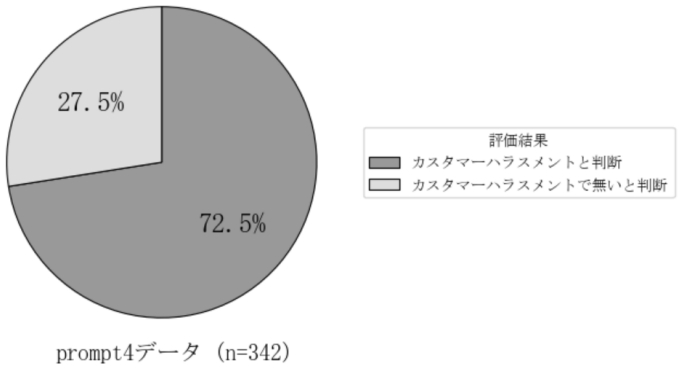

図 3　prompt4 におけるカスタマーハラスメント判定割合（2024 年度・1 名評価）

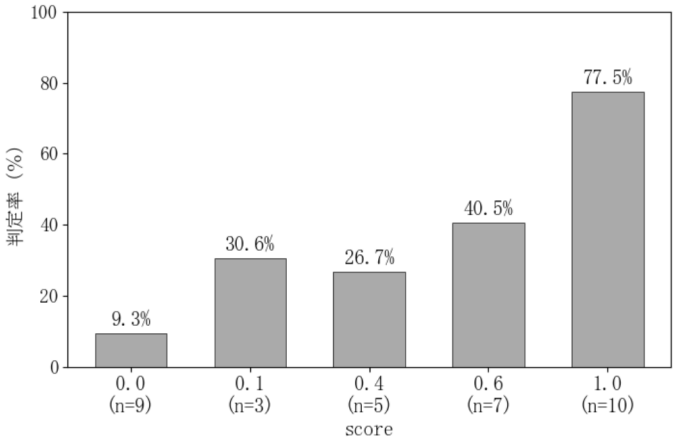

図 4　score 別：カスタマーハラスメント判定割合

また，小川等の研究[5]では，prompt4（全 342 件）で生成された会話データの評価が行われており，その結果でも 72.5%がカスタマーハラスメントと判断された（図 3 参照）．本研究における 12 名の評価（16 件）も，この傾向と概ね一致していた．

score 別に見ると，score が高いほど会話データがカスタマーハラスメントと判断される傾向があり，score=1.0 では判定率が 77.5%と最も高かった．一方，score＝0.4（26.7%）および 0.6（40.5%）といった中間帯では，判定にばらつきが見られた．これは，該当する会話がカスタマーハラスメントに該当するか否かの判断が分かれる，いわゆるグレーゾーンに位置していた可能性が高く，アノテーター間での解釈の揺れが生じやすい領域であったことを示唆する．また，score=0.0 では判定率が 9.3%にとどまり，出力制御の意図と概ね一致していた．このことから，score 設計はアノテーターの判断と一定の整合性を持つ可能性が示唆される．なお，本分析では，score 項目が設計意図に基づき付与された 34 件のみを対象としており，score 制御を行っていない prompt4 の会話データは分析対象から除外している（図 4 参照）

また，score=0.1 に分類された会話データは 3 件と件数が少ないため一般化は避けるべきだが，カスタマーハラスメント判定率は 30.6%と相対的に高かった．

5　考察

本研究を通じて，LLM によるカスタマーハラスメント会話データの自動生成は，適切なプロンプト設計によって一定の制御が可能であることが示された．特に，語調や要求内容，文脈的な要素が出力に反映され，評価結果に大きく影響する傾向が見られた．これにより，実用的な人工会話データを一度に大量に生成できる手法としての有用性が示された．

本研究では，score・tone・reqest などのパラメータを網羅的に組み合わせ，生成結果を実験的に検証している．

4 章の実験結果の分析にて，score=0.1 に分類された会話のカスタマーハラスメント判定率は 30.6%と相対的に高かった．対象はわずか 3 件であり，一般化は困難だが，うち 1 件では「顧客の不満のはけ口」や「威圧的な話し方」といったパラメータが影響し，多くのアノテーターがカスタマーハラスメントと判断した．このように，特定条件が判定率を押し上げることからも，文脈や語調の微細な設計がプロンプト制御において重要であることが示唆される．あわせて，LLM の確率的な生成特性により，意図通りの出力を得ることには限界がある可能性も示している[10].

また，prompt4 では，「顧客の話し方」や「顧客の要求内容」の指示が明確に出力へ反映され，攻撃的・否定的な表現が会話によく現れていた．このため，アノテーターの判断が揃いやすく，判定率も他条件より高かったと考えられる．

これらの結果は，score と話し方等の要素の組み合わせにより出力の印象が大きく左右されることを示しており，プロンプト設計においてはこうしたバランスの考慮が重要である．

この傾向は，LLM の出力がプロンプト設計に大きく依存すること，また，LLM の限界として「文脈理解や意図把握の不完全さ」や「意図しない出力の可能性」を明示しており，こうした特性を踏まえたモデル設計および運用が求められるとする OpenAI 社の記述[11]と整合する．

6 まとめ

本稿では，LLM を用いたカスタマーハラスメント会話データの自動生成手法を提案した．カスタマーハラスメントはプライバシーや機密保持の観点から実際の会話データの外部開示が困難であり，学習用データの不足が課題となっている．本研究では，この課題への対応策として LLM による疑似データ生成を行い，設計条件に基づいた生成結果の傾向と実用性について評価・分析を行った．

LLM を活用することで，設計意図に基づいた多様な会話データを効率的に生成することができる．これにより，事案ごとの事情を踏まえた汎用的な判断基準の策定に資する疑似データの確保が可能となり，実用的価値がある．

このようにして得られた多様な会話データを用いて，本研究では出力内容の傾向を評価・分析した結果，生成内容は概ね設計意図と整合しており，語調や要求に関する指示が出力に反映されていることが確認された．また，文脈的ニュアンスが受け手の印象に影響する可能性も示唆された．

実社会での応用には再現性や説明可能性が求められるが，LLM のブラックボックス性ゆえに，生成データのレビューとプロンプトの調整を専門家や実務担当者が担うことが不可欠である．このような課題に対しては，プロンプト設計の工夫と人によるレビュー体制を併用することで，実務応用における信頼性の担保に有効である．

今後は，会話データの多様化とドメイン知識を有するアノテーターによる検証を通じて，プロンプト設計指針のさらなる精緻化が期待される．あわせて，一部の会話に対しては顧客ストレスおよび就業者ストレスに関する印象評価も収集しており，カスタマーハラスメント判定との関連性についての分析も今後の検討課題となる．

謝辞

東京都デジタルサービス局の皆様には TDPF[12]の活動やカスタマーハラスメント・データについて，東京都産業労働局の皆様には東京都カスタマー・ハラスメント防止条例およびその背景について，全日本空輸株式会社 CX 推進室 CS 推進部の皆様には企業におけるカスハラ対策について，それぞれご教示いただいた．ここに厚く御礼申し上げる．

本研究は小川等の先行研究[5]を基に，評価作業を補完・発展させたものである．研究の引継ぎ，知見や記録の活用に際し，協力いただいた著者の皆様に感謝申し上げる．

参考文献

1. 厚生労働省：「カスタマーハラスメント対策企業マニュアル」，厚生労働省，(2022-2)
2. 東京都：「東京都カスタマー・ハラスメント防止条例」，東京都，(2024-10)
3. 全日本空輸株式会社・日本航空株式会社：「ANA・JAL グループ共同で『カスタマーハラスメントに対する方針』を策定」，プレスリリース，(2024-6)
4. 池内裕美：「なぜ『カスタマーハラスメント』は起きるのか－心理的・社会的諸要因と具体的な対処法－」，情報の科学と技術，70, 10, pp.486–492（2020-10）
5. 小川和浩・池田貴康・荒川健・中塚晶仁・富樫健太・河野吉宏・浪岡 保男：「LLM を用いたカスタマーハラスメントならびにストレスの自動スコアリング手法」, DEIM2025, 1B-01, (2025-3)
6. Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroglu, and Marco Guerini: "Human–Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering", Proc. 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022), pp. 8031-8049, Abu Dhabi, United Arab Emirates (2022-12)
7. Rajkumar Pujari and Dan Goldwasser: "LLM–Human Pipeline for Cultural Grounding of Conversations", Proc. 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2025), pp. 1029-1048, Albuquerque, NM, USA (2025-6)
8. 中塚晶仁・小川和浩・池田貴康・富樫健太・河野吉宏・荒川健・浪岡保男：「LLM と IndexedDB を活用した対話的ベクトルデータ分析基盤の検討 COVID-19 不満データを用いた社会動態の時系列分析への適応」, DEIM2025, 7F-03, (2025-3)
9. H. Nakayama: “doccano: Text annotation tool for machine learning practitioners”, GitHub, https://github.com/doccano/doccano (accessed 2025-07)
10. OpenAI:“GPT-4 Technical Report”, https://cdn.openai.com/papers/gpt- 4.pdf (2023-3)
11. OpenAI: "Best Practices for Deploying Language Models", https://openai.com/index/best-practices-for-deploying-language-models/ (2022-6)
12. 東京都, “東京データプラットホーム”, https://www.tdpf-hp.metro.tokyo.lg.jp/, 2024.

付録　使用したプロンプトの詳細

表 A prompt1 の構成

System prompt	入力された値に基づいて顧客と就業者の会話を作って JSON で返して． ルール ・会話はすべて日本語で行うこと． ・顧客と就業者が交互に 5 回発話して計 10 回の会話にすること． ・score（range0.0-1.0）の値に基づいてカスタマーハラスメントの度合いを調整する こと．1.0 のときは悪質なカスタマーハラスメント，0.0 のときは通常の会話. JSON Format output{ situation:入力した situation の値,customer:入力した customer の値,worker:入力した worker の値, score （range0.0-1.0）:入力した score（range0.0-1.0）の値, dialogue:[{"speaker":,"text":},{"speaker":,"text":},{"speaker":,"text":},{"speaker":,"text":},{"speaker":,"text":}]}
User prompt	situation: {situation } customer: {Customer} worker: {worker} score（range0.0-1.0）:{ score（range0.0-1.0）}
Input parameters	Situation：コンビニのレジ，介護施設，スーパーマーケットなど（全 43 パターン） customer：20 代〜60 代の男女，および 70 代・80 代の男性（全 12 パターン） worker：10 代〜50 代の男女，60 代・70 代の男性（全 10 パターン） score：0, 0.2, 0.4, 0.6, 0.8, 1.0（全 6 段階）

表 B prompt2 の構成

System prompt	入力された値に基づいて顧客と就業者が苦情・クレームのやりとりをしている会話を作って． ルール ・10 回発話させること． ・会話はすべて日本語で行うこと． ・score はカスタマーハラスメントの度合いを表す．0.0 は通常の会話，1.0 は悪質なカスタマーハラスメント にすること． ・score が 0.4 以上のとき，request の内容を考慮すること．0.7 以上のときは， さらに reason，tone の内容も考慮すること． ・結果は JSON で返すこと． JSON Formatdialogue: [{"speaker":,"text":},]
User prompt	situation: {situation } customer_sex: {customer_sex} customer_age: {customer_age} request: {request} reason: {reason} tone: {tone} worker_response: {worker_response} score: {score}
Input parameters	situation：スーパーマーケット，タクシー，コールセンター（全 3 パターン） customer_sex：男，女（全 2 パターン） customer_age：10 代から 70 代以上までの各年代（全 7 パターン） request：商品・サービスに見合った現金の要求，上司・上長による謝罪の要求など（全 8 パターン） reason：接客やサービス提供のミス，顧客の不満のはけ口・嫌がらせなど（全 7 パターン） tone：一方的に話す，威圧的に話す，攻撃的な話し方，理詰めなど（全 7 パターン） worker_response：毅然と対応する，謝りつづける，何もできないなど（全 6 パターン） score：0.1, 0.4, 0.7, 1.0（全 4 段階）

表 C prompt3 の構成

System prompt	入力したパラメータに基づいて顧客（男性）と就業者（女性）とのやりとり会話を作って． 就業者の対応は気弱な対応/通常の対応/毅然とした対応からランダムに選択して． ルール ・10 回発話させること． ・会話はすべて日本語で行うこと． ・セクハラ度は就業者へのセクシャルハラスメントの度合いを表す．0.0 ならセクハラの無いやりとり，1.0 は悪質なセクハラのやりとりとし，それ以外の値は値に応じてセクハラ度を調整すること． ・結果は JSON Format で指示した形式で返すこと． ・"speaker"は"顧客"か"就業者"で出力すること JSON Format {dialogue: [{"speaker":,"text":}]}
User prompt	状況：{状況} 会話のきっかけ：{会話のきっかけ} セクハラ度：{スコア}
Input parameters	状況：タクシー，コールセンター，保険のセールス，市役所，薬局など（全 16 パターン） 会話のきっかけ：通常のやりとり，商品やサービスに対する苦情・クレームなど（全 4 パターン） セクハラ度：0, 0.2, 0.4, 0.6, 0.8, 1.0（全 6 段階）

表 D prompt4 の構成

System prompt	入力された値に基づいて顧客と就業者が苦情・クレームのやりとりをしている会話を作って． ルール ・10 回発話させること． ・会話はすべて日本語で行うこと． ・結果は JSON で返すこと． JSON Format dialogue: [{"speaker":,"text":},]
User prompt	シチュエーション: {シチュエーション} 要求内容: {要求内容} 話し方: {話し方}
Input parameters	シチュエーション：コンビニ，スーパーマーケット，市役所，銀行の窓口など（全 35 パターン） 要求内容：不手際などに関する謝罪，上司・上長からの謝罪（全 5 パターン） 話し方：攻撃的な話し方，大声をあげる，人格を否定する