修士論文

レビュアの評価傾向を用いた個人差のある リッカート尺度間のスケール統合手法

川上 大凱

2026 年 1 月 30 日

岐阜大学大学院 自然科学技術研究科 知能理工学専攻 知能情報学領域
鈴木研究室

本論文は岐阜大学大学院自然科学技術研究科
修士（工学）授与の要件として提出した修士論文である.

川上 大凱

指導教員：
鈴木 優 教授

レビュアの評価傾向を用いた個人差のあるリッカート尺度間のスケール統合手法*

川上 大凱

内容梗概

本研究の目的は，リッカート尺度に基づく主観評価に内在する個人差の解消である．本研究の目的を達成するため，我々はリッカート尺度に基づく主観評価を多数の評価者による平均的な判断基準に対応するスケールへと正規化する手法を提案する．主観評価には，同一の評価値であっても評価者ごとに判断基準が異なるという問題がある．特に，実際に付与された満足度と評価理由を示すレビューから推測される満足度が大きく異なる例外的なレビューケースにおいてはその傾向が顕著である．このことから我々は，例外的なレビューケースを主観評価に個人差が内在する要因の一つであると考えた．よって本稿において我々は，既存研究ではノイズとして扱われてしまう例外的なレビューケースの存在を構造的な情報として保持することが可能な混同行列を用いてレビュアごとの評価傾向を表現する．実験では，例外的なレビューケースとレビュアの評価傾向における基礎的検証や混同行列による評価傾向表現の妥当性検証，そして本提案手法が個人差解消の正規化タスクにおいて有効かどうかを検証した．実験結果より，意味解釈の個人差を構造として扱うことによって，既存研究ではノイズとして扱われていた個人差が体系的に補正可能であることが示した．一方で，混同行列における個人差を表す構造によって補正効果が変わることを確認した．

キーワード

レビュー，評価傾向，SSIM，LLM，Few-Shot 学習，リッカート尺度，正規化

*岐阜大学大学院 自然科学技術研究科 知能理工学専攻 知能情報学領域 修士論文, 学籍番号: 1241751485, 2026 年 1 月 30 日.

目次

図目次

表目次

第 1 章　はじめに

本研究で我々は，リッカート尺度に基づく主観評価を多数の評価者による平均的な判断基準に対応するスケールへと正規化する手法を提案する．本提案手法により，リッカート尺度に基づく主観評価における意味解釈の個人差を解消することが可能となる．これにより，本提案手法は評価値の直接的な比較や集計から導かれる分析結果の安定した解釈を可能とする．

リッカート尺度は，EC サイトに投稿されるサービス評価やアンケート調査などで用いられる満足不満足といった主観評価を表す順序尺度の一つである．リッカート尺度は，企業のマーケティング分析や商品改善のための重要なデータ資源としても活用されている．またマーケティングだけでなく，心理学のような主観評価を定量的に利用する研究分野においてもリッカート尺度は多く活用されている．しかし，リッカート尺度に基づく主観評価には大きな問題がある．その一つが，主観評価に内在する個人差である．主観評価は，評価者の評価基準や考え方によって個人差が生まれる．例えば評価が全体的に厳しい評価者と寛容な評価者では，同一の評価値が必ずしも同一の意味を内包しない．そのためリッカート尺度に基づく主観評価における個人差は，評価値の直接的な比較や集計から導かれる分析結果の解釈を不安定なものにしてしまう．したがって，本研究は主観評価における意味解釈の個人差を解消することを目的とする．

我々はこの問題を解消するために，リッカート尺度に基づく主観評価を意味解釈の個人差が内在する主観尺度から意味解釈の個人差が存在しない参照尺度へと変換する必要があると考えた．我々は，評価値とは独立に存在するものではなく評価者が持つ評価理由に基づいて付与された結果であるという立場に立つ．この立場に基づく場合，主観評価に内在する個人差は評価理由と評価値における対応関係の違いとして捉えることが可能となる．我々は評価理由と評価値における対応関係の違いを基に評価値を再解釈することによって，評価者ごとに異なる意味解釈を持つ主観的な尺度を意味解釈の個人差が存在しない参照尺度へと変換することが可能だと考えた．

本研究において我々は，リッカート尺度に基づく主観評価に内在する個人差を解消するために，評価者の評価傾向を用いる．本稿において我々は，評価者の評価傾向を評価理由と評価値における対応関係と定義する．我々は評価者の評価傾向を定量的に表現するために，評価者が付与した評価値と評価理由を表すレビューから推測される評価値の共起頻度を利用する．既存の Z-score 正規化は，平均や分散といった統計量のみに基づいて個人差を補正する．そのため，評価者の評価傾向や考え方に関する特徴的な情報は失われてしまう．一方で既存の IRT 系モデルによる正規化は，評価者の評価傾向を能力や厳格さといったスカラ値を用いて統計確率的に推定する．そのため，多数の評価者の平均的な評価基準に基づいた個

人差の補正が可能となる．しかし，IRT 系モデルは評価者の評価傾向を単一のスカラ値に落とし込んでいるため，例外的なレビューケースがノイズとして平均化されてしまう．例外的なレビューケースとは，レビュー内では不満点や改善点といったネガティブな言及がされている一方でポジティブな評価値が付与されるといったレビューと評価が大きく異なるケースのことである．我々は，例外的なレビューケースを主観評価に個人差が内在する要因の一つであると考えた．したがって，例外的なレビューケースがノイズとして平均化されてしまう場合，評価者の正確な評価基準をもとに個人差を補正しているとは言い難い．一方で，評価者が付与した評価値と評価理由を表すレビューから推測される評価値の共起頻度は例外的なレビューケースを共起パターンとして保持することが可能である．

本稿において我々は，本手法の検証対象として EC サイトのレビューデータを扱う．そのため我々はレビュアの評価傾向を，共起頻度が集約されたデータであるレビュア単位の満足度予測における混同行列を用いて表現する．混同行列は，そのレビュアの感じた実際の満足度がどの程度レビューに内包されているのかを要素の大きさや位置で表現可能である．そのため，我々は混同行列による評価傾向の定義が妥当であるという仮説を立てた．この仮説を検証するため，我々は混同行列に基づいて選択されたレビュアが持つデータ群を Few-Shot Prompting に用いる参考データとした満足度予測を実施する．また本提案手法の有効性を検証するため，混同行列とレビュアが付与した満足度を入力とし，意味解釈の個人差が存在しない参照尺度へと正規化した満足度を出力とするニューラルネットワークを構築する．本稿において我々は，意味解釈の個人差が存在しない参照尺度，つまり多数のレビュアに共通する平均的な評価傾向を近似的に表現する参照尺度を LLM による Zero-Shot 推論の予測値で定義した．

混同行列による評価傾向定義の妥当性検証に関する実験結果より，LLM に対して与える参考データとして推論対象レビュアと評価傾向が類似しているレビュアを用いるほど満足度予測精度が向上することを確認した．この結果は，混同行列による評価傾向の定義が妥当であるという仮説が正しいことを根拠づけるものである．また，本提案手法の有効性検証に関する実験結果より，提案手法が主観評価に内在する個人差の解消に有効であることを確認した．

本研究の貢献は以下の通りである．

- 主観評価に内在する個人差の補正を予測問題ではなく正規化問題として捉えた．
- 主観評価に内在する個人差を意味解釈の個人差が内在しない参照尺度へと正規化可能であることを示した．
- 主観評価に内在する個人差の要因として，例外的なレビューケースに着目した．

- 例外的なレビューケースをノイズではなく構造的な特徴として捉えた.
- 例外的なレビューケースは想像上の概念ではなく実データ上に存在する現象であることを確認した.
- 例外的なレビューケースはレビュー投稿数に関係していることを確認した.
- 混同行列による評価傾向の表現が妥当であることを確認した.
- 提案手法が主観評価に内在する個人差の解消に有効であることを確認した.

第 2 章　基本的事項

2.1　レビュー

レビューとはサービス使用後にユーザが記述する感想・批評全般のことを指す．ユーザは EC サイトや地図アプリ，SNS などの電子媒体上での投稿や手紙やアンケートのような紙媒体上での投稿が可能である．レビュー対象は商品や施設など多岐にわたる．ユーザはレビューを読むことにより他人の考えや感じたものを知ることが可能となる．そのため，企業戦略を考える際の問題点分析やユーザの購買意思決定に役立つ客観的な指標であり情報源となる．

レビュー構成は主に評価・感想の 2 種が主要素である．主要素はレビューが投稿される媒体やサイトによって，項目別評価や長所短所のような形に細分化される．評価は星数値や満足/不満足で表現されることが多い．そのため，評価値はデータ分析時に順序尺度として扱われる．

2.2　リッカート尺度

リッカート尺度とは，主に EC サイトやアンケート調査での満足/不満足といった主観評価を段階的に表すための指標である．またリッカート尺度は，順序尺度の一つである．そのため各評価値間の意味的距離は等しくないことに注意する必要がある．例えば，リッカート尺度の評価値に対する分析では，意味的距離関係が定義できないにもかかわらず評価値を平均化してしまうことがある．しかし，各評価値が等間隔性を持って付与されているわけではない．数値上では 4 と 3 の間，3 と 2 の間は両者とも 1 の意味的距離があるが，リッカート尺度においては 4 と 3 の間，3 と 2 の間は等間隔性を持たないため，同一の意味的距離を持たない．つまり，平均や差分といった一般的な分析がリッカート尺度においては問題となる．また，リッカート尺度は主観評価を表現しているため評価者によって同一の評価値であったとしても意味解釈に個人差が生まれてしまう．そのため，分析結果の解釈が不安定となることがある点に注意する必要がある．しかしその一方で，リッカート尺度は主観評価を 2 値ではなく 5 段階や 7 段階といった段階的な表現が可能である．そのため，分析に活用可能な細かいフィードバックが得られるというメリットがある．

2.3　レビュー評価値予測タスク

レビュー評価値予測タスクとは，製品やサービスに対するレビューを入力としてレビューの評価値を予測するタスクである．企業はマーケティングや顧客が持つ不満や改善点の調査など，多種多様な用途でレビュー評価値予測タスクを利用することがある．一般的に 2.4.2 節で示す分類タスクが該当する．

2.4　機械学習

機械学習とは，大量のデータからパターンやルールを自動的に学習することによってモデルがタスクに沿った判断を実施可能にする技術を指す．学習方法については 2.4.1 節で述べる．画像やテキスト，音声など多種多様なドメインに対応することが可能であるため，様々な分野で実践的に用いられている．

2.4.1　学習方法

機械学習の手法は，大まかに三つ存在する．一つ目の手法は，教師なし学習である．教師なし学習とは，正解のないデータから潜在的なルールやパターンを見つけ出す学習手法である．主にクラスタリングや異常検知が該当する．クラスタリングはデータ間の距離やパターンの違いからデータを分類し，最終的にいくつかのグループへとデータを配置していくものである．異常検知は，データ間の距離やパターンの違いからデータの中で異質なものを検知するものである．教師なし学習は教師データが存在しないため，教師あり学習と比較してデータ数が少なくて済む．

二つ目の学習手法は，教師あり学習である．教師あり学習とは，入力と出力のペアデータをもとに，入力に対してどのような出力をするべきかを学習する手法である．学習したルールやパターンをもとに未知のデータに対する予測を実施する．主に分類タスクや回帰タスクが該当する．タスクの詳細に関しては，2.4.2 節で述べる．2.3 節で述べたレビュー評価値予測タスクはこの手法に分類される．教師あり学習はルールやパターンを学習するために膨大な学習データが必要となる．少ない学習データでも精度を向上させるため，交差検証 (Cross-Validation) と呼ばれる手法が用いられることもある．

三つ目の学習手法は，強化学習である．強化学習とは，報酬関数という行動に対する報酬を決定する関数を利用し，報酬が最も高くなる出力方法を学習する手法である．この学習は自律的に行われる．教師データを必要としない上，多種多様なタスクに対応することが可能である．そのため，実世界において医療や機械制御からゲームなどの娯楽まで色々なものに活

用されている．

2.4.2　教師あり学習におけるタスクの種類

教師あり学習のタスクは主に 2 種類存在する．一つ目は，分類タスクである．分類タスクとは，複数のカテゴリがあるときデータがどのカテゴリに属するのかを予測するタスクである．モデルの出力層ノードはカテゴリ数分存在し，各ノードの値がそのカテゴリに該当する確率的な振る舞いをする．ロジスティック回帰やサポートベクターマシン (SVM)[1] が代表的なモデルである．

二つ目は，回帰タスクである．回帰タスクとは，連続値のラベルを予測するタスクである．モデルの出力層ノードは一つであり，そのノードの値がそのまま予測値となる．重回帰分析などの線形回帰や非線形回帰が代表的なモデルである．

2.5　項目反応理論 (Item Response Theory；IRT)

項目反応理論とは，IRT とも呼ばれる統計学的なテスト理論である．従来のテスト理論では，回答者の能力を考慮しない状態で各項目をどれだけ正解したかという点に焦点を当てていた．しかし回答者の能力を考慮せずに適切なテストを作成しようとした場合，テストの難易度に応じた形で回答者全体の正答率分布が変化する．具体的には，高難易度のテストがある場合，高い能力を持つ回答者のみが高い点数を取り，その他の回答者は全体的に低い点数を取ることになる．この場合，低い能力から中程度までの能力を持つ回答者が分布内で集約されてしまい適切な能力を計ることが困難になる．この問題を解決するための方法が IRT である．

IRT は回答者ごとの能力とテストの項目別難易度といった項目特性を分離し，回答者の能力に応じた確率的な点数設定を実施する．この際，項目特性曲線という指標を用いる．項目特性曲線とは，回答者の能力とテストの項目別正答率を対応させたグラフである．縦軸がある項目の正答率である．横軸が回答者の能力である．IRT は，このグラフを用いることによって適切な項目別難易度を設定し，回答者の能力に対する公正な判断を可能とする．

既存研究においては，評価者の判断基準を表現する際に IRT 系のモデルが用いられた．IRT 系のモデルでは，評価者の厳格さなどの特徴をスカラ値の潜在パラメータとして推定する．この際，例外的なレビューケースはノイズとして平均化されてしまう．本稿において我々は，例外的なレビューケースが主観評価に個人差が内在する要因の一つだと考えた．よって我々は，評価傾向の表現に例外的なレビューケースの存在を構造的な情報として保持する混

同行列を用いる．この点が既存研究との差別点の一つである．

2.6　ニューラルネットワーク

ニューラルネットワークとは，人間の脳内細胞であるニューロン間での情報伝達を模して作られた非線形な機械学習モデルを指す．本稿において我々は，非線形な写像を出力する意味解釈の個人差解消という正規化タスクにニューラルネットワークを利用する．

ニューラルネットワークは，ニューロンを模したノードの結合構造を複数層にわたって重ねることによって複雑な学習を可能とする．この時，ニューラルネットワークの層は大まかに三つに分かれる．一つ目は，入力層である．入力層は学習に用いるデータの入力に用いられる層であり，基本的に複数のノードからなる．二つ目は，中間層である．中間層とは，隠れ層とも呼ばれる入力と出力の間にある層を指す．ニューラルネットワークの中間層には，非線形な出力を可能とするために活性化関数という関数に値を通す働きがある．活性化関数は，主にシグモイド関数や ReLU 関数 [2] が用いられる．中間層が 2 層以上重なるニューラルネットワークをディープニューラルネットワーク (DNN) とも呼ぶ．三つ目は，出力層である．出力層はタスクに合わせてノード数を変える．分類タスクでは，分類対象のカテゴリ数分のノードを持つ．そのため，分類タスクの出力層で得られる出力は各カテゴリに該当する確率のような振る舞いをする．回帰タスクでは，ノードの数が一つに固定される．そのため，回帰タスクの出力層で得られる出力は直接予測値として扱うことが可能である．

ニューラルネットワークの学習は予測誤差に基づくパラメータの更新によって行われる．具体的には，ノード間の重みパラメータを誤差逆伝播法 [3] と呼ばれる学習アルゴリズムによって更新し最適化する．誤差逆伝播法とは，予測誤差から導出される損失関数を基に出力層から入力層へ向けてパラメータを遡及的に更新していく方法である．これは勾配降下法と呼ばれる学習アルゴリズムに基づく．勾配降下法とは，損失関数の微分結果と設定された学習率から損失が減少するようにパラメータを更新する学習アルゴリズムである．学習率とは，一回あたりにどれだけ広く重みを更新するかを示すパラメータである．学習率が大きいほど少ない学習回数や訓練データでモデルパラメータの最適化が可能である一方，大きすぎると更新結果が発散・振動し，最適なパラメータに辿り着くことが困難になる可能性がある．

2.7　大規模言語モデル (Large Language Model；LLM)

大規模言語モデルとは，LLM とも呼ばれる大規模なパラメータと豊富な学習資源を用いて訓練した言語モデルのことである．言語モデルとは，単語系列におけるそれぞれの条件付き

生成確率を割り振る確率モデルである．つまり，言語モデルは単語を確率的に決定して適切な文章を生成する．モデルの多くは Transformer[4] と呼ばれるアーキテクチャを基礎としている．Transformer とは，2017 年に Google が発表した深層学習用のアーキテクチャである．Transformer の中核に Attention 機構が存在する．Attention 機構とは入力されたテキストにおける各トークンが他のどのトークンと関連づけられているものなのかを把握する機構である．Transformer は Attention 機構を導入することによって，RNN[5] が構造的に抱えていた長距離依存関係の学習困難性を緩和することを可能とした．計算時は各トークンに対して Key,Value,Query の 3 行列を作成し，Key 行列と Query 行列の類似度を計算したものと Value 行列の加重平均を算出した値を用いてトークン間の依存関係を獲得する．トークン間の依存関係の把握によって，LLM は長文に対する自然で適切な応答を生成することが可能となる．

LLM の学習について述べる．LLM の学習段階は大まかに三つに分けられる．一つ目は，事前学習である．事前学習では，大規模コーパスを用いた自己教師あり学習によって語彙や文法のような基礎的な言語理解を獲得する．LLM は，自己教師あり学習として Next Token Prediction を実施する．Next Token Prediction とは，入力されたトークンの次にあたるトークンに対して条件付き生成確率を予測するタスクである．二つ目は，Fine-Tuning である．Fine-Tuning では，教師あり学習によってモデルの性能改善や特定タスクへの適応を実現する．LLM は，教師あり学習として Instraction Tuning を実施する．Instraction Tuning とは，タスク指示とそれに付随した補足情報から与えられたプロンプトに対する理想的な応答を学習する手法である．これにより，LLM は多種多様な入力プロンプトに対する理想的な出力を学習する．三つ目は，Reinforcement Learning from Human Feedback(RLHF)[6] である．RLHF では，人間によるフィードバックを用いた強化学習によって応答時の方向性や価値観を人間に沿うよう調整する．具体的には同じテキストに対する応答を複数パターン生成し，それぞれの回答に対する人間の評価を利用して強化学習を実施する．これにより，LLM はユーザへ正確な応答を行うだけでなく，差別や悪意のある発言を行わないようになる．

本稿において我々が利用する LLM は，Gemma3[7] の 27B モデルである．Gemma3 は 2025 年に Google から公開されたオープン LLM モデルであり，Ollama や Llama.cpp，LM Stadio など複数の開発者ツールで公開されている．Gemma3 は，128,000 トークンに及ぶコンテキスト長を誇る．また，Gemma3 は 140 を超える言語に対応している．パラメータサイズは，1B，4B，12B, 27B の 4 種類が公開されている．量子化は，32bit から 4bit まで 5 種類のレベルが公開されている．

2.8 文脈内学習 (In-Context Learning；ICL)

文脈内学習とは ICL とも呼ばれる LLM の回答調整手法である．LLM は，与えられたプロンプトにあるタスクの補助情報や推論挙動の指示を利用し，出力を柔軟に変更する．そのため LLM は，ICL を用いることによってプロンプトを通じて与えた学習用データや推論挙動の変更指示に対応し，まるで再学習を行ったかのような上質な回答を出力することが可能となる．ICL は，GPT-3[8] の発表時に広く認知されるようになった．代表的な手法として，Few-Shot Prompting が存在する．Few-Shot Prompting とは，モデルの推論時に少数の参考となる例を与える ICL である．つまり，Few-Shot Prompting を実施した場合，LLM はプロンプト内で与えられた参考データに基づいた条件付きの推論を実施する．

Few-Shot Prompting は，推論対象データと指示のみを与えて参考データを与えない Zero-Shot Prompting や 一つだけ参考データをモデルに例示する One-Shot Prompting と比較して，より柔軟で参考データに沿った応答を可能とする．また，Few-Shot Prompting は Fine-Tuning と異なりパラメータを更新しない．そのため，Few-Shot Prompting は Fine-Tuning ほどのデータを必要としない．つまり，Few-Shot Prompting は Fine-Tuning と異なり，少数データにしか存在しない局所的な推論方針をモデルに提示することも可能である．

2.9 類似度計算手法

ベクトル間の類似度計算は主に 4 種類の計算方法に分かれる．一つずつ説明する．

2.9.1 距離ベース

一つ目は，距離ベースの類似度計算手法である．代表的なものに，L1 ノルムと L2 ノルムが存在する．L1 ノルムは，マンハッタン距離とも呼ばれる距離計算の方法である．二つのベクトル $\mathbf{x} = [x_1, x_2, \cdots, x_d]$ と $\mathbf{y} = [y_1, y_2, \cdots, y_d]$ がある時，L1 ノルムは以下の式で表せる．

$$L_1 = \|\mathbf{x} - \mathbf{y}\| = \sum_{i=1}^{d} |x_i - y_i| \tag{2.9.1}$$

L1 ノルムは，この式にあるように絶対値による距離計算を実施する．そのため外れ値に強い一方で，ベクトル間の要素ごとの差を重視するという特徴がある．

L2 ノルムは，ユークリッド距離とも呼ばれる距離計算の方法である．L2 ノルムは以下の式

で表せる．

$$L_2 = \|\mathbf{x} - \mathbf{y}\| = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2} \tag{2.9.2}$$

L1 ノルムは，この式にあるように差の二乗による距離計算を実施する．そのため外れ値に弱い一方で，ベクトル間の直線的な距離を重視するという特徴がある．両計算では，ベクトル間の距離が 0 に近いほど類似度が高いことを示す．

2.9.2 角度ベース

二つ目は，角度ベースの計算手法である．代表的なものに，cos 類似度が存在する．cos 類似度は以下の式で表せる．

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{d} x_i y_i}{\sqrt{\sum_{i=1}^{d} x_i^2} \sqrt{\sum_{i=1}^{d} y_i^2}} \tag{2.9.3}$$

cos 類似度は，ベクトルの内積を利用する．そのため，ベクトルの大きさに依存せずベクトルの向きがどれほど類似しているか計算できる．cos 類似度は-1 から 1 の値を取り，0 に近いほど類似度が高いことを示す．

2.9.3 分布ベース

三つ目は，分布ベースの計算手法である．代表的なものに，KL ダイバージェンスが存在る．KL ダイバージェンスは，正確には類似度計算手法ではない．しかし，ベクトルが確率分布と見做せる場合は類似度計算手法として扱うことが可能となる．KL ダイバージェンスは以下の式で表せる．

$$D_{\mathrm{KL}}(\mathbf{x} \,\|\, \mathbf{y}) = \sum_{i=1}^{d} x_i \log \frac{x_i}{y_i} \tag{2.9.4}$$

この時，KL ダイバージェンスは $\mathbf{x}$ を真の分布，$\mathbf{y}$ を近似分布として基準を設定する．そのため，$D_{\mathrm{KL}}(\mathbf{p} \,\|\, \mathbf{q}) \neq D_{\mathrm{KL}}(\mathbf{q} \,\|\, \mathbf{p})$ である．KL ダイバージェンスは，分布がどれだけ異なっているかを表す．そのため，0 に近いほど類似度が高いことを示す．

2.9.4 構造ベース

四つ目は，構造ベースの計算手法である．代表的なものに SSIM[9] や MS-SSIM[10] ，Block-SSIM が存在する．SSIM の式は以下のように表せる．

$$\mathrm{SSIM}(\mathbf{x}, \mathbf{y}) = l(\mathbf{x}, \mathbf{y})^{\alpha} c(\mathbf{x}, \mathbf{y})^{\beta} s(\mathbf{x}, \mathbf{y})^{\gamma} = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2.9.5}$$

μ_x, μ_y は各行列の要素平均値，σ_x, σ_y は各行列の標準偏差，σ_{xy} は二つの行列における共分散である．また，C_1 と C_2 は計算時に 0 で除算しないよう設定するための安定化定数である．α, β, γ は $\alpha, \beta, \gamma \in \mathbb{R}_{\geq 0}$ の重みである．基本的な SSIM では，$\alpha = \beta = \gamma = 1$ である．SSIM は前述した式の通り，輝度 $l(\mathbf{x}, \mathbf{y})$，コントラスト $c(\mathbf{x}, \mathbf{y})$，構造 $s(\mathbf{x}, \mathbf{y})$ の 3 要素に着目して類似度計算を実施する．それぞれは以下の式で表される．

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{2.9.6}$$

$$c(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{2.9.7}$$

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{2.9.8}$$

この時，C_3 は C_1, C_2 と同様に安定化定数である．本稿において我々は，SSIM を混同行列の構造的な情報を維持した類似度計算のために使用する．SSIM は理論上-1 から 1 の範囲を取り，1 に近いほど類似度が高いことを示す．

MS-SSIM の式は以下の式で表せる．

$$\text{MS-SSIM}(\mathbf{x}, \mathbf{y}) = [l_N(\mathbf{x}, \mathbf{y})]^{\alpha_N} \prod_{j=1}^{N} \left[c_j(\mathbf{x}, \mathbf{y})^{\beta_j} s_j(\mathbf{x}, \mathbf{y})^{\gamma_j}\right] \tag{2.9.9}$$

この時，N はスケール数を表す．MS-SSIM は，ガウシアンフィルタを用いたダウンサンプリングによって解像度を落としたしたもの，つまり複数のスケールを乗算する計算手法である．そのため，単一スケールの SSIM と比較して人間の評価と近しい結果が得られるという特徴がある．本稿において我々は，考察の幅を広げるために擬似的な MS-SSIM を利用した．擬似的である理由は，混同行列において解像度という概念が存在しないためである．MS-SSIM も SSIM と同様に，1 に近いほど類似度が高いことを示す．

Block-SSIM は以下の式で表せる．

$$\text{Block-SSIM}(\mathbf{x}, \mathbf{y}) = \frac{1}{M} \sum_{j=1}^{M} \text{SSIM}(x_j, y_j) \tag{2.9.10}$$

この時，M は分割したブロックの総数である．Block-SSIM は，小領域のブロックごとに計算した SSIM に対して平均を取る計算手法である．本稿において我々は，MS-SSIM と同様に考察の幅を広げるため Block-SSIM を利用した．Block-SSIM も SSIM と同様に，1 に近いほど類似度が高いことを示す．

2.10　精度評価指標

精度評価指標はタスクの種類に応じてそれぞれ複数存在する．

一つ目は，分類タスクで用いられる精度評価指標である．代表的なものに，Accuracy や Precision，Recall，F1-score が存在する．これらの指標は，2.11 節で述べる混同行列を基に計算される．二つ目は，回帰タスクで用いられる精度評価指標である．代表的なものに，決定係数 (R^2) や RMSELoss，MAELoss が存在する．一つずつ説明する．

2.10.1　分類タスク

Accuracy は，以下の式で表せる．

$$Accuracy = \frac{1}{n}\sum_{k} T_k \tag{2.10.1}$$

k はラベル，T_k はあるラベル k に対して予測ラベルと正解ラベルが一致しているデータの個数，n はデータの総数である．Accuracy は予測ラベルが正解ラベルと合致しているものの割合を示す指標である．そのため，最も直感的な理解が可能な精度指標である．一方で，Accuracy は特定のラベルのみを予測してしまう場合においてもある程度の精度が担保できてしまうという問題点がある．そのため，正確な分析には他の指標を追加で用いる必要がある．Accuracy は 1 に近いほど精度が高いことを示す．

Recall は，以下の式で表せる．

$$\mathrm{Recall} = \frac{1}{|K|}\sum_{k\in K} \frac{T_k}{T_k + F_k^{\mathrm{FN}}} \tag{2.10.2}$$

F_k^{FN} は正解ラベルが k である場合において予測されたラベルが異なるデータの個数，$|K|$ は分類タスクにおけるカテゴリ集合の要素数である．Recall は正解ラベルに属するデータのうち，どれだけ正しく予測できたかを示す指標である．そのため，予測時における取りこぼしの少なさを評価する際に有効である．Recall は 1 に近いほど精度が高いことを示す．

Precision は，以下の式で表せる．

$$\mathrm{Precision} = \frac{1}{|K|}\sum_{k\in K} \frac{T_k}{T_k + F_k^{\mathrm{FP}}} \tag{2.10.3}$$

F_k^{FP} は予測ラベルが k である場合において正解ラベルが異なるデータの個数である．Precision は予測されたデータのうち，どれだけ正解であったのかを示す指標である．そのため，誤検出の少なさを評価する際に有効である．Precision は 1 に近いほど精度が高いことを示す．

F1-score は，Precision と Recall の調和平均として定義される．そのため，以下の式で表される．

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{2.10.4}$$

F1-score は，総合的な予測精度の良さを示す指標である．F1-score は 1 に近いほど精度が高いことを示す．また，F1-score は，macro-f1-score と micro-f1-score の 2 種類が存在する．

macro-f1-score は，あるカテゴリを陽性，それ以外のカテゴリを陰性として算出した F1-score を平均することで求められる．そのため，macro-f1-score は特定カテゴリを重要視する必要のある分類問題に対する評価として適している．一方，micro-f1-score は，全カテゴリにおける F1-score をまとめて計算することで求められる．つまり，多クラス分類においては Accuracy と同義である．

2.10.2 回帰タスク

決定係数（R^2）は，以下の式で表される．

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.10.5}$$

y_i は正解値，$\hat{y}_i$ は予測値，$\bar{y}$ は正解値の平均，n はデータ数である．決定係数 (R^2) は，目的変数の分散がどの程度モデルによって説明されているかを示す指標であり，また，決定係数 (R^2) は 1 に近いほど精度が高いことを示す．

RMSELoss は，以下の式で表される．

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.10.6}$$

RMSE は，予測誤差を元の値と同じ単位で解釈することが可能な指標である．RMSELoss は 0 に近いほど精度が高いことを示す．

MAELoss は，以下の式で表される．

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.10.7}$$

MAE は，予測誤差の絶対値の平均を示す指標である．そのため，外れ値の影響を比較的受けにくい．MAELoss は 0 に近いほど精度が高いことを示す．

2.11　混同行列

混同行列は、機械学習の分類タスクにおいてモデルの性能を評価・可視化するために用いられる正解値と予測値の対応表である．この対応表を基に，2.10.1 節で述べた精度評価を算出する．混同行列によって Accuracy 以外の精度評価指標も容易に算出可能であるため，誤検出や予測の取りこぼしを正確に評価することが可能である．本稿においては，混同行列をレビュアの評価傾向を表現するために用いる．評価傾向を表現する際，我々は Acuuracy や Precision，Recall などの評価指標による表現ではなく，混同行列そのものを用いる．Accuracy や Precision，Recall はそれぞれ対角線，行，列といった一つの方向に対する要素間のみしか対応関係を考慮できない．我々は，例外的なレビューケースのように対角線上から離れた位置に要素が点在するケースに対して，これらの評価指標による対応が困難であると考えた．また我々は，同一の行や列に例外的なレビューケースが集合していた場合においても，Precision や Recall は平均を取るため要素の構造的な情報を維持することが不可能であると考えた．一方で，混同行列を直接用いる場合，要素の位置やばらつきといった構造的な情報を維持したままの表現が可能となる．

2.12　対応のある 2 標本 t 検定

対応のある 2 標本 t 検定は，対応関係にある二つの母集団から抽出した標本に対して母平均の差が存在するかどうかを調査するための検定方法である．対応関係にあるとは，同一の被検対象において母集団を形成した時期が異なる場合など，二つの標本抽出における基盤が共通である状況を指す．この検定における帰無仮説は「二つの母集団の平均に差がない」である．検定統計量 t は以下の式で定義される．

$$t = \frac{\sqrt{n}(\bar{d} - \mu)}{s} \tag{2.12.1}$$

この時，n はデータ数，$\bar{d}$ は二つの標本データに対する差の平均，μ は差の母平均，s は二つの標本データに対する差の標準偏差である．μ は帰無仮説が「二つの母集団の平均に差がない」であるため，0 が代入される．t 検定では上記の検定統計量から導かれる p 値に対して，有意水準を満たすかどうかを調査する．p 値とは，統計的検定において帰無仮説が正しいと仮定した際に得られたデータ外の結果が偶発する確率である．t 検定に用いるデータは正規分布に従っている必要がある．

第 3 章　関連研究

本研究は，レビュアの評価傾向を用いてリッカート尺度に代表される段階的な評価尺度に基づく主観評価に内在する個人差を補正することを目的とする．主観評価およびレビュー分析に関する既存研究は研究目的や評価対象に応じて異なる観点から問題設定がなされてきた．そのため，主観評価およびレビュー分析に関する既存研究は多岐にわたる．

本稿では，既存研究を以下の三つの観点から整理する．

1. 段階的な評価尺度に基づく主観評価の扱いに関する研究
2. 主観評価やレビュー情報を用いた評価値予測タスクに関する研究
3. 評価者やユーザの特徴表現および類似度計算に関する研究

この整理により，既存研究の多くが主観評価に内在する個人差や評価者特徴を評価値分布や評価者の厳しさ・信頼度といった低次元の指標によって表現していることを示す．また，既存研究がレビュー内容と評価値の乖離として現れる例外的なレビューケースを構造的に保持できていないという限界を示す．

3.1　段階的な評価尺度に基づく主観評価の扱いに関する研究

本節では，リッカート尺度に相当する段階的評価尺度そのものの意味解釈に着目した研究を整理する．リッカート尺度に相当する段階的な評価尺度に基づく主観評価を正規化する試みは十分に議論されていない．しかし既存研究では，段階的な評価尺度に基づく主観評価を対象に評価者バイアスの分析や主観評価に内在する不確実性の定量化を目的としているものが多い．Yeşilçınar ら [11] は，リッカート尺度に相当する段階的な評価に基づくピア評価を対象とし，評価環境が評価者バイアスに与える影響を定量的に明らかにすることを目的とした．ピア評価とは，同一または類似した専門分野に属する評価者同士が互いの成果物を評価し合う活動を指す．評価環境とは，対面やオンライン，匿名といった評価時の環境を指す．評価者バイアスとは，段階評価において中央付近の評価を付与しやすいという中央化効果 [12] や，相手の学歴や自分との関係性といった他の目立つ情報によって正確な評価が歪められてしまうハロー効果 [13] などを指す．Yeşilçınar らは，評価者同士の関係性と評価環境に応じて主観に基づいたピア評価の信頼性が変動すると考えた．同研究において Yeşilçınar らは，IRT を拡張したモデルである MFRM を用いた．MFRM は，評価者の厳しさや評価環境の影響をスカラ値の潜在パラメータとして推定する．そのため，各レビュー単位における評価理由と評価値の乖離は誤差項として吸収される．つまり，MFRM は評価理由と評価値

が大きく乖離する例外的なレビューケースを構造的に保持することが困難なモデルである．また，MFRM はレビュー内容そのものをモデル化の対象としていないことも特徴である．実験では学生同士のピア評価および教員による評価を比較し，評価環境を含む誤差要因のうちどの要因が評価の信頼性を損なうのか定量的に調査した．実験結果より，学生同士の評価においては匿名性の有無によって評価信頼性に有意な差が生じることを示した．また，教員による評価では匿名評価の有無によって評価信頼性に有意な差は確認できなかった．

Anjaria ら [14] は，リッカート尺度に代表される主観評価に内在する個人差から生まれた尺度の不確実性や曖昧さを解決することを目指した．その際 Anjaria らは，z-number[15] を用いることによって段階的な評価尺度における曖昧さを明示的に取り扱うことが可能であると考えた．この考えを基に，Anjaria らは z-number を用いた主観評価の表現手法を提案した．z-number は，評価値に対応する制約とその評価に対する信頼度をファジィ数として表現する枠組みである．制約は，段階的な尺度の各項目において回答がどのように分布しているのかを表す．信頼度は，評価者がどれだけ確信を持って評価したのかを表す．ファジィ数とは，大きさや信頼性といった概念的な値を離散値ではなくどの程度大きいのか，信頼できるのかといった連続値として表現した数を指す．つまり z-number に基づく表現は，評価者の評価傾向を評価値の集合および信頼度のスカラ値を用いて集約した要約的な情報で表現する．そのため，個々のレビューにおける評価理由と評価値の対応関係を構造的に保持することは困難である．また，前述した MFRM はと同様に例外的なレビューケースは曖昧さの一部として吸収される．実験では，講義に対する教師と学生の認識の違いを z-number を用いて分析した．実験結果より，教師は学生と比較して高い信頼度を持って主観評価を行っていることを確認した．また学生が否定評価をしている場合は信頼性が低い，つまり反対の意味だけではなく理解ができていないという状態であることを確認した．

これらの研究は，主観評価に対する影響や主観評価の信頼性を定量的に評価すること目的としており，主観評価において同一の評価値が必ずしも同一の意味を持たないことを示している．またこれらの研究では，評価理由と評価値が大きく異なる例外的なレビューケースをノイズとして平均化することが暗黙的に仮定されている．しかし，いずれの研究においても段階的な主観評価を意味解釈の個人差が内在しない参照尺度に基づく意味評価へと正規化する試みは十分に議論されていない．本研究はこの点に着目し，主観評価に内在する個人差を構造として扱うことによって評価値に対する意味的正規化を行う点に新規性がある．特に本研究は既存研究と異なり，例外的なレビューケースを個人差の解消のための重要な要素であると位置付ける．つまり，本研究は例外的なレビューケースの存在をノイズとして平均化するのではなく，例外的なレビューケースの存在を構造情報として維持したまま主観評価を扱う点に特徴がある．本稿において我々は，評価者の判断傾向をレビューと評価値の対応関係を集約した混同行列で捉える．

3.2　主観評価やレビュー情報を用いた評価値予測タスクに関する研究

主観評価やレビュー情報を用いた評価値予測タスクに関する研究は現在までに複数行われている．これらの研究の主目的はレビュー文に含まれる感情表現やユーザ・製品固有の特徴量を用いて評価値を高精度に予測する点にある．つまりこれらの研究において，主観評価に内在する個人差は予測精度向上のための要因の一つとして扱われる場合が多い．したがって本節では，主観評価に内在する個人差を予測精度向上のための要因の一つとして扱う側面に着目し，主観評価やレビュー情報を用いた評価値予測タスクに着目した研究を整理する．

Wang ら [16] は，レビューに加えてユーザおよび製品固有のコンテキスト情報を潜在表現として統合することにより，評価値予測精度の向上を目指した．コンテキスト情報とはユーザや製品に固有の特徴を表すものである．同時に，コンテキスト情報とはレビューや製品説明に含まれる感情表現が各ユーザや各製品で異なる影響を持つことを潜在的に表現したものである．同研究では，ユーザや製品ごとの評価の偏りはモデル内部の潜在変数として暗黙的に表現される．つまり，コンテキスト情報は Wang らの研究における主観評価の個人差であり，明示的に設定されたパラメータではない．実験では，6 つのベースライン手法とユーザや製品の固有特徴をレビューに統合した提案手法を比較した．実験結果より，コンテキスト情報を統合することが評価値予測精度向上の観点で有効であることを確認した．

坂本ら [17] は，レビューからレビュアの評価値を予測するモデルを構築し，その予測誤差を用いてユーザの嗜好に適合したレビューを抽出する手法を提案した．この研究は，レビュアが付与したアイテムへの評価値とレビューからモデルが推測したアイテムへの評価値間の差が小さいものを，レビュアの嗜好に合うレビューとしている．そのため，坂本らの研究において評価値予測はレビュアの嗜好を含むレビューを識別するための手段として位置づけられている．実験では，実際の評価値と推測の評価値に差があるものは具体的な内容のないレビュー，もしくはユーザの嗜好のレビューであるという仮定を検証するために，Fine-Tuning を活用した BERT[18] モデルによる評価値予測を実施した．実験結果より，提案手法によって具体的な内容のある役に立つレビューが抽出できることを確認した．

これらの研究は，多数のデータに基づく平均的な評価傾向を捉えるためにユーザや製品に固有の特徴をモデル内部の潜在表現として扱うことの有効性を議論している．つまり，評価値を当てることを主目的としたタスクの一つとして満足度予測が用いられている．また，これらの研究は 3.1 節の研究と同様に，評価値とレビューが大きく異なるような例外的なレビューケースを損失最小化の過程においてノイズとして平均化してしまうという側面を持つ．これらの研究で用いられた Fine-Tuning は，例外的なデータをノイズとして平均化する側面を持つ代表的な手法である．Fine-Tuning は全体の平均的な損失を最小化することに

よって学習を進める．そのため Fine-Tuning は，多数のデータに基づく平均的な評価傾向を捉える点で有効である．しかし，例外的なデータは全体的な最適化方針に適さないため，外れ値として扱われることになる．つまり，例外的なデータの存在がモデル出力の方針にほとんど影響を与えない．一方で本研究は既存研究と異なり，評価傾向と例外的なレビューケースの関係性調査や混同行列による評価傾向表現の妥当性検証，そして主観評価における意味解釈の個人差を解消した参照尺度に基づく主観評価への正規化のために満足度予測タスクを活用する．つまり，他概念に対する間接的な評価や評価値の意味的正規化を実施するためのタスクとして満足度予測が用いられている．また本稿において我々は，例外的なレビューケースの存在を重要な要素として扱うために Fine-Tuning ではなく評価傾向が反映されたデータを推論時の文脈として提示する枠組みを採用する．具体的には，我々は LLM に対する Few-Shot Prompting を用いて例外的なレビューケースの存在を含めた評価傾向を反映した推論を実施する．Few-Shot Prompting は，提示された例を判断の参考とする条件付き推論である．つまり，参考データに例外的なレビューケースの存在を含めた評価傾向を反映する場合，LLM はレビュアの評価傾向を参考にして回答を出力する．これにより，我々は Few-Shot Prompting が例外的なレビューケースを考慮した柔軟な判断に基づく出力を可能とすると考えた．本稿における Few-Shot Prompting は，評価傾向をモデルに学習させることを目的としたものではない．本稿における Few-Shot Prompting は評価傾向を構造情報として保持したまま，評価傾向と例外的なレビューケースの関係性調査や混同行列による評価傾向表現の妥当性検証，そして主観評価における意味解釈の個人差を解消した参照尺度に基づく主観評価への正規化を実施するための手段として位置付ける．

3.3　評価者やユーザの特徴表現および類似度計算に関する研究

レビューや主観評価に基づいて推定したレビュア間の類似度を推薦や検索に応用する研究も複数行われている．これらの研究は，レビュアの嗜好や関心といった主観的な要素の類似性を推定することに主眼を置いている．本節では，レビュアの嗜好や関心といった主観的な要素に内在する個人差を類似レビュア計算に扱う側面に着目し，主観的な要素に対する類似度計算によって類似レビュアを推定する既存研究について整理する．

松岡ら [19] の研究では，肌の特徴や悩みといった個人差を考慮した類似ユーザ判定に基づくコスメアイテム推薦手法を提案した．松岡らは，ユーザ間の類似性を推定するために二種類のユーザ特徴ベクトルを構築した．一つ目は，ユーザが自己申告した肌悩みにおいて気になる部位を表す特徴ベクトルである．二つ目は，ユーザが過去に投稿したレビューから評価表現を抽出して計算した，部位ごとの肌特徴や悩みを反映した特徴ベクトルである．その後，松岡らは構築した二種類のユーザ特徴ベクトルに対して cos 類似度を適用し，推薦に活用し

た．実験では，提案手法の有効性を確認するために，作成した推薦システムに対してユーザによる評価実験を実施した．実験結果より，提案手法の有効性に加え，二つのユーザ特徴ベクトルを連結する方式での類似度計算を活用した提案手法の有用性を確認した．

宮本ら [20] の研究では，感性の違いに基づく印象の差異が共通したレビュアにおける観点の類似性を利用した書籍推薦手法を提案した．宮本らは，レビュー中に含まれる感情語に対して tf-idf[21] を適用することによって，ユーザごとの感性ベクトルを構築した．そして構築した感性ベクトル間の類似度を cos 類似度によって計算し，類似した印象を持つ書籍の推薦を行った．この際，宮本らは類似度計算の対象を同一ユーザが書いたレビューに限定した．それにより，ユーザの特定観点における主観的な要素の類似性を捉えることを試みた．実験結果より，提案手法の有効性が明確に確認できなかった一方で，ユーザが読みやすいと感じる書籍傾向が考慮されている可能性は示唆された．

これらの研究は，ユーザ固有の感性や特性の意味的類似性を推定することを目的として，レビューや評価表現から抽出したユーザ特徴を対象とした類似度計算を実施する．そのためいずれの研究においても，ユーザ特徴を表す一次元の特徴ベクトルを対象に cos 類似度を用いた類似度計算を実施する．つまり，これらの研究においてはユーザ固有の感性や特性の構造的な類似性を推定する問題は扱われておらず，構造的な情報を維持したまま類似度を計算する枠組みは提示されていない．本研究はこの点に着目し，混同行列に対する構造的な類似度計算を用いて混同行列による評価傾向表現の妥当性を定量的に評価する点に新規性がある，つまり，本稿における混同行列を対象とした類似度計算は構造的な類似性を計算することが目的であり，レビュアごとの評価傾向における意味的な類似性を推定すること自体が目的ではないという点で既存研究と異なる．したがって本研究において我々は，混同行列という二次元のベクトルを対象に SSIM を用いた類似度計算を実施する．我々は，既存研究と同様の方法で混同行列の類似度を計算する場合，例外的なレビューケースに基づく評価のばらつきや特定評価値への偏りといったレビュアによって異なる構造的な評価傾向の情報が類似度計算の過程で失われてしまうという問題点があると考えた．我々はその問題点を解決するために，混同行列を平滑化せず構造的な情報を維持したまま用いる．また，我々は構造的な情報を考慮可能な類似度計算指標として，画像分野のノイズ判定タスクなどで用いられる SSIM を採用する．我々は混同行列に対して SSIM を使用することによって，混同行列における要素の大きさや位置，ばらつきといった構造的な評価傾向の情報損失を最低限に抑えることが可能となると考えた．

第 4 章　提案手法

本研究の目的は，リッカート尺度に基づく主観評価に内在する意味解釈の個人差を解消することである．我々は本研究の目的を達成するために，レビュアごとの評価傾向を利用した意味解釈の個人差が内在しない参照尺度への正規化手法を提案する．

既存研究における評価者の評価基準は，3.1 節で述べたように集合やスカラで表される．しかし既存研究で用いられる方法を正規化に適用する場合，例外的なレビューケースにおける主観評価はノイズとして平均化されてしまう．例外的なレビューケースの一例を図 4.1 に示す．例外的なレビューケースとは，レビュアが実際に付与した満足度とレビューから推測される満足度が大きく異なるレビューのことを指す．例外的なレビューケースにおける主観評価がノイズとして扱われてしまう場合，我々は評価者の正確な評価傾向をもとに個人差を補正しているとは言い難いと考えた．本稿において我々は，レビュアの評価傾向をレビュアが付与した満足度と評価理由を表すレビューから LLM の Zero-Shot 推論によって推測される満足度の共起頻度，つまりレビュア単位の Zero-Shot 満足度予測によって得られる混同行列として定義する．詳細は 4.1.1 節で述べる．また本稿において我々は，正規化後の意味解釈の個人差が存在しない参照尺度に基づく評価値を LLM による Zero-Shot 推論の予測値で定義する．詳細は 4.1.2 節で述べる．

4.1　提案手法概要

提案手法のフローは，主に二つの Step に分けられる．

Step 1　混同行列を用いた評価傾向の用意
Step 2　ニューラルネットワークを用いたリッカート尺度の正規化

===レビュー===
スタッフには何の不満もない。が、グループ利用者がうるさい.
部屋もいつもより汚い。前に来た時はよかったのに。
グループを角部屋にするという配慮くらいしてほしい。
===評価値===
レビュアが付与した実際の評価値：5
レビューから推察するレビュアの評価予測値：2

図 4.1: 例外的なレビューケースの一例

レビュアの集合を $U = \{u_0, u_1, \cdots, u_i, \cdots, u_n\}$，レビューの集合を $R(u_i) = \{r_0, r_1, \cdots, r_j, \cdots, r_m\}$ とする．この時，$0 \leq i \leq n$，$0 \leq j \leq m$ である．レビュー r_j にはそのレビューに付与された評価値 y_j が付随する．評価値カテゴリ集合を Y とする時，評価値 y_j は $y_j \in Y = \{1, 2, \ldots, k\}$，$k \in \mathbb{N}$ である．したがって，あるレビュア u_i が持つデータ集合は $D(u_i) = \{(r_0, y_0), (r_1, y_1), \cdots, (r_j, y_j), \cdots, (r_m, y_m)\}$ で表す．また，サンプリングした複数のレビュア $U_s = \{u_0^s, u_1^s, \cdots\}$ が持つデータ集合は $D(U_s) = \{D(u_0^s), D(u_1^s), \cdots\}$ で表す．この時，$|D(u_i)| \geq 1, |D(U_s)| \geq 1$ である．評価傾向は $|D(u_i)|$ や $|D(U_s)|$ が大きいほど明瞭に表現することが可能となる．

Step 1 で我々はレビュアごとの評価傾向を用意する．Step 2 で評価傾向と実際に付与された評価値をもとに，ニューラルネットワークを用いた評価値の意味的正規化を行う．詳細は 4.1.1 節と 4.1.2 節で述べる．

4.1.1 評価傾向の用意 (Step 1)

本稿において我々は主観評価に内在する個人差を定量的に解釈するために，レビュアが書いたレビュー r_j とそれに付随して付与された満足度 y_j の対応関係に着目する．対応関係を定量的に示す際，我々はレビュアが持つデータ集合 $D(u_i)$ における満足度間の差ではなく満足度間の共起関係を用いる．満足度間の差ではなく共起関係を用いる理由は，順序尺度であるリッカート尺度に基づく評価値間の差が必ずしも同一だと言い切れないためである．このことから，本稿において我々はデータ集合 $D(u_i)$ の共起構造を集約したデータであるレビュア単位の満足度予測における混同行列 $C^{(u_i)}$ を，レビュアの評価傾向を表す情報として利用する．混同行列 $C^{(u_i)}$ は表 4.1 に示すように，評価値カテゴリ集合の要素数 k に依存した行列サイズをもつ．つまり $C^{(u_i)} \in \mathbb{R}^{k \times k}$ であり，混同行列の各要素 $C_{p,q}^{(u_i)}$ は $1 \leq p \leq k, 1 \leq q \leq k$ である．具体例を表 4.2a と表 4.2b，表 4.2c，表 4.2d に示す．混同行列 $C^{(u_i)}$ の対角線上にある要素を濃い灰色で示す．それ以外の非ゼロ要素を薄い灰色で示す．表 4.2a に示すように対角線上に要素が集合している場合，つまり $C_{p,q}^{(u_i)}$ における $p = q$

表 4.1: レビュア u_i に対する満足度予測の混同行列

予測満足度 $\hat{y}_j$	実際の満足度 y_j			
	1	2	$\cdots$	k
1	$C_{1,1}^{(u_i)}$	$C_{1,2}^{(u_i)}$	$\cdots$	$C_{1,k}^{(u_i)}$
2	$C_{2,1}^{(u_i)}$	$C_{2,2}^{(u_i)}$	$\cdots$	$C_{2,k}^{(u_i)}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
k	$C_{k,1}^{(u_i)}$	$C_{k,2}^{(u_i)}$	$\cdots$	$C_{k,k}^{(u_i)}$

の位置に要素が集合している時，そのレビュアはレビューを順当に満足度へと反映している．反対に表 4.2b に示すように，対角線上に要素が集合せず要素がばらついている場合，そのレビュアは言及した要素とは関係のない評価を付与している．表 4.2b における，$C_{1,4}^{(u_i)}$ や $C_{3,5}^{(u_i)}$ が例外的なレビューケースに該当する要素となる．また表 4.2c や表 4.2d に示すように，混同行列の要素が少ない場合や極端な評価が行われる場合も存在する．加えて 3.1 節で述べた既存研究のように，一次元のベクトルやスカラ値は隣接した評価値間に共通するポジティブ／ネガティブといった感情情報が同一であるか否かを表現できない．このことから，我々は一次元のベクトルやスカラ値による評価傾向の表現では，近しい評価値間の意味的な類似性が失われてしまうと考えた．一方で混同行列は，隣接した評価値間に共有される感情情報を実際に付与された満足度 y_j と推測された満足度 $\hat{y_j}$ の対応関係である要素の位置やばらつきによって構造的に保持することが可能となる．以上の事実より，我々は混同行列 $C^{(u_i)}$ が主観評価に内在する個人差を定量的に表現するために適切であると考えた．本稿において我々は，$C_{p,q}^{(u_i)}$ における $p = q$ の位置に要素が集合しているレビュアを多数のレビュアが持つ平均的な判断基準に基づいた評価傾向を持つレビュアだと定義する．また本稿において我々は，混同行列 $C^{(u_i)}$ において対角線上に要素が集合せず要素がばらついているレビュアを特殊な評価傾向を持つレビュア，つまり個人差が内在する主観評価を行うレビュアだと定義する．

表 4.2: 評価傾向の違いを表す混同行列の例

(a) 順当に満足度を反映している場合

$\hat{y}\backslash y$	1	2	3	4	5
1	1	2	0	0	0
2	0	4	3	0	0
3	0	0	11	9	0
4	0	0	0	38	25
5	0	0	0	0	7

(b) 評価とレビューが乖離している場合

$\hat{y}\backslash y$	1	2	3	4	5
1	0	0	1	2	0
2	1	2	7	4	0
3	0	1	4	6	1
4	0	0	2	6	0
5	0	0	0	3	0

(c) レビュー数が極端に少ない場合

$\hat{y}\backslash y$	1	2	3	4	5
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	2	0
4	0	0	0	1	0
5	0	0	0	0	2

(d) 極端な判断を行う場合

$\hat{y}\backslash y$	1	2	3	4	5
1	0	0	0	0	0
2	4	0	0	0	2
3	16	1	4	0	4
4	3	0	0	0	14
5	0	0	0	0	2

本稿において我々は混同行列 $C^{(u_i)}$ を計算するために，実際に付与された満足度 y_j と Zero-Shot 推論によって推測された満足度 $\hat{y_j}$ の共起頻度を用いる．共起頻度とは，特定の事象が他の事象と一緒に出現する回数やその傾向を数値化したものである．Zero-Shot 推論とは，LLM の判断時に参考となるデータをプロンプト上で与えないまま，推論したい対象とタスク指示のみを与える推論方法である．つまり本稿における共起頻度は，レビュアが実際に付与した満足度 y_j とレビューのみから推測される満足度 $\hat{y_j}$ の対応関係をレビュア単位で集計したものである．

我々は評価傾向に内在する個人差を，多数のレビュアが持つ平均的な評価基準からそのレビュアの評価基準がどれほど離れているのかを表す指標だと考える．しかし，レビュアの評価基準が平均的な評価基準からどれほど離れているのかを定量的に解釈するためには，平均的な評価基準を定量的に表現することが必要となる．多数のレビュアにおける平均的な評価基準を定量的に表現する際，我々は全てのレビュアに対して判断基準を推定するための調査を行う必要があると考えた．しかし，実際に同じサービスを受けたレビュア全員に対して調査を行うことは現実的ではない．そこで我々は，事前学習済み LLM を用いて多数のレビュアに共通する平均的な評価傾向を近似的に表現する参照尺度を定義する．事前学習済み LLM は，多様な文脈や表現を含む大規模コーパスに基づいて学習されている．そのため，事前学習済み LLM は一般的知識や典型的な判断傾向に基づく推論を行う能力を有する．このことから我々は，一般的知識や典型的な判断傾向に基づく推論を行う能力を有する事前学習済み LLM が多数のレビュアに共通する平均的な評価傾向を近似的に表現する参照尺度として適切であると考えた．よって本稿において我々は，LLM の持つ一般的知識や典型的な判断傾向を崩さずに判断させるため，事前学習済み LLM による Zero-Shot 推論によって得られる満足度予測値 $\hat{y_j}$ を用いて参照尺度を定義する．

レビュアが実際に付与した満足度 y_j は，レビュアの評価傾向に基づいて尺度空間上に直接射影された満足度である．事前学習済み LLM による Zero-Shot 推論によって得られる満足度予測値 $\hat{y_j}$ は，多数のレビュアに共通する平均的な評価傾向を近似的に表現する参照尺度に基づいて意味空間上に射影された満足度である．つまり，レビュアが実際に付与した満足度 y_j とレビューから推測される満足度 $\hat{y_j}$ の対応関係を集計した混同行列 $C^{(u_i)}$ は，あるレビューに基づいて意味空間上に射影した満足度に対してそのレビューを書いたレビュアがどの満足度を付与しやすいかという共起構造を表現している．このことから我々は，実際に付与された満足度 y_j と Zero-Shot 推論によって推測された満足度 $\hat{y_j}$ の共起頻度を集計することによって，レビュアの評価傾向が多数のレビュアに共通する平均的な評価傾向を近似的に表現する参照尺度からどれほど離れているのかを表現することが可能だと考えた．

4.1.2　リッカート尺度の正規化 (Step 2)

我々は，個人差の内在するリッカート尺度に基づく満足度 y_j とレビューを書いたレビュアの混同行列 $C^{(u_i)}$ を用いて，多数のレビュアに共通する参照尺度に基づく満足度 y'_j を射影する．この際，我々はニューラルネットワークを構築し，多数のレビュアに共通する参照尺度に基づく満足度 y'_j への射影，つまり尺度の正規化へと活用する．

ニューラルネットワークの入力 $\mathbf{x}_j^{(u_i)}$ は，正規化対象の満足度 y_j とそのレビューを書いたレビュアの混同行列 $C^{(u_i)}$ を結合した一次元のベクトル $v(C^{(u_i)})$ である．この時，$\mathbf{x}_j^{(u_i)} = [y_j, v(C^{(u_i)})] \in \mathbb{R}^{\dim(\mathbf{x}_j^{(u_i)})}$ である．入力次元数 $\dim(\mathbf{x}_j^{(u_i)})$ は，$\dim(\mathbf{x}_j^{(u_i)}) = N(y_j) + |C^{(u_i)}|$ である．つまり，$\dim(\mathbf{x}_j^{(u_i)}) = 1 + k^2$ である．ニューラルネットワークの出力 $\hat{y_j}$ は，入力 $\mathbf{x}_j^{(u_i)}$ に基づく意味的な正規化写像である．以上より，本研究のタスクにおけるニューラルネットワークの入出力は以下の式で表される．

$$y'_j = f_\theta\left(\mathbf{x}_j^{(u_i)}\right), \quad f_\theta : \mathbb{R}^{\dim(\mathbf{x}_j^{(u_i)})} \to \mathbb{R} \tag{4.1.1}$$

この時，θ は正規化写像 f_θ を構成するニューラルネットワークの学習可能パラメータである．また，$\hat{y_j} \in Y$ である．本研究のタスクは，回帰タスクとして設計する．分類タスクではなく回帰タスクとした理由は，分類モデルの構造に関係する．分類タスクに用いるモデルは，出力層に評価値カテゴリ集合の数だけノードを持つ．それに伴い，出力は各クラスに属する確率となる．このことから我々は，分類モデルにおける最終的な予測満足度はモデル内部で計算された確率に基づく名義尺度のようなものであると考えた．正規化された後の満足度は，多数のレビュアに共通する平均的な評価傾向を近似的に表現する参照尺度に基づく満足度である．そのため，我々は隣接した満足度評価値を近い満足度評価値だと認識できるよう，順序と距離がない名義尺度でのシステム化は不適切だと考えた．我々は，順序と距離を明示的に扱うことが可能な回帰予測値によって，参照尺度に基づく評価値が順序関係や意味的距離を正確に扱えるようになることを期待する．

学習時の教師信号は，多数のレビュアに共通する平均的な評価傾向を近似的に表現する参照尺度として定義した Zero-Shot 推論による満足度予測結果 $\hat{y_j}$ を利用する．前述した通り，事前学習済み LLM は多様な文脈や表現を含む大規模コーパスに基づいて学習されている．そのため，事前学習済み LLM は一般的知識や典型的な判断傾向に基づく推論を行う能力を有する．これにより，特定のレビュアの評価傾向に依存しない平均的な評価傾向を近似的に表現する参照尺度を教師信号として与えることが可能となる．

次に，ニューラルネットワークを用いた理由を説明する．我々がニューラルネットワークを選択した理由は，二つある．一つ目は，参照尺度への意味的正規化という特殊なタスクへの対応が可能であるためである．自然言語を用いたタスクでは，ニューラルネットワークでは

なく LLM が用いられることも多い．LLM は多様な自然言語タスクに対して高い汎化性能を示すためである．しかし，LLM は本研究で対象とするような評価傾向とリッカート尺度に基づく主観尺度の対応関係からレビューを基盤とした多数のレビュアに共通する参照尺度への正規化を実施することを目的とした設計をなされていない．一方で，ニューラルネットワークは一からパラメータを調整することによって，固有の評価傾向に基づく満足度から多数のレビュアに共通する平均的な参照尺度に基づく満足度への写像を学習することが可能である．そのため，我々はニューラルネットワークならば本研究のタスクに十分対応可能だと考えた．またタスクへの対応力を向上させるために，LLM に対する Fine-Tuning や Few-Shot Prompting という選択肢も存在する．しかし，LLM の Fine-Tuning や Few-Shot Prompting では個々のレビュアが持つ評価傾向をベクトルとして構造的な情報を維持したまま明示的に与えることが難しい．そのため，本研究のタスクに適切ではないと考えた．

二つ目は，線形なモデルで特殊な評価傾向に対応することが困難なためである．本タスクにおいて，LLM 以外に重回帰分析のような線形モデルを用いるという選択肢が存在する．重回帰分析は，各入力特徴が出力に対して独立かつ線形に寄与することを仮定したものである．しかし，本研究で扱う入力は実際に付与された満足度 y_j と，レビュアの評価傾向を表す混同行列 $C^{(u_i)}$ である．つまり，入力と出力の関係は混同行列 $C^{(u_i)}$ による条件付けで変化する．したがって，入力が同一の満足度 y_j であったとしても評価傾向を示す混同行列 $C^{(u_i)}$ が異なる場合は正規化後の満足度 y'_j が変化する可能性がある．この現象は，表 4.2d のような評価傾向を持つレビュアに顕著である．このことから我々は，本研究のタスクが満足度 y_j と評価傾向 $C^{(u_i)}$ の相互作用を考慮した非線形写像であると考える．非線形写像を表現するためには，特徴間の非線形な関係を学習可能なモデルが必要である．よって我々は，非線形な活性化関数を介することによって混同行列 $C^{(u_i)}$ に内在する構造的な情報と満足度 y_j との関係を柔軟に統合できるニューラルネットワークが，本研究のタスクに適していると判断した．

4.2 提案手法の適用範囲

本提案手法は，評価傾向を混同行列 $C^{(u_i)}$ によって表現する．また本提案手法は，事前学習済み LLM の Zero-Shot 推論による満足度予測結果 $\hat{y_j}$ を多数のレビュアに共通する平均的な評価傾向を近似的に表現する参照尺度として定義する．そのため，混同行列 $C^{(u_i)}$ や Zero-Shot 推論による満足度予測結果 $\hat{y_j}$ に関わる満足度予測の入力であるレビュー r_j が手法全体の基盤となる．したがって，本提案手法はレビュー r_j の影響を大きく受けてしまう．そのため我々は，本提案手法を用いる上で以下の三つの条件が満たされる必要があると考えた．

一つ目は，レビュー r_j が評価値 y_j と対になっていることである．EC サイトの満足度評価値において，評価理由を示すレビュー r_j が存在しておらず，評価値 y_j が単体で投稿されているケースは確かに存在する．しかし本提案手法において我々は，評価値 y_j は評価者が持つ評価理由を表すレビュー r_j に基づいて付与された結果であるという立場に立つ．そのため，レビュー r_j が評価値 y_j と対になっていないケースは本提案手法の適用範囲外とする．

二つ目は，レビュー r_j が平均的な判断傾向のもとで解釈可能であることである．レビュー r_j が平均的な判断傾向のもとで解釈不可能である場合，レビュー r_j を基に付与された満足度 y_j の意味解釈における違いを個人差という範囲で扱うことは難しい．そのため本研究において我々は，クレーマーを代表とした平均的な判断枠組みのもとで解釈することが困難な難癖をつけるようなレビューや，満足度評価を実施しているレビュアを本提案手法の適用範囲外とする．具体的には，平均的に些細とされる程度の欠点を過剰に批判するレビュアや，個人的な信念や考えに強く依存するレビューを書くレビュアを指す．また，表 4.2c で示すようにレビュー数が極端に少ないレビュアは評価傾向の解釈が不安定になるため，本提案手法の適用範囲外とする．手法の適用範囲外となる具体的なレビュー数は，5.1.3 節で調査する．

三つ目は，Zero-Shot 推論が妥当な評価を返すことである．Zero-Shot 推論を行う LLM が何らかの追加学習によって偏りを持ってしまった場合，その LLM は一般的知識や典型的な判断傾向に基づく推論を行わない．つまり，偏った LLM による Zero-Shot 推論結果が参照尺度となるため，偏った LLM による Zero-Shot 推論による参照尺度に基づいて正規化が実行される．その場合，満足度 y_j は多数のレビュアに共通する平均的な評価傾向と異なる意味空間上に射影されることとなる．それは主観評価における意味解釈の個人差を解消するという本研究の目的と合致しない．よって本研究の参照尺度形成において，我々は追加学習や参考情報の添付を実施していない事前学習済み LLM が Zero-Shot 推論によって妥当な評価を返す場合のみを本提案手法の適用範囲とする．本稿において我々は，追加学習や参考情報の添付を実施していない事前学習済み LLM が Zero-Shot 推論によって妥当な評価を返すという仮定の下，以降の実験を実施する．

第 5 章　実験

我々は，三つの実験を実施した．実験概要は以下の通りである．

実験 1： 例外的なレビューケースと評価傾向に関する基礎検証
実験 2： 混同行列を用いた評価傾向定義の妥当性検証
実験 3： 提案手法の有効性検証

実験 1 の目的は，提案手法の基盤となる例外的なレビューケースと評価傾向に関する仮説を検証することである．この実験により，例外的なレビューケースと評価傾向が示す個人差構造が実データ上に存在することを示す．我々は，三つの仮説を立てた．

一つ目の仮説は，提案手法の核となる例外的なレビューケースや個人差は実データ上に存在する現象であるというものである．本提案手法において我々は，例外的なレビューケースの存在に着目してレビュアの評価傾向を表現する．つまり，実際のレビューデータにおいて例外的なレビューケースが存在していることが前提となる．したがって，我々は例外的なレビューケースに該当する基準を明確に定義する必要があると考えた．また，我々は例外的なレビューケースが一部の特殊事例ではなく，無視できない頻度で存在する現象であるかを確認する必要があると考えた．

二つ目の仮説は，例外的なレビューケースはレビュー投稿数に関係しているというものである．例外的なレビューケースが一定数存在する場合，次に問題となるのはデータ数と出現率の関係性である．投稿数が極端に少ないレビュアに例外的なレビューケースが集中している場合，我々はレビュアごとの評価傾向を安定的に推定することが困難となると考えた．したがって，我々は例外的なレビューケースが不規則に発生するものなのか，あるいはレビュー投稿数に応じて発生が偏るものなのかを明らかにする必要があると考えた．

三つ目の仮説は，レビュアの評価傾向が例外的なレビューケースと関係しているというものである．本提案手法において，我々はレビュアの評価傾向をスカラ値ではなく混同行列を用いて表現する．混同行列を用いた評価傾向の表現は，例外的なレビューケースの存在を構造的な情報として保持する．これは，スカラ値のような少数パラメータを用いた評価傾向の表現を行う既存手法との差別点である．我々は既存手法と異なり，例外的なレビューケースを単なるノイズとして扱うべき対象ではないと考えた．したがって，我々は例外的なレビューケースによる評価値への影響がスカラ値で表せるような一様なものではないことを示す必要があると考えた．

実験 2 の目的は，評価傾向を混同行列で定義することが妥当であるという仮説を検証する

ことである．混同行列による評価傾向の定義は，尺度正規化の際も重要な要素となる．したがって，我々はこの評価傾向の定義が妥当であるかどうか実験的に検証する必要があると考えた．この実験により，実験 1 で示された評価傾向が示す個人差構造を混同行列で捉えることが可能であることを示す．この実験は，実験 1 と比較して複雑な実験である．そのため，5.2.1 節で概要を説明する．

実験 3 の目的は，提案手法が主観評価に内在する個人差の解消に有効であるという仮説を検証することである．この実験では，実験 2 で示された混同行列による評価傾向表現を用いることによって主観評価に内在する個人差の補正とそれに伴う意味空間上での正規化が可能であるかを検証・考察する．

5.1 実験 1: 例外的なレビューケースと評価傾向に関する基礎検証

本実験において我々は，以下の三つの分析を実施する．

1. 例外的なレビューケースに該当する基準の調査及び出現率の確認
2. 例外的なレビューケース出現率とレビュー投稿数の関係性調査
3. レビュアの評価傾向と例外的なレビューケースの関係性調査

一つ目は，例外的なレビューケースに該当する基準の定義及び出現率に対する分析である．我々は，提案手法の核となる例外的なレビューケースや個人差は実データ上に存在する現象であるという仮説を立てた．この仮説の検証により，後続の分析における例外的なレビューケースや個人差といった前提条件を明確化する．また 4.2 節で述べたように，手法の適用範囲外とする具体的なレビュー数帯を明確化する．

二つ目は，例外的なレビューケース出現率とレビュー投稿数の関係性に対する分析である．我々は，例外的なレビューケースはレビュー投稿数に関係しているという仮説を立てた．この仮説の検証により，後続の実験における実験対象レビュアをどのデータ数帯に設定するのか決定する．

三つ目は，レビュアの評価傾向と例外的なレビューケースの関係性に対する分析である．我々は，レビュアの評価傾向が例外的なレビューケースと関係しているという仮説を立てた．この仮説の検証により，既存手法のように例外的なレビューケースをノイズとして平均化することが問題であることを示す．

5.1.1 実験手順

実験手順は以下の通りである．

1. 例外的なレビューケースに該当する基準の調査及び出現率の確認：

   Step 1 レビューデータ全体を抽出
   Step 2 全レビューデータの分布に関する調査
   Step 3 レビューデータの準備・前処理
   Step 4 例外的なレビューケースの基準設定と目視確認
   Step 5 Step 3 で設定された基準を基に例外的なレビューケースの出現率を計算

2. 例外的なレビューケース出現率とレビュー投稿数の関係性調査：

   Step 1 各データ数帯における例外的なレビューケースを持つレビュアのみを抽出
   Step 2 レビューデータの準備・前処理
   Step 3 各データ数帯における例外的なレビューケース出現率が $> 0\%$ の時と $\geq 5\%$ の時の関係性を確認

3. レビュアの評価傾向と例外的なレビューケースの関係性調査：

   Step 1 指定したデータ数帯における評価傾向特性の異なるレビュアを 1 人ずつ抽出
   Step 2 例外的なレビューケースに該当するデータとそうでないデータをそれぞれ用意
   Step 3 レビューデータの準備・前処理
   Step 4 Step 2 で用意した 2 種類のデータ割合を変更しつつ複数パターンの参考データを構築
   Step 5 Step 4 で構築した複数パターンの参考データを利用して満足度予測タスクを実施
   Step 5 評価傾向特性の異なるレビュアごとの予測精度を比較

一つ目の分析手順を整理する．Step 1 ではレビューデータ全体を抽出する．詳細なデータ数は 5.1.2 節で述べる．Step 2 ではレビューデータ全体の分布に関する調査を実施する．この調査の目的は，実データにおいてどれだけのレビュアがどの程度のレビュー数を投稿しているかを把握することである．我々は，実際の EC サイトにおけるレビュアにレビュー投稿数の偏りがあると考えた．レビュー投稿数が少ないレビュアが多い場合，我々は本提案手法における手法適用範囲としてどの程度のレビュー投稿数を持つレビュアが妥当であるのか考察する必要があると考えた．その考察にしたがって，以降の実験で用いるデータ数帯を設定す

る．Step 3 では本分析に用いるレビューデータの準備と前処理を行う．レビューデータは設定したデータ数帯ごとで扱う．前処理については全実験で共通であるため，5.1.2 節で述べる．Step 4 では例外的なレビューケースの基準設定と目視確認を行う．例外的なレビューケースは 4 節で述べたとおり，レビュアが実際に付与した満足度とレビューから推測される満足度が大きく異なるレビューのことを指す．レビュアが実際に付与した満足度とレビューから推測される満足度のズレが 1 の時，2 の時，3 の時，4 の時をそれぞれ分析する．Step 5 では，設定された基準を基に例外的なレビューケースの出現率を調査する．この結果が，主観評価に内在する個人差の存在を示す根拠となる．

二つ目の分析手順を整理する．Step 1 では一つ目の分析で決定したデータ数帯それぞれにおいて，例外的なレビューケースを持つレビュアを抽出する．Step 2 ではレビューデータの準備と前処理を行う．Step 3 では出現率が $> 0\%$ のグラフと $\geq 5\%$ のグラフをそれぞれ三種類作成する．$> 0\%$ のグラフは，Step 1 で抽出した例外的なレビューケースを持つレビュア全てを対象とする．$\geq 5\%$ のグラフは，Step 1 で抽出した例外的なレビューケースを持つレビュアの内，例外的なレビューケースが 5% 以上出現するレビュアを対象とする．5% の閾値は，一つ目の分析結果を踏まえ，例外的なレビューケースの偶発的な出現を考察時に考慮するため設定した．

三つ目の分析手順を整理する．Step 1 では二つ目の分析で決定したデータ数帯それぞれにおいて，評価傾向特性の異なるレビュアを抽出する．本分析における評価傾向特性の異なるレビュアは，混同行列の要素が対角線上に集合したレビュアと混同行列の要素がばらけたレビュアを指す．2 種類のレビュアをデータ数帯ごとに一人ずつ抽出する．Step 2 では例外的なレビューケースに該当するデータとそうでないデータをそれぞれ用意する．Step 3 ではレビューデータの準備・前処理を行う．Step 4 では Step 2 で用意した 2 種類のデータを参考データとした満足度予測タスクを実施する．参考データ数は 5.1.2 節で述べる．Step 5 では Step 4 で構築した複数パターンの参考データを利用して満足度予測タスクを実施する．例外的なレビューケース割合は推移関係を確認するため 0.0 ,0.25 ,0.50 ,0.75 ,1.0 の 5 段階とする．Step 6 では 評価傾向特性の異なるレビュアごとの予測精度を比較する．我々は，例外的なレビューケースの割合という条件のみを変更した上で評価傾向特性の異なるレビュアの予測精度を比較することによって，例外的なレビューケースによる満足度予測の精度への影響が一様でないことを示すことが可能だと考えた．

5.1.2 実験準備

本実験において我々は，実際のレビューデータに対する三種類のデータ分析を実施する．使用するデータセットは楽天トラベルデータセット*である．一つ目の分析は，2,044,166 人のレビュアが書いた 6,560,157 件のレビューを分析対象とする．二つ目の分析は，一つ目の分析結果を踏まえて設定した 7 種類のデータ数帯において，ランダムに抽出した各 50 人ずつのレビュアを分析対象とする．データ抽出にはランダムサンプリングを実施する．詳細なデータ数分布は付録 A.1 で示す．また，定義した 7 種類の各データ数帯において 50 人ずつのデータが収集できなかった場合は，±1 のデータ数を持つレビュアから人数を補完する．三つ目の分析は，二つ目の分析結果を踏まえて設定した 4 種類の実験対象データ数帯ごとに，評価傾向を示す混同行列の要素が対角線上に集合しているレビュアとばらけているレビュアをそれぞれ一人ずつ抽出し，分析対象とする．また三つ目の分析における参考データは，例外的なレビュケースに該当するデータとそうでないデータの中から設定した割合に従ってランダムサンプリングを実施する．参考データのサンプリングに用いるシード値は，全推論で異なる．つまり，推論対象のレビュー数 ×5×5 個のシード値を利用する．参考データ数は合計 50 データとなるよう設定した．前処理では，欠損しているデータを除外する．使用する LLM は，Gemma3:27B†である．量子化は行わない．三つ目の分析における目的はレビュアの評価傾向と例外的なレビューケースの関係性調査であるが，両者の関係性はレビュア単位での例外的なレビューケース割合を変更した予測精度推移によって間接的に評価可能である．精度の評価指標は Accuracy と RMSELoss を用いる．RMSELoss は，LLM の予測値が実際の満足度評価値から大きく外れている場合とそうでない場合を反映させるために用いる．統計的検定として対応のある 2 標本 t 検定を実施した．

5.1.3 結果・考察

例外的なレビューケースに該当する基準の調査及び出現率の確認

実データ全体に対し，レビュア単位のレビュー投稿数分布を集約した結果を図 5.1 に示す．図 5.1 は使用した実データ全体におけるレビュアの投稿数分布を示している．図 5.1 に示す通り，実データ全体においてレビューの投稿数には大きな偏りがある．一部のレビュアは 10 件以上のレビューを投稿しているものの，ほとんどのレビュアは 1 から 5 件しか投稿していない．そのため，楽天トラベルデータセットに対して手法を適用する際は大多数を占める 1 から 5 件投稿しているレビュアを対象とすることが理想である．しかし，レビュー投稿数が 10 件未満のレビュアは評価傾向の表現が困難である．具体例を表 5.1 に示す．表 5.1a は，

*楽天グループ株式会社 (2014): 楽天データセット．国立情報学研究所情報学研究データリポジトリ．(データセット)．https://doi.org/10.32130/idr.2.0

†`https://ollama.com/library/gemma3:27b`

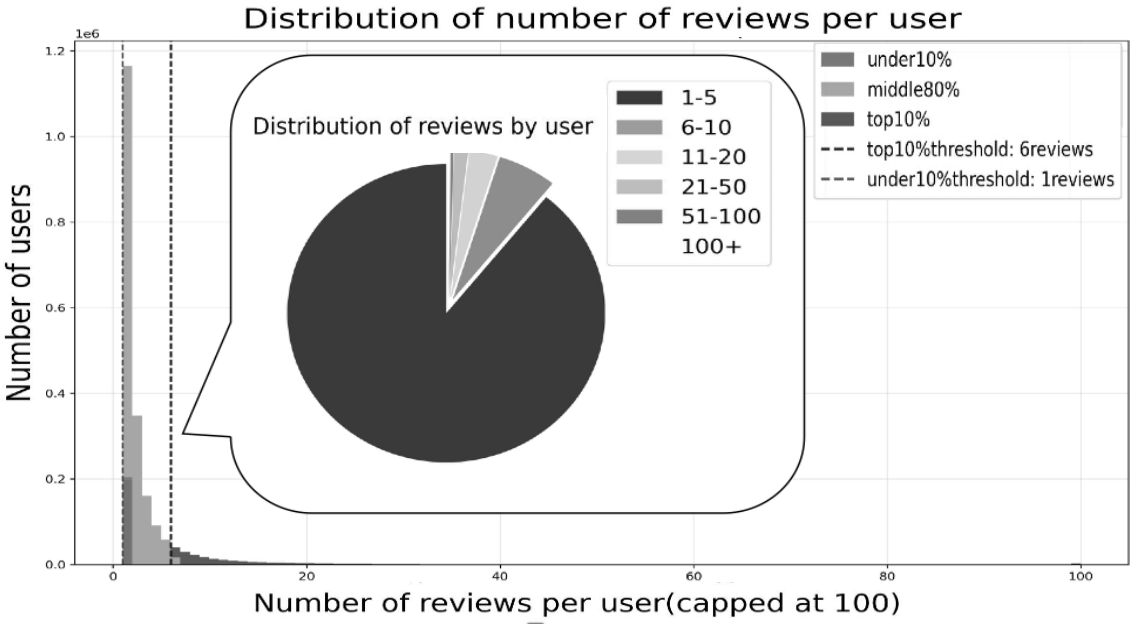

図 5.1: 実データ全体におけるレビュアの投稿数分布

レビュー投稿数が 5 件のレビュアにおける混同行列を示している．表 5.1b は，レビュー投稿数が 1 件のレビュアにおける混同行列を示している．これらの混同行列は極端に要素数が少ないため，限られたレビューに依存した偶発的な満足度付与が反映されている可能性がある．このことより我々は，レビュー投稿数が極端に少ないレビュアの混同行列で評価傾向を網羅的に反映することが困難であると考えた．一方，レビュー投稿数が 10 件以上のレビュアにおける混同行列の具体例を表 5.2 に示す．表 5.2a，表 5.2b は，レビュー投稿数が 10 件のレビュアにおける混同行列を示している．レビュー投稿数が 10 件以上のレビュアにおける混同行列は，各要素の大きさに明確な差が存在している傾向にある．加えて，要素の位置が分散している傾向にある．このことから，レビュー投稿数が 10 件未満のレビュアと比較

表 5.1: 投稿数が極端に少ないレビュアの混同行列

(a) user_559162(投稿数 5)

$\hat{y}\backslash y$	1	2	3	4	5
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	1	1
4	0	0	0	1	1
5	0	0	0	0	1

(b) user_1700335(投稿数 1)

$\hat{y}\backslash y$	1	2	3	4	5
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	1

して，限られたレビューに依存した偶発的な満足度付与が反映されている可能性は比較的少ないと考えられる．以上のことより，我々は本稿における手法適用範囲を投稿数が 10 件以上のレビュアと定義する．手法適用範囲の設定に基づき，我々は今後の分析に用いるデータ数帯を{10,20,30,40,50,100,> 100}の 7 種類で定義する．図 5.1 より，レビュー投稿数が多いレビュアは最低限とした．

次に，例外的なレビューケースの基準を定義する．例外的なレビューケースを定義するため，我々は実データ全体における実際の満足度と推測された満足度の差を調査した．結果を表 5.3 に示す．また，1 から 4 の予測誤差が発生したレビューの具体例を表 5.4，表 5.5，表 5.6，表 5.7 に示す．実際の満足度とレビューから推測される満足度における差が 1 である

表 5.2: 投稿数 10 以上のレビュアが持つ混同行列

(a) user_190911(投稿数 10)

$\hat{y}\backslash y$	1	2	3	4	5
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	1	0	0
4	0	0	0	4	1
5	0	0	0	3	1

(b) user_353297(投稿数 10)

$\hat{y}\backslash y$	1	2	3	4	5
1	0	0	0	0	0
2	0	0	1	0	0
3	0	0	1	1	0
4	0	0	1	1	3
5	0	0	0	0	2

表 5.3: 実際の満足度とレビューから推測される満足度における差の分布

差 (diff)	0.0	1.0	2.0	3.0	4.0
データ数	22,411	16,890	1,286	118	9
割合 (%)	55.0	41.5	3.2	0.3	0.02

表 5.4: 差が 1.0 のレビューにおける具体例

レビュー	実際の満足度	推測された満足度	差
フロントの対応が良く、チェックアウトもスムーズでした。総合的に良いホテルの感じです。	5.0	4	1.0
従業員の愛想が悪い。	3.0	2	1.0
ホテルは全体的に綺麗。	3.0	4	1.0

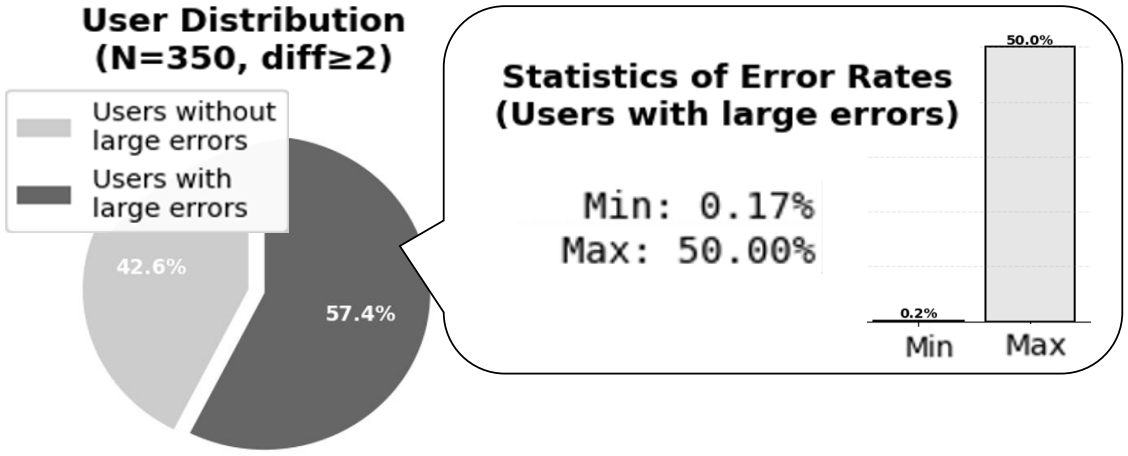

図 5.2: 例外的なレビューケースを持つレビュアの割合

データは，全体の 41.5% を占める．また，表 5.4 にも示す通り，レビューと実際の満足度の間に大きな乖離は見られない．そのため，予測誤差が 1 のレビューは例外的なレビューケースに該当しない．次に予測誤差が 2 であるレビューについて考察する．実際の満足度とレビューから推測される満足度における差が 2 であるデータは，全体の 3.2% である．加えて表 5.5 にも示す通り，レビュー内の言及と実際の満足度の間に大きな乖離が見られる．この例における予測誤差が 2 のレビューは，ネガティブなことを言及しているのにも拘らずポジティブな満足度が付与されているものが複数存在している．また三つ目のレビューのように，一部を除いて満足だという言及があるもののその一部が大きな欠点であるケースなど，文面だけでは満足度が付与しにくいレビューも存在した．表 5.6，表 5.7 でも同様のケースが見て取れる．以上の結果より，我々は本稿における例外的なレビューケースを実際の満足度と推測された満足度の差が 2 以上あるものと定義する．

最後に，定義した例外的なレビューケースが実データにどれほど出現するのか調査する．結果を図 5.2 と図 5.3 に示す．図 5.2 は，本実験で使用した計 350 レビュアのレビューデータにおいて，例外的なレビューケースを持つレビュアがどれだけ出現するのかを示したものである．図 5.3 は，例外的なレビューケースの出現率に応じたレビュア数の分布を示したものである．図 5.2 に示す通り，350 人のレビュアのうち 57.4% のレビュアが例外的なレビューケースを持つ．この結果より，例外的なレビューケースは一部の特殊事例ではないということが分かる．また図 5.3 に示す通り，350 人のレビュアのレビュー内には平均 4%，中央値 5.93% ほど例外的なレビューケースが存在する．つまり，レビュアの評価傾向を表現する情

報のうち，平均 4% ほどが例外的なレビューケースに関連する情報であることがわかるまた，例外的なレビューケースの割合が最も高いレビュアは，50% もの例外的なレビューケースを持つ．つまり，このレビュアの評価傾向を表現する情報のうち，50% が例外的なレビュケースに関連している．したがって，例外的なレビューケースは主観評価の個人差補正に用いる評価傾向を正確に表現する際に無視するべきではない情報であることが分かる．

以上の結果より，例外的なレビューケースは想像上の概念ではなく実データ上に存在する現象であるという仮説は正しいと言える．また以降の実験における前提として，例外的なレビューケースの存在は主観評価に内在する個人差の存在を明確に示していると言えることが明らかとなった．

例外的なレビューケース出現率とレビュー投稿数の関係性調査

実験の結果を図 5.4，図 5.5，図 5.6 に示す．図 5.4 は，各データ数帯に属する 50 人のレビュアが例外的なレビューケースをどの程度保有しているのかを示したものである．図 5.5 は，各データ数帯における例外的なレビューケースの合計数を示したものである．図 5.6 は，各データ数帯における 50 人のレビュアの内，何人が例外的なレビューケースを持つか示した

表 5.5: 差が 2.0 のレビューにおける具体例

レビュー	実際の満足度	推測された満足度	差
テナントの弁当屋が撤退していたのが、非常に残念です。	4.0	2	2.0
改装中とはいえ、廊下どころか客室内でも溶剤の異臭がキツかったのが残念でした。	4.0	2	2.0
ある点を除けは非常に高いレベルで満足できました。それは 520 号室ですが、ウォシュレットが「時々」故障する。始末が悪い。風呂のシャワーの切り替えが甘く、強弱はありますが常に出っ放しの状態で、打たせ湯です。520 号室は避けましょう。	1.0	3	2.0

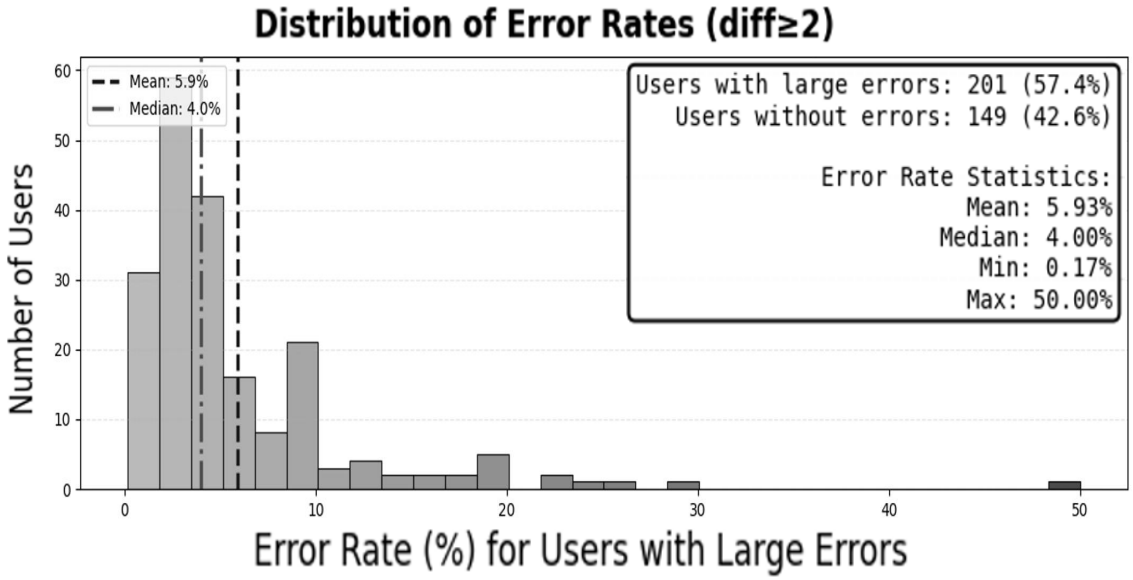

図 5.3: 例外的なレビューケースの出現率に応じたレビュア数分布

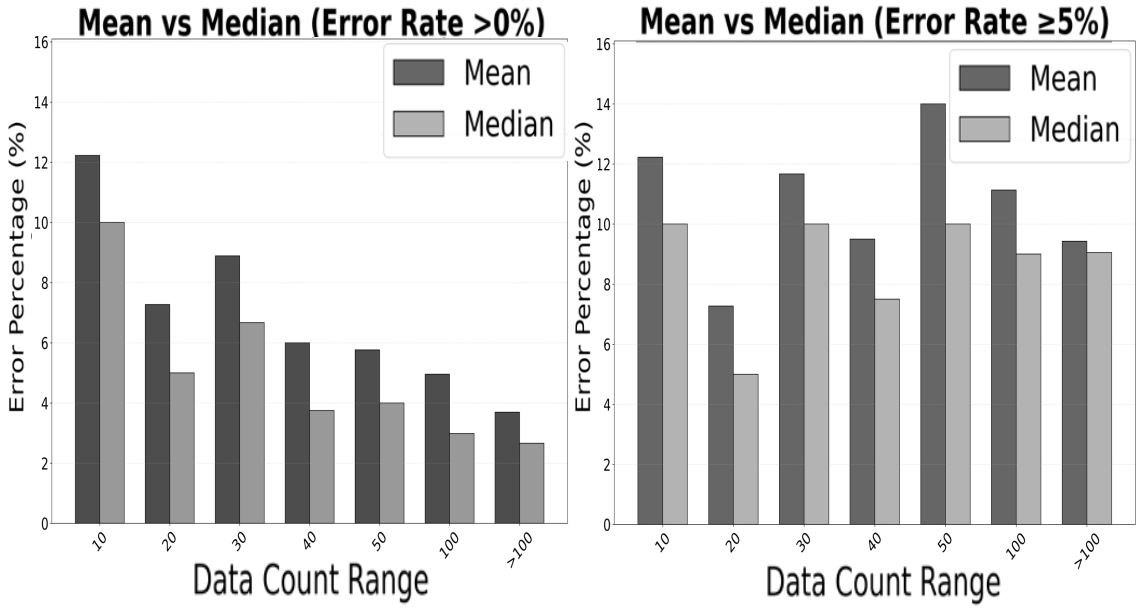

図 5.4: 各データ数帯における例外的なレビューケースの保有率

ものである．左のグラフは，例外的なレビューケースを持つレビュア全体に関連した結果を示している．つまり，図 5.6 の左グラフは図 5.2 を各データ数帯に細分化したものである．右のグラフは，例外的なレビューケースを 5% 以上持つレビュアに関連した結果を示している．

図 5.4 における $> 0\%$ のグラフが示す通り，データ数帯が少ないレビュアほど例外的なレビューケースの出現率は高い値を示した．特にデータ数帯が 10 のレビュアは，平均値・中央値ともに 10% 以上である．他のデータ数帯に関しては，データ数帯が大きくなるにつれて例外的なレビューケースの出現率が減少している．この結果は，データ数が少ないために一つ一つのデータがもたらす影響が大きくなることを示している．一方，$\geq 5\%$ のグラフでは，データ数帯が 10,20 であるレビュアを除いて例外的なレビューケースの出現率が大きく増加した．この結果は，例外的なレビューケースの出現率が 5% に満たないレビュアが大きく減ったことに起因する．この結果は，図 5.6 からも読み取れる．したがってこの結果は，データ数帯が小さいレビュアほど例外的なレビューケースが比較的多く出現することを示す．

次は例外的なレビューケースの出現数に着目する．図 5.5 に示す通り，データ数帯が大きいほど例外的レビューケースの総量は増加している．しかし前述した通り，図 5.4 においてはデータ数帯が大きいレビュアほど例外的なレビューケースの出現率が減少している．この結

表 5.6: 差が 3.0 のレビューにおける具体例

レビュー	実際の満足度	推測された満足度	差
2 人での宿泊でした。カードキーを 2 枚用意しないのはなぜでしょうか？	5.0	2	3.0
隣のひとの目覚まし時計が鳴り止まず・・・	4.0	1	3.0
駅の近くにあるが、雨が降ったら傘が要る距離です。部屋は広くて気持ちよい。デスクスペースも十分。ベッドは柔らかすぎず気持ちよい。風呂も容量十分。ただひとつエアコンは今風ではなく古い。弱いというか頼りない。今のシーズンは不便は感じないが、エアコンをどうにかしないと寒暖厳しい時には泊まりたくない。	1.0	4	3.0

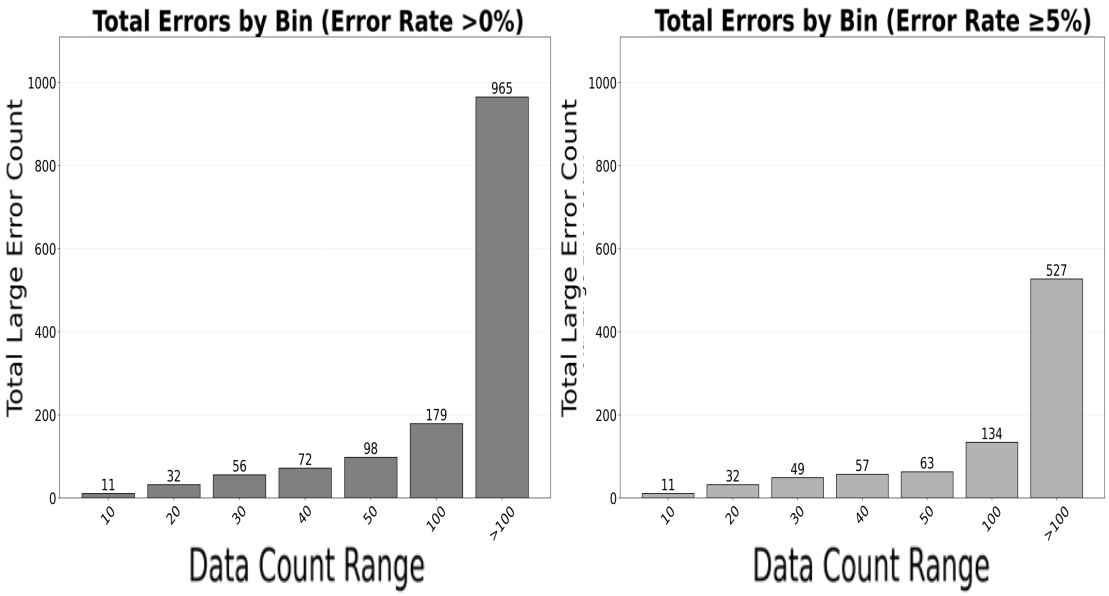

図 5.5: 各データ数帯における例外的なレビューケース数

果は，データ数が非常に多いために例外的なレビューケースの出現率が低くとも絶対数としてはデータ数が多くなることに起因する．つまり，データ数帯が大きいレビュアは例外的なレビューケース以外の要素が大部分を占めているため，評価傾向を示す混同行列において例外的なレビューケースに該当する要素が構造として現れにくいことを示唆している．

本分析では実験結果より，大きく二つの考察が得られた．一つ目は，データ数帯が小さいレビュアほど，データ一つに対する依存度が大きくなるということである．この結果はデータ数帯が小さいレビュアほど，例外的なレビュケースの偶発的な発生によって評価傾向の推定が不安定になることを示唆している．二つ目は，データ数帯が小さいレビュアほど，例外的なレビューケースが明確に現れるということである．そのためデータ数帯が小さいレビュアほど，例外的なレビューケースに該当する要素が混同行列上の構造として現れやすいと言え

表 5.7: 差が 4.0 のレビューにおける具体例

レビュー	実際の満足度	推測された満足度	差
海外から帰国しますとこのホテルは最高にホットします。	5.0	1	4.0
文句内で賞を授与します。	5.0	1	4.0
早くお風呂に入りたいです。	5.0	1	4.0

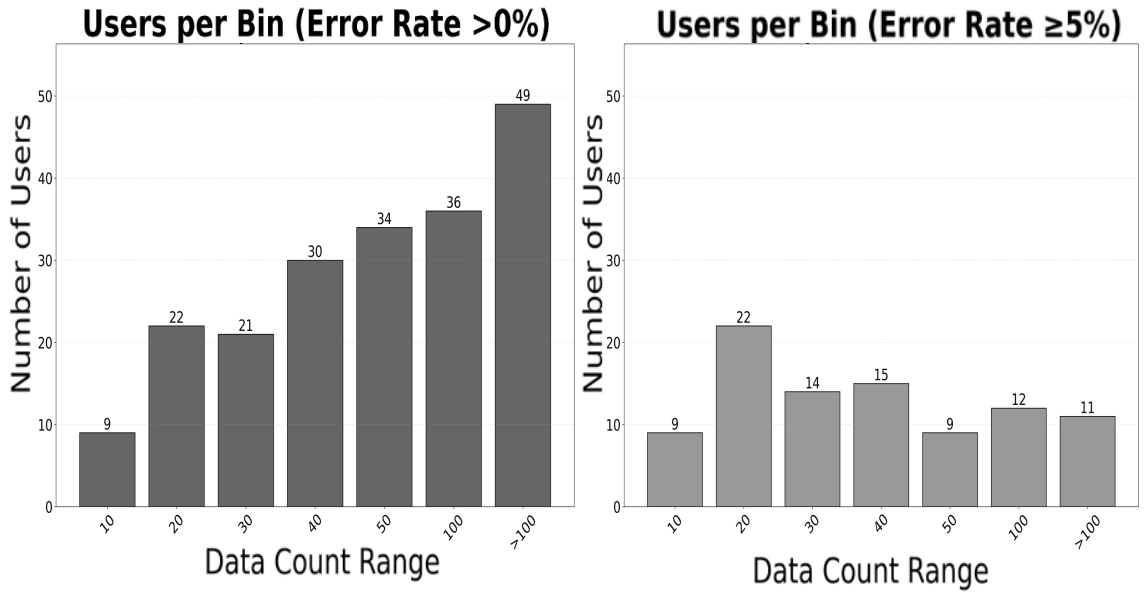

図 5.6: 各データ数帯における例外的なレビューケースを持つレビュア数

る．以上より，例外的なレビューケースはレビュー投稿数に関係しているという仮説は正しいと言える．

我々はこれらの結果と考察を踏まえ，以降の実験においてデータ数帯が 20，30，40 のレビュアを用いる．また，比較対象として評価傾向の構造において例外的なレビューケースが平均化されるであろうデータ数帯が 100 のレビュアも用いる．

レビュアの評価傾向と例外的なレビューケースの関係性調査

実験結果を図 5.7 と図 5.8 に示す．図 5.7 は，設定した例外的なレビューケース割合に基づいて作成した各参考データを利用した際の満足度予測における全レビュアの平均 Accuracy 推移をグラフ化したものである．図 5.8 は，設定した例外的なレビューケース割合に基づいて作成した各参考データを利用した際の満足度予測における全レビュアの平均 RMSELoss 推移をグラフ化したものである．グラフ内の実線は，評価傾向を示す混同行列の要素が対角線上に集合しているレビュアを対象とした際の平均精度である．グラフ内の破線は，評価傾向を示す混同行列の要素がばらけているレビュアを対象とした際の平均精度である．

実験結果より，混同行列の要素が対角線上に集合しているレビュアと混同行列の要素がばらけているレビュアで Accuracy と RMSELoss 両精度指標における増減の仕方が異なることが確認できる．混同行列の要素が対角線上に集合しているレビュア，つまり多数のレビュアが持つ平均的な判断基準に基づいた評価傾向を持つレビュアは，例外的なレビューケースの

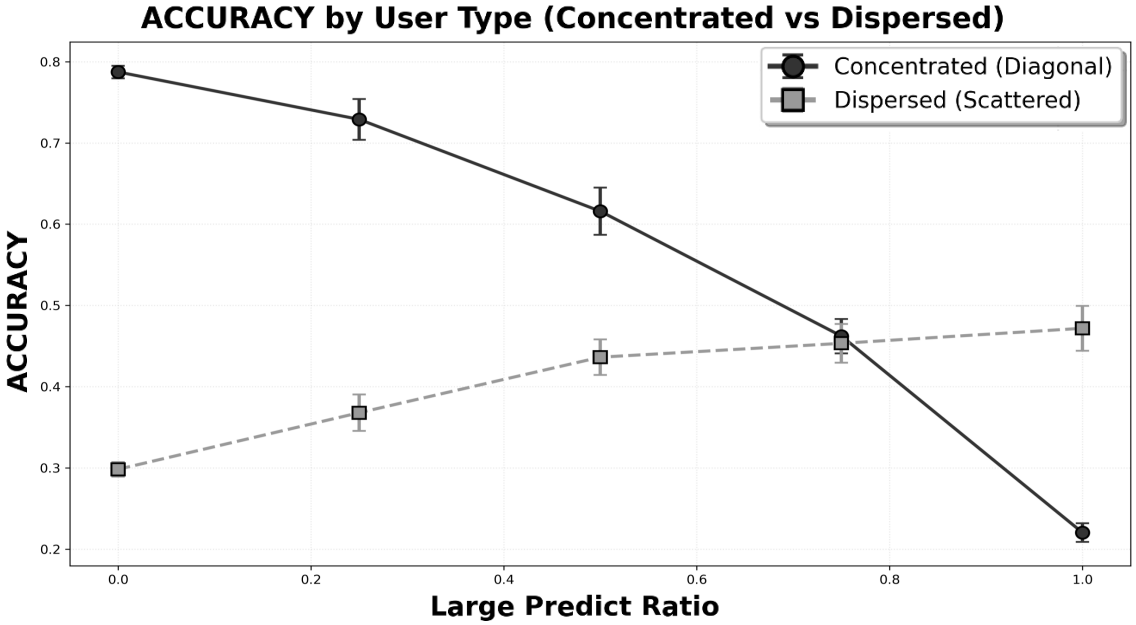

図 5.7: 例外的なレビュケース割合の変化に伴う評価傾向タイプ別平均精度推移 (Accuracy)

割合を増加させるにつれて Accuracy が低下していく．その一方で，RMSELoss は例外的なレビューケースの割合を増加させるにつれて増加していく．これらの結果は，多数のレビュアが持つ平均的な判断基準に基づいた評価傾向を持つレビュアのデータを推論対象とした場合において，例外的なレビューケースが判断のノイズになることを示唆している．我々は例外的なレビュケースがノイズとして扱われてしまう要因を，多数のレビュアが持つ平均的な判断基準に基づいた評価傾向のレビュアが例外的なレビューケースを多く持たないためだと考える．また，評価傾向を示す混同行列要素がばらけているレビュア，つまり特殊な評価傾向に基づき個人差が内在する主観評価を行うレビュアは，例外的なレビューケースの割合を増加させるにつれて Accuracy が向上していく．その一方で，RMSELoss は例外的なレビューケースの割合を増加させるにつれて減少していく．これらの結果は，特殊な評価傾向に基づき個人差が内在する主観評価を行うレビュアのデータを推論対象とした場合において，例外的なレビューケースが判断の有益な参考例となることを示唆している．我々はこの要因を，特殊な評価傾向に基づき個人差が内在する主観評価を行うレビュアが例外的なレビューケースを多く持つためだと考える．以上の結果より，レビュアの評価傾向が例外的なレビューケースと関係しているという仮説は正しいと言える．

なお，詳細な精度や割合変化前後の精度変化に対する p 値については付録 B.2 と付録 B.4 に示す．また，本分析に用いたレビュアの混同行列は付録 B.2 と付録 B.3 に示す．

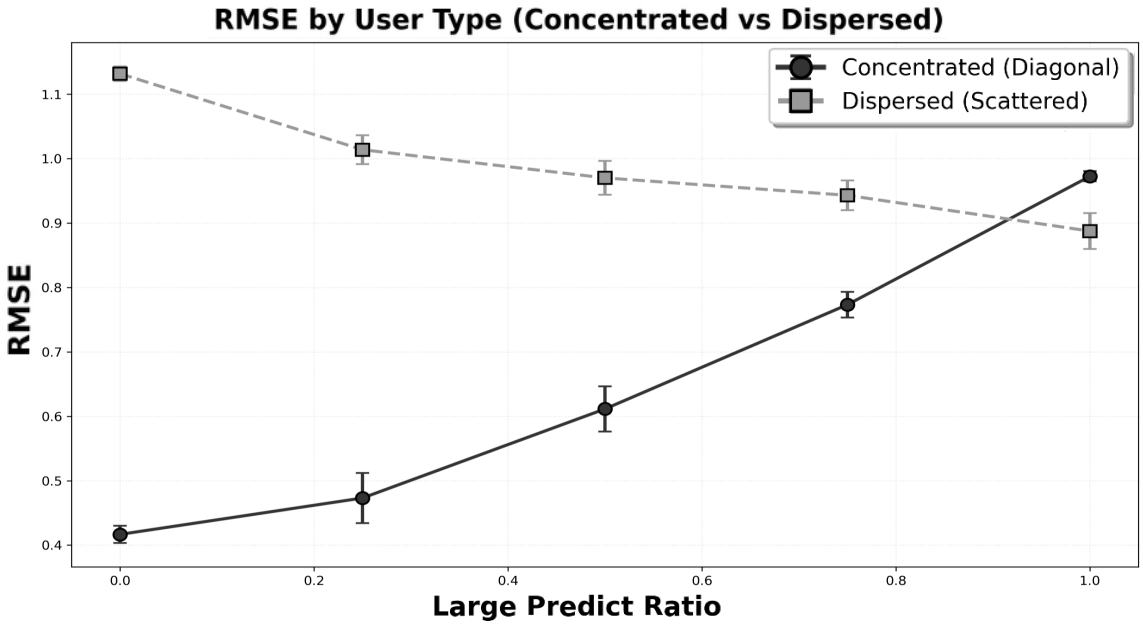

図 5.8: 例外的なレビュケース割合の変化に伴う評価傾向タイプ別平均精度推移 (RMSELoss)

5.2 実験 2: 混同行列を用いた評価傾向定義の妥当性検証

本実験において我々は，評価傾向を混同行列で定義することが妥当であるという仮説に対する検証実験を実施する．我々は，評価傾向を用いた参考データ選択の条件を変更し，条件ごとの満足度予測タスク精度を比較する．実験概要を 5.2.1 節で示す.

5.2.1 実験概要

レビューに付与される満足度は，主観評価である．そのため，例え同じ内容のレビューであってもレビュアごとの評価傾向に依存した個人差が生じる．評価傾向に依存した個人差が生じるということは，すなわち同一人物が書いたレビューでは同一の評価傾向に基づいて同一の満足度が付与されるということである．この考えに基づくと，評価傾向が類似した人物の書いたレビューは類似した評価傾向に基づいて類似した満足度が付与される．そこで我々は，LLM に対して与える参考データとして推論対象レビュア u_t と評価傾向が類似しているレビュアを用いるほど満足度予測精度が向上するという仮説を立てた．我々は，精度の比較結果を基にこの仮説を検証する．仮説検証により，我々は混同行列 $C^{(u_i)}$ による評価傾向の定義がレビュアの評価傾向を適切に捉えていることを示す.

実験手順

実験手順は以下の通りである．

Step 1　指定したデータ数帯におけるレビュアをランダムに 3 人ずつ抽出
Step 2　参考データ選択方針に基づいて参考データを選択
Step 3　レビューと参考データを LLM へ入力し満足度予測

Step 1 では実験 1 の二つ目の分析で決定したデータ数帯それぞれにおいて，レビュアをランダムに 3 人ずつ抽出する．Step 2 では 4 種類の参考データ選択方針に基づいて参考データ $D_s(U_s)$ を選択する．$D_s(U_s)$ は参考データとしてサンプリングされた複数のレビュアが持つデータ集合を示す．4 種類の参考データ選択方針については後述する．Step 3 では推論対象のレビュー r_j^t と Step 2 で選択された参考データ $D_s(U_s)$ を入力とした Few-Shot Prompting により満足度予測を行う．Few-Shot Prompting で使用するプロンプトの一例を図 5.9 として示す．Few-Shot Prompting を用いた理由は，参考データとして用いられたレビュアの評価傾向に基づいて予測値を出力するためである．LLM を任意のタスクに対応させる際，Fine-Tuning と呼ばれる学習法が多く用いられる．この方法は全体の平均的な損失を最小化することによって学習を進める．そのため我々は，全体的な損失最適化方針に沿わない例外的なレビューケースに LLM がうまく対応できなくなってしまうと考えた．一方で，Few-Shot Prompting では提示された例を判断の参考とする条件付き推論である．したがって LLM は，Few-Shot Prompting を用いることによって与えられた参考データに対応した柔軟な回答が可能となる．

参考データ選択方針

本実験において我々は，4 種類の参考データ選択方針を提案する．そしてその後，4 種類の参考データ選択方針に従う計 14 種類の参考データ選択手法を実行する．その後．それぞれの参考データ選択方針に従って選択したレビュアのデータを参考とした Few-Shot Prompting による満足度予測を実施し，その精度を比較する．14 種類の参考データ選択手法についての詳細は，本節で後述する．4 種類の参考データ選択方針とは以下の通りである．

1. 推論対象レビュアと評価傾向が同一のレビュアを利用
2. 推論対象レビュアと評価傾向が類似したレビュアを利用
3. 評価傾向に関係なくランダムにサンプリングしたレビュアを利用
4. 推論対象レビュアと評価傾向が類似していないレビュアを利用

まずは，推論対象レビュアと評価傾向が同一のレビュアを利用する方針について述べる．我々は評価傾向を考慮した満足度予測タスクにおいて，同一レビュアのレビューと満足度を

あなたはレビュー評価システムです.

【重要】以下の指示に厳密に従ってください：

1. 参考例は学習用です。これらを評価してはいけません
2.「評価対象」セクションのレビュー 1 件のみを評価してください
3. 回答は 1〜5 の整数を 1 つだけ出力してください
4. 複数の評価は禁止です

【参考例】（これらは評価しません）

レビュー：日曜日に宿泊しましたが，大変コスパが良かったです.

評価値：5

⋮

レビュー：部屋は広くて清潔．近くにコンビニもあり不便も感じませんでした.

評価値：3

【評価対象】（この 1 件のみを評価してください）

レビュー：前日に低価格で予約できたのはここだけでした.

上記レビューの評価値（1〜5 の整数のみ）：

図 5.9: Few-Shot Prompting の一例

参考データとして使用する場合が最も高い精度を示すと考える．推論対象レビュア u_t と評価傾向が同一のレビュアを選択するために，我々は同一レビュア ID，つまり参考データとして選択されたレビュア集合 U_s が $U_s = \{u_t\}$ となるようにレビュアの検索と選択を行う．その後，我々は選択されたレビュア U_s が持つデータ $D(U_s)$ の中からレビューとそれに対応する評価値をランダムに l_s 件サンプリングし，LLM の推論を補助する参考データ $D_s(U_s)$ とする．つまり，$|D_s(U_s)|{=}l_s$ である．この時，$D_s(U_s)^{l_s} \subset D(u_t)$ であり，$r_j^t \notin D_s(U_s)^{l_s}$ である.

次に，推論対象レビュアと評価傾向が類似，また類似していないレビュアを利用する方針について述べる．我々は評価傾向を考慮した満足度予測タスクにおいて，類似レビュアのレビューと満足度を参考データとして使用する場合が同一レビュアを用いた場合の次に高い精度を示すと考える．また，我々は類似していないレビュアのレビューと満足度を参考データとして使用する場合が最も低い精度を示すと考える．我々は推論対象レビュアと評価傾向が類似，また類似していないレビュアを検索するために，レビュアの評価傾向である混同行列 $C^{(u_i)}$ を利用する．我々は推論対象レビュア u_t の混同行列 $C^{(u_t)}$ と全てのレビュアの混同行列 $C^{(U)}$ で総当たりでの類似度計算を実施する．総当たりでの類似度計算は，$Similar(v(C^{(u_\alpha)}), v(C^{(u_\beta)}))$，$u_\alpha, u_\beta \in U$ と表す．具体的な類似度計算手法については本

節にて後述する．その後，我々は評価傾向の類似した順にレビュアをランキングする．類似度順にランキングされたレビュア列を $U^r = \{u_{r_1}, u_{r_2}, \ldots, u_{r_n}\}$ とする．ランキングの後，選択対象のレビュアに応じて参考データ選択時の操作が変化する．

推論対象レビュアと評価傾向が類似しているレビュアを選択する場合，我々は U^r における上位のレビュア u_{r_1} から順に順位 h までのレビュアのデータ集合を統合する．この時，統合されたデータ集合は $D(U_h) = \bigcup_{i=1}^{h} D(u_{r_i})$ と表せる．また，レビュア集合は $U_h = \{u_{r_1}, u_{r_2}, \ldots, u_{r_h}\}$ と表せる．この時，$h \geq 1$ である．推論対象レビュアと評価傾向が類似していないレビュアを選択する場合，我々は U^r における下位のレビュア u_{r_n} から順に順位 l までのレビュアのデータ集合を統合する．この時，統合されたデータ集合は $D(U_l) = \bigcup_{i=1}^{l} D(u_{r_i})$ と表せる．また，レビュア集合は $U_l = \{u_{r_n}, u_{r_{n-1}}, \ldots, u_{r_l}\}$ と表せる．この時，$l \geq 1$ である．最後に，我々は $D(U_h)$，または $D(U_l)$ のデータ数が規定数 l_s を満たす最小の順位を求め，l_s 件のレビューと評価値を LLM の推論を補助する参考データ $D_s(U_s)^{l_s}$ として抽出する．この時，規定数 l_s を満たす最小の順位が

$$h^* = \min\{h \mid |D(U_h)| \geq l_s\} \tag{5.2.1}$$

$$l^* = \min\{l \mid |D(U_l)| \geq l_s\} \tag{5.2.2}$$

であるならば，$D_s(U_s)^{l_s} \subseteq D(U_{h^*})$，$D_s(U_s)^{l_s} \subseteq D(U_{l^*})$ である．

本実験で我々が選択する類似度計算手法について述べる．類似度計算の対象である混同行列 $C^{(u_i)}$ はベクトル化することにより，空間的な比較が可能である．そのため，ベクトル間の相対的な方向の類似性を評価する指標として，角度ベースの類似度計算手法である cos 類似度が広く用いられている．また混同行列 $C^{(u_i)}$ は正規化により，確率分布的に扱うことが可能である．そのため，分布間の差異を評価する指標として，距離ベースの類似度計算手法である L1 ノルム，L2 ノルムや分布ベースの類似度計算手法である KL ダイバージェンスなどの類似度計算指標も用いられている．しかし，前述した手法はいずれも入力を一次元のベクトルとして扱う．我々は，前述した手法による類似度計算では混同行列 $C^{(u_i)}$ が持つ要素の位置やばらつきといった構造関係に関する情報が失われ，適切な類似レビュア計算が不可能になると考えた．よって本稿では，我々は類似レビュアを計算するアプローチとして構造ベースの類似度計算手法である SSIM が有効であるという仮説を立てた．SSIM の一般的な計算式は以下の通りである．

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5.2.3}$$

二つの行列を x と y で表す時，μ_x, μ_y は各行列の要素平均値，σ_x, σ_y は各行列の標準偏差，σ_{xy} は二つの行列における共分散である．また，C_1 と C_2 は計算時に 0 で除算しないよう設定するための安定化定数であり，$C_1 = (K_1 L)^2, C_2 = (K_2 L)^2$ である．本実験において，K_1,K_2,L はそれぞれ $K_1 = 0.01$，$K_2 = 0.03$，$L = 255$ である．また本実験において，我々は考察の幅を広げるために一般的な SSIM 以外に 擬似的な ms-ssim(pseudo Multi Scale-SSIM;pms-ssim) と Block-SSIM を実施した．pms-ssim ではなく，pms-ssim とした理由は，混同行列 $C^{(u_i)}$ において自然な解像度の概念が存在しないためである．これら三種類の SSIM における詳細は 2.9.4 節で述べる．

また，我々は類似度計算における比較手法として cos 類似度も計算する．cos 類似度の一般的な計算式は以下の通りである．

$$cos(\mathbf{x}, \mathbf{y}) = \frac{\sum_{n=1}^{d} x_n y_n}{\sqrt{\sum_{n=1}^{d} x_n^2} \sqrt{\sum_{n=1}^{d} y_n^2}} \tag{5.2.4}$$

最後に，評価傾向に関係なくランダムにサンプリングしたレビュアを利用する方針について述べる．我々は評価傾向を考慮した満足度予測タスクにおいて，ランダムにサンプリングしたレビュアのレビューと満足度を参考データとして使用する場合が 3 番目に高い精度を示すと考える．ランダムにサンプリングしたレビュアを選択する際，評価傾向は考慮しない．そのため，参考データとして選択されたレビュア集合 U_s は要素として推論対象と同一のレビュア u_t や類似したレビュアを含む可能性もある．サンプリングの条件は，参考データとして選択されたレビュア集合の要素数 $|U_s|$ が $|U_s| \geq 1$ を満たすかつ，参考データ $D_s(U_s)$ が $|D_s(U_s)| = l_s$ を満たすことである．

参考データ選択手法

本実験では 4 種類の参考データ選択方針に基づく計 14 種類の参考データ選択手法を使用する．計 14 種類の参考データ選択手法は以下の通りである．

- 推論対象レビュアと評価傾向が同一のレビュアを利用
    1. same
- 推論対象レビュアと評価傾向が類似したレビュアを利用
    2. cosine
    3. ssim
    4. pms-ssim
    5. block-ssim

6. combined-cosine
7. combined-ssim

- 評価傾向に関係なくランダムにサンプリングしたレビュアを利用

8. random

- 推論対象レビュアと評価傾向が類似していないレビュアを利用

9. alter cosine
10. alter ssim
11. alter pms-ssim
12. alter block-ssim
13. alter combined-cosine
14. alter combined-ssim

same テクニックは，推論対象レビュアと評価傾向が同一，つまり推論対象レビュアの他データを参考データとして利用する参考データ選択手法である．random テクニックは，レビュアの評価傾向に関係なくランダムにサンプリングしたレビュアを参考データとして利用する参考データ選択手法である．

残りの 12 テクニックは類似度計算結果を基に参考データを選択する．cosine テクニックは，評価傾向が類似したレビュアを選択するために混同行列間の類似度計算に直接 cos 類似度を用いる．ssim テクニックは，評価傾向が類似したレビュアを選択するために混同行列間の類似度計算に直接 SSIM を用いる．pms-ssim テクニックは，混同行列における全ての要素を中心とした 3×3 の小領域にて SSIM を計算して平均を取ったものと，元の 5×5 行列から SSIM を計算したものを同等の重みで乗算し，それをレビュア間類似度とする．block-ssim テクニックは，混同行列内の 3×3 のブロック各所で SSIM を計算した後に平均を取得したものをレビュア間類似度とする．combined-cosine テクニックは，混同行列内の 3×3 のブロック各所で平均を計算した後に cos 類似度計算を適用する．combined-ssim テクニックは，混同行列内の 3×3 のブロック各所で平均を計算した後に SSIM を適用する．cosine テクニックから combined-ssim テクニックまでの参考データ選択手法では，得られた類似度が高いレビュアを参考データとする．alter 系列の参考データ選択手法は，得られた類似度が低いレビュアを参考データとする．参考データ選択手法によって選択される混同行列が変わる具体例を付録 C に示す．

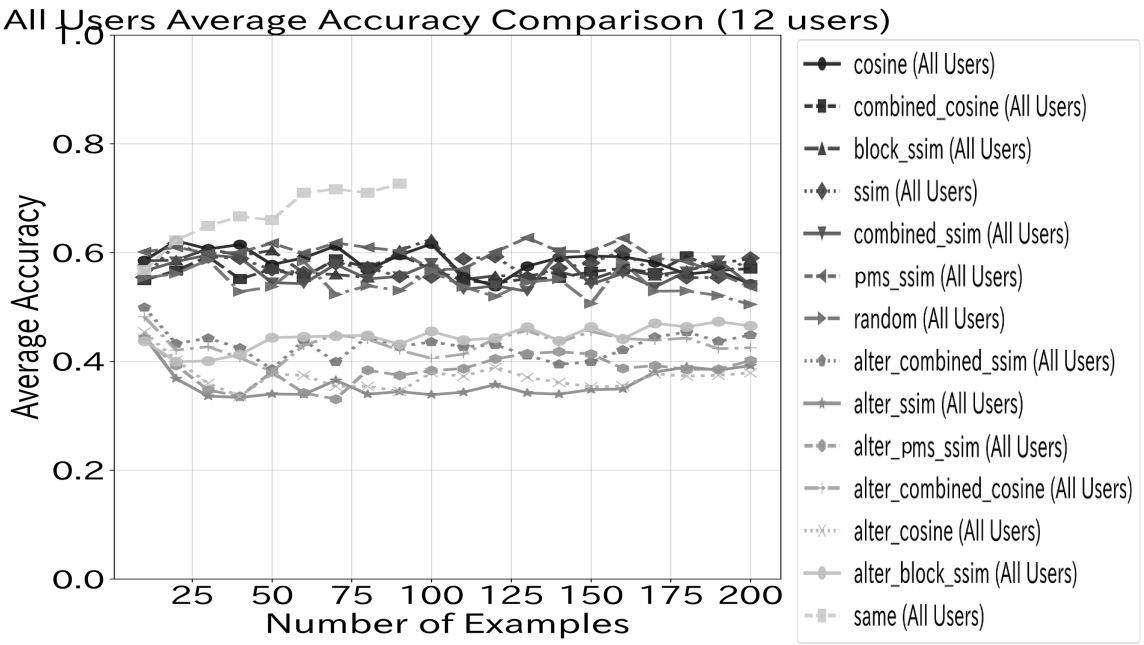

図 5.10: 全データ数帯における各参考データ選択手法の精度比較 (Accuracy)

5.2.2 実験準備

本実験では，混同行列を用いた参考データ選択とそれを利用したレビュー満足度予測タスクを実施する．入力はレビューと参考データである．出力はレビュアの満足度である．出力される満足度は 1〜5 の整数値である．使用するデータセットは，実験 1 と同様に楽天トラベルデータセットである．対象データは実験 1 の二つ目の分析結果を踏まえて設定した四つのデータ数帯から，ランダムサンプリングで抽出した各 3 人ずつのレビュアとする．使用する LLM は，実験 1 の三つ目の分析と同様に Gemma3:27B である．量子化は行わない．本実験の目的は混同行列による評価傾向の表現妥当性の検証であるが，混同行列の妥当性は混同行列を用いた参考データ選択における予測性能の比較によって間接的に評価可能である．そのため，本実験では精度の評価指標として Accuracy と RMSELoss を用いる．RMSELoss は実験 1 と同様に，LLM の予測値が実際の満足度評価値から大きく外れている場合とそうでない場合を反映させるために用いる．統計的検定として対応のある 2 標本 t 検定を実施した．参考データ選択手法は 5.2.1 節で述べたものに従う．

5.2.3 結果・考察

実験の結果を図 5.10 と図 5.11 に示す．図 5.10 は全データ数帯計 12 人のレビュアに対し，14 種類の参考データ選択手法による Few-Shot Prompting を用いた満足度予測の Accuracy を比較した結果である．図 5.11 は全データ数帯計 12 人のレビュアに対し，14 種類の参考

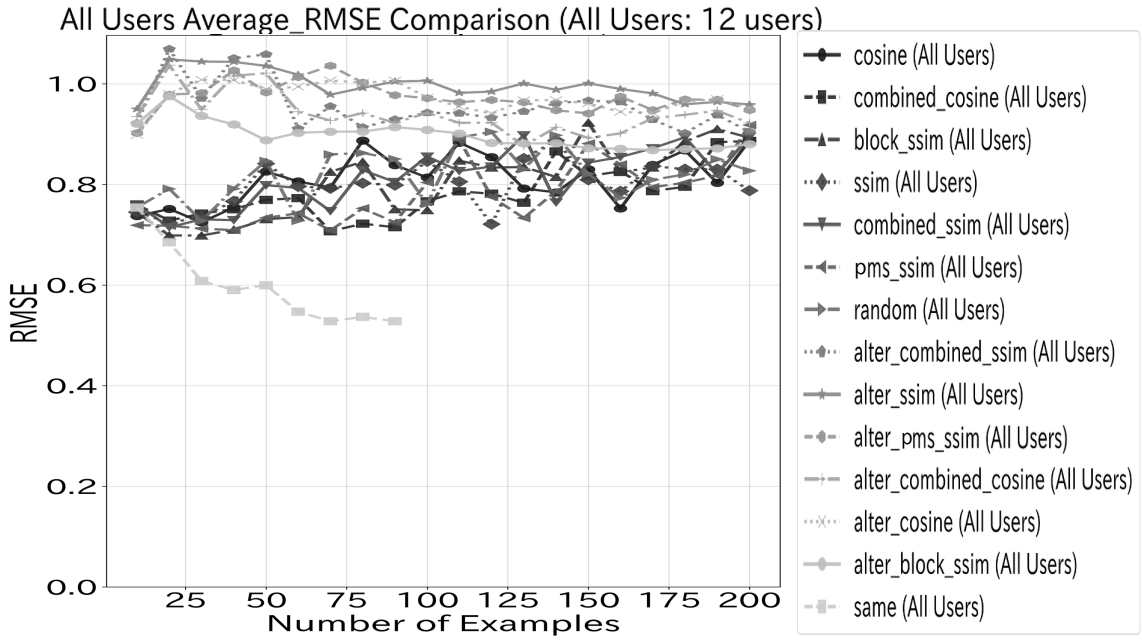

図 5.11: 全データ数帯における各参考データ選択手法の精度比較 (RMSE)

データ選択手法による Few-Shot Prompting を用いた満足度予測の RMSELoss を比較した結果である．

図 5.10 や図 5.11 では，3 段階の精度帯が確認できる．最も良い精度帯は，same テクニックによる参考データ選択を実施した場合が該当する．same テクニックは，Accuracy が 0.6 から 0.8 の間，RMSELoss が 0.8 から 0.4 の間に位置している．また，与える参考例数が増加するにつれて予測精度は向上する．次に良い精度帯は，ssim 系列や cosine 系列，そして random テクニックによる参考データ選択を実施した場合である．この精度帯では，Accuracy が 0.5 付近から 0.7 付近の間，RMSELoss が 0.9 付近から 0.7 付近に位置している．そして最も悪い精度帯は，alter 系列の類似度計算による参考データ選択を実施した場合である．この精度帯では，Accuracy が 0.3 付近から 0.5 付近の間，RMSELoss が 1.1 付近から 0.9 付近に位置している．

より詳細に考察を実施するため，統計的検定結果を表 5.8 や表 5.9 に示す．表 5.8 は，14 種類のテクニックに対して Accuracy と RMSELoss がどれほど有意な差を持つのかを示したものである．表 5.9 は，各テクニック間の主要な有意関係性を示したものである．表 5.8 より，same テクニックは一つを除いて全てのテクニックに対して有意に精度が高い．また，alter 系列は全体的に劣位である．特に，alter ssim は一つを除いて全てのテクニックに対して有意に精度が低い．これらの結果は，same が最も良い精度帯であり，alter 系列が最も悪い精度帯であることを根拠付けるものである．そして表 5.9 より，cosine 系列や ssim

系列の参考データ選択手法は．random テクニックと比較して有意に精度が高いことが読み取れる．したがって，前述した精度帯の中では random テクニックが比較的劣位な参考データ選択手法であることが示された．これは，図 5.10 や図 5.11 からも読み取れる．また，random テクニックは alter 系列のテクニックと比較して有意に精度が高いことが読み取れる．一方で，cosine 系列と ssim 系列の比較ではどちらかが有意に優れているという結果は確認できなかった．この結果は，ssim 系列のテクニックによって構造的な情報を維持することが評価傾向間の類似度計算を実施する際に有効であるとは言い切れないことを示唆する．以上の結果から，テクニック間の精度における関係性が same>cosine 系列・ssim 系列>random>alter 系列の順になることがわかる．つまりこれらの結果は，LLM に対して与える参考データとして推論対象レビュア u_t と評価傾向が類似しているレビュアを用いるほど満足度予測精度が向上するという仮説が正しいことを示している．つまり，混同行列による評価傾向の表現は妥当であると言える．各データ数帯における精度や表 5.9 に示した以外の有意関係は付録 C に示す．

表 5.8: 14 種類の参考データ選択手法における有意関係 (all_users_average)

参考データ選択手法	Accuracy		RMSE	
	有意差あり (Win/Lose)	有意差なし	有意差あり (Win/Lose)	有意差なし
same	12/0	1	12/0	1
cosine	9/2	2	7/5	1
pms-ssim	9/2	2	7/5	1
block-ssim	8/3	2	7/5	1
combined-cosine	8/3	2	9/3	1
ssim	8/3	2	9/3	1
combined-ssim	8/4	1	8/4	1
random	8/5	0	7/5	1
alter combined-ssim	5/8	0	5/7	1
alter combined-cosine	5/8	0	5/7	1
alter block-ssim	4/9	0	4/8	1
alter pms-ssim	3/10	0	3/9	1
alter cosine	1/11	1	1/11	1
alter ssim	0/12	1	0/12	1

Win: 当該手法が有意に優れている比較数

Lose: 当該手法が有意に劣っている比較数

表 5.9: 主要な手法間の統計的有意差比較 (all_users_average)

比較	Accuracy			RMSE		
	平均差	p 値	有意差	平均差	p 値	有意差
cosine/ssim 系列 vs. random						
cosine vs. random	0.0495	<0.0001	***	-0.0821	<0.0001	***
pms-ssim vs. random	0.0582	<0.0001	***	-0.0359	<0.0001	***
ssim vs. random	0.0386	0.000020	***	-0.0650	<0.0001	***
com-ssim vs. random	0.0216	0.002786	**	-0.0422	<0.0001	***
com-cosine vs. random	0.0274	0.002503	**	-0.0343	0.027109	*
block-ssim vs. random	0.0362	0.000197	***	-0.0133	0.377819	-
random vs. alter 系列						
random vs. alter cosine	0.1779	<0.0001	***	-0.1469	<0.0001	***
random vs. alter ssim	0.1881	<0.0001	***	-0.1678	<0.0001	***
random vs. alter pms-ssim	0.1712	<0.0001	***	-0.1533	<0.0001	***
cosine 系列 vs. ssim 系列						
cosine vs. pms-ssim	-0.0087	0.079441	-	0.0015	0.776508	-
cosine vs. ssim	0.0109	0.091272	-	-0.0096	0.016732	*
cosine vs. block-ssim	-0.0133	0.034583	*	0.0130	0.438365	-
cosine vs. com-ssim	0.0280	<0.0001	***	-0.0399	<0.0001	***
com-cosine vs. pms-ssim	-0.0203	0.011313	*	0.0126	0.055360	-
com-cosine vs. ssim	0.0112	0.103009	-	-0.0190	0.347054	-
com-cosine vs. block-ssim	-0.0032	0.664859	-	0.0012	0.856179	-
com-cosine vs. com-ssim	0.0253	0.010559	*	-0.0258	0.005359	**
ssim 系列同士の比較						
pms-ssim vs. ssim	0.0196	0.008415	**	-0.0081	0.040967	*
pms-ssim vs. block-ssim	0.0172	0.000090	***	0.0114	0.012137	*
pms-ssim vs. com-ssim	0.0457	<0.0001	***	-0.0384	0.000066	***
ssim vs. block-ssim	-0.0062	0.128376	-	0.0033	0.435206	-
ssim vs. com-ssim	0.0347	<0.0001	***	-0.0303	0.000293	***
block-ssim vs. com-ssim	0.0285	0.000813	***	-0.0270	0.000149	***

*** p<0.001, ** p<0.01, * p<0.05, -: 有意差なし

平均差 = 各精度指標の平均差 (前者 - 後者)

5.3 実験 3: 提案手法の有効性検証

本実験において我々は，提案手法が主観評価に内在する個人差の解消に有効であるという仮説に対する検証実験を実施する．ニューラルネットワークへ混同行列と正規化対象の満足度を入力し，多数のレビュアに共通する平均的な評価傾向を近似的に表現する参照尺度として

定義した Zero-Shot 推論による満足度予測結果を出力する．実験手順は 4 節で述べた通りである．

5.3.1 実験準備

本実験におけるタスクは，正規化満足度の回帰タスクである．ニューラルネットワークの入力層は 4.1.2 節で述べたとおり $1 + k^2$，つまり 26 次元である．中間層は 13 次元 1 層とした．出力層は回帰タスクであるため，1 次元である．活性化関数は ReLU 関数である．層間は全結合とする．使用するデータは，実験 1 の二つ目の分析で用いた 350 人のレビュアにおけるレビューと満足度，混同行列である．付与された満足度の割合を維持したまま，訓練用: 検証用: テスト用=8:1:1 としてデータを分割した．割合を維持した理由は，正解ラベルが多数のレビュアに共通する平均的な評価傾向を近似的に表現する参照尺度に基づく満足度であるためである．訓練用データは 32615 件である．検証用データは 4028 件である．テスト用データは 4027 件である．入力する混同行列の要素は正規化した．正規化した理由は，レビュアごとのレビュー投稿数の違いが特徴量のスケールとして出力に反映されてしまうことを防ぐためである．earlystopping は 100 エポックとした．バッチサイズは 32 である．10 分割交差検証を実施した．本実験の目的は主観尺度に内在する個人差の補正であるが，補正の有効性は補正後の予測性能として間接的に評価可能である．そのため，本実験では評価指標として RMSELoss，決定係数 (R^2)，MAELoss といった回帰用の評価指標に加え，Accuracy を使用した．Accuracy を使用した理由は，本来の満足度が離散値であるためである．Accuracy を計算する際は，得られた連続値の出力を四捨五入して整数とした．

表 5.10: 10 分割交差検証での結果

Fold	Best Epoch	評価指標			
		RMSE	MAE	決定係数 (R^2)	Accuracy
1	142	0.5626	0.4084	0.3988	0.6687
2	135	0.5625	0.4047	0.3995	0.6744
3	270	0.5667	0.4100	0.3905	0.6805
4	92	0.5753	0.4152	0.3719	0.6722
5	882	0.5695	0.4184	0.3815	0.6667
6	112	0.5548	0.4024	0.4130	0.6809
7	53	0.5646	0.4031	0.3921	0.6809
8	171	0.5543	0.4016	0.4139	0.6922
9	205	0.5576	0.3948	0.4069	0.6961
10	329	0.5402	0.3943	0.4454	0.6841
平均	**239.1**	**0.5608**	**0.4053**	**0.4014**	**0.6797**
標準偏差	**240.6**	**0.0097**	**0.0079**	**0.0204**	**0.0096**

5.3.2 結果・考察

実験結果を表 5.10 に示す．10 分割交差検証の結果，決定係数は平均して 0.4014 であり，Accuracy は平均して 0.6797 であった．決定係数はそこまで高い値を取っていない．しかしこの原因は，本来離散値であるカテゴリ変数を回帰で表現したためである．よって我々は，決定係数が低い値を取ること自体は問題ないと考える．一方で，Accuracy は約 7 割ほどである．5 段階の分類タスクにおいてこの結果は悪くないものである．この結果は，主観評価に内在する個人差の解消というタスクにおいて本提案手法が有効に働く傾向を有するということを示唆する．また，RMSELoss は平均して 0.5608 であり，MAELoss は平均して 0.4053 であった．両 Loss の差について考察するため，最も予測誤差が大きいケースと次点で予測誤差が大きいケースを抽出する．また，最も予測誤差が小さいケースと次点で予測誤差が小さいケースも抽出する．

それぞれのケースを表 5.11，表 5.12，表 5.13 と表 5.14 に示す．表 5.11 と表 5.12 では，いずれのレビュアも混同行列に例外的なレビューケースが複数存在する．またいずれのレビュアにおいても，予測誤差が大きいケースは例外的なレビューケースを対象としたときである．これは，例外的なレビューケースに対する個人差の補正が本提案手法において有効ではないことを示唆している．しかし，例外的なレビューケースに該当しない満足度に対する個人差の補正ではいずれの予測誤差も 0.25 を切っている．この結果は，混同行列が構造的な個人差を捉える際に有効であることを示している．したがって，以上の結果は例外的なレ

表 5.11: 予測誤差が最大のケースを持つレビュア（最大予測誤差 = 3.64)

(a) 予測情報

タイプ	入力満足度	予測満足度	正解	予測誤差
予測誤差 (大)	1.0	1.36	5.0	3.64
予測誤差 (小)	5.0	4.77	5.0	0.23

(b) 混同行列の全要素

$\hat{y}\backslash y$	1	2	3	4	5
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0.05
4	0	0	0	0.25	0.2
5	0.05	0	0	0	0.45

ビューケースのような大きな個人差が内在する主観評価を行うレビュアほど個人差の補正が困難であることを示唆している．また，表 5.13 と表 5.14 では，いずれのレビュアも混同行列に例外的なレビューケースが存在していない．そしていずれのレビュアにおいても，予測誤差が小さいケースはほぼ完全な個人差の正規化がなされている．これは，多数のレビュアが持つ平均的な判断基準に基づいた評価傾向を持つレビュアほど完全に近い形で個人差を補正することが可能であることを示唆している．これらの考察は，予測誤差の大きさに基づいて抽出した 4 種類のレビュアサンプルから得られたものである．そのため，得られた考察は抽出したサンプルに依存している可能性がある．しかしながら，例外的なレビューケースの割合や評価傾向における固有構造の有無によって補正効果が変化するという傾向は，混同行列による個人差補正の適用条件を示す重要な示唆であると考えられる．

本実験では，主に二つの考察が得られた．一つ目は，本提案手法は主観評価に内在する個人差の解消というタスクにおいて有効に働く傾向があるということである．二つ目は，本提案手法は例外的なレビューケースのような大きな個人差の解消において改善する余地があるということである．しがたって，提案手法が主観評価に内在する個人差の解消に有効であるという仮説は正しいと言える．

表 5.12: 予測誤差が二番目に大きいケース（最大予測誤差 = 3.39）

(a) 予測情報

タイプ	入力満足度	予測満足度	正解	予測誤差
予測誤差 (大)	5.0	4.39	1.0	3.39
予測誤差 (小)	3.0	3.79	4.0	0.211

(b) 混同行列の全要素

$\hat{y}\backslash y$	1	2	3	4	5
1	0.00135	0	0	0.00135	0.00270
2	0	0.00135	0	0.00270	0
3	0	0	0.00270	0.03374	0.00540
4	0	0	0.00810	0.44130	0.16464
5	0	0	0.00270	0.26451	0.06748

表 5.13: 予測誤差が最小のケース（最小予測誤差 = 0.0003）

(a) 予測情報

タイプ	入力満足度	予測満足度	正解	予測誤差
予測誤差 (大)	3.0	3.19	2.0	1.19
予測誤差 (小)	5.0	4.0	4.0	0.0003

(b) 混同行列の全要素

$\hat{y}\backslash y$	1	2	3	4	5
1	0	0	0	0	0
2	0	0	0.033	0	0
3	0	0	0.033	0.033	0
4	0	0	0.133	0.5	0.067
5	0	0	0	0.2	0

表 5.14: 予測誤差が二番目に小さいケース（最小予測誤差 = 0.0004）

(a) 予測情報

タイプ	入力満足度	予測満足度	正解	予測誤差
予測誤差 (大)	3.0	3.16	2.0	1.16
予測誤差 (小)	5.0	4.0	4.0	0.0004

(b) 混同行列の全要素

$\hat{y}\backslash y$	1	2	3	4	5
1	0	0	0	0	0
2	0	0	0.01	0	0
3	0	0	0	0.03	0
4	0	0	0.01	0.31	0.58
5	0	0	0	0.01	0.05

第 6 章　おわりに

本研究の目的は，リッカート尺度に基づく主観評価に内在する意味解釈の個人差を解消することである．我々は本研究の目的を達成するために，リッカート尺度に基づく個人差の内在した主観評価値を多数のレビュアに共通する個人差の内在しない参照尺度に基づく評価値へと正規化する手法を提案した．つまり本研究は，満足度予測タスクを予測ではなく主観評価に内在する意味解釈の個人差解消のための正規化タスクとして問題設定を再定義した．また本研究は，既存研究でノイズとして扱われる例外的なレビューケースの存在をレビュアの混同行列が保有する個人差として構造的に扱った．

我々は，評価値とは独立に存在するものではなく評価者が持つ評価理由に基づいて付与された結果であるという立場に立つ．この立場に基づく場合，主観評価に内在する個人差は評価理由と評価値における対応関係の違いとして捉えることが可能となる．よって，我々はレビュアごとに異なる評価理由と評価値の対応関係をレビュアの評価傾向として捉えた．また評価傾向を定義する際，我々は例外的なレビュケースの存在に着目した．例外的なレビュケースとは，レビュアが実際に付与した満足度とレビューから推測される満足度が大きく異なるレビューのことを指す．我々は，例外的なレビューケースを主観評価に個人差が内在する要因の一つであると考えた．以上のことから本稿において我々は，レビュアの評価傾向をレビュアが付与した満足度と評価理由を表すレビューから LLM の Zero-Shot 推論によって推測される満足度の共起頻度，つまり混同行列として定義した．これにより我々は，例外的なレビュケースの存在を構造的な情報として保持した評価傾向の表現が可能になると考えた．

本稿において我々は，三つの仮説を立てた．一つ目の仮説は，提案手法の核となる例外的なレビューケースや意味解釈の個人差は実データ上に存在する現象であるということである．本提案手法は，実際のレビューデータにおいて例外的なレビューケースが存在していることが前提となる．したがって，我々は例外的なレビューケースに該当する基準を明確に定義する必要があると考えた二つ目の仮説は，例外的なレビューケースはレビュー投稿数に関係しているということである．投稿数が極端に少ないレビュアに例外的なレビューケースが集中している場合，我々はレビュアごとの評価傾向を安定的に推定することが困難となると考えた．したがって，我々は例外的なレビューケースが不規則に発生するものなのか，あるいはレビュー投稿数に応じて発生が偏るものなのかを明らかにする必要があると考えた．三つ目の仮説は，レビュアの評価傾向が例外的なレビューケースと関係しているということである．我々は既存手法と異なり，例外的なレビューケースを単なるノイズとして扱うべき対象ではないと考えた．したがって，我々は例外的なレビューケースによる評価値への影響がス

カラ値で表せるような一様なものではないことを示す必要があると考えた.

我々は，三つの仮説に対する検証と提案手法の有効性確認のために，五つの実験を実施した．一つ目の実験では，例外的なレビューケースに該当する基準の調査及び例外的なレビュケースの出現率を確認した．二つ目の実験では，例外的なレビューケースの出現率とレビュー投稿数の関係性を調査した．三つ目の実験では，満足度予測タスクを用いてレビュアの評価傾向と例外的なレビューケースの関係性を調査した．四つ目の実験では，評価傾向によって参考データ選択の条件を変更した満足度予測タスクを用いて混同行列による評価傾向の定義妥当性を検証した．五つ目の実験では，ニューラルネットワークによる評価値の正規化タスクを用いて提案手法が主観評価に内在する個人差の解消に有効であるか検証した.

一つ目の実験結果より，実際の満足度とレビューから推測された満足度の差が 2 以上あるものを例外的なレビュケースと定義した．また，例外的なレビューケースは想像上の概念ではなく実データ上に存在する現象であることを確認した．二つ目の実験結果より，例外的なレビューケースの出現はレビュー投稿数に関係していることを確認した．また，投稿レビュー数が少ないレビュアほど例外的なレビュケースの偶発的な発生によって評価傾向の推定が不安定になるという考察が得られた．加えて，投稿レビュー数帯が少ないレビュアほど例外的なレビューケースに該当する要素が混同行列上の構造として現れやすいという考察が得られた．三つ目の実験結果より，レビュアの評価傾向が例外的なレビューケースと関係していることを確認した．これにより，レビュアの固有特徴情報を内包する例外的なレビューケースの存在をノイズとして扱うべきではないことを確認した．以上の実験結果より，例外的なレビューケースと評価傾向が示す個人差構造が実データ上に存在することを示した．四つ目の実験結果より，混同行列によってレビュアの評価傾向を定義することの妥当性を確認した．この結果より，上記の結果で示された評価傾向が示す個人差構造を混同行列で捉えることが可能であることを示した．五つ目の実験結果より，本提案手法が主観評価に内在する個人差の解消というタスクにおいて有効に働く傾向にあることを確認した．一方で，例外的なレビューケースに対する個人差の解消は改善の余地があることを確認した．この結果より，四つ目の実験で示された混同行列による評価傾向表現を用いることによって主観評価に内在する個人差の補正とそれに伴う意味空間上での正規化が可能であることを示した.

つまり，五つの実験の結果から意味解釈の個人差を評価傾向の構造として扱うことによって，既存研究ではノイズとして扱われていた個人差が体系的に補正可能であることが示された．しかし，評価傾向に個人差構造が存在する場合において有効性が示唆された一方，個人差構造が乏しい場合においては補正効果が限定的であることも明らかとなった．また，本提案手法の補正により主観尺度は共通した尺度上の客観的評価値として扱うことが可能となるため，評価値間の直接的な比較や集計から導かれる分析結果の安定した解釈を可能とするこ

とが示された．一方で本稿での実験結果より，本提案手法にはいくつかの懸念事項が存在している．

一つ目は本稿において我々が，多数のレビュアに共通する平均的な評価傾向を近似的に表現する参照尺度を LLM による Zero-Shot 推論の予測値で定義したことである．4.2 節でも述べた通り，本稿において我々は追加学習や参考情報の添付を実施していない事前学習済み LLM が Zero-Shot 推論によって妥当な評価を返すという仮定をした．参照尺度の定義は，この仮定の下で成り立つ．しかし，我々はこの仮定自体の妥当性を確認できていない．そのため，我々はこの仮定の妥当性を確認するか，多数のレビュアに共通する平均的な評価傾向を近似的に表現する参照尺度を LLM による Zero-Shot 推論の予測値以外で定義する必要があると考える．我々は前者に関して，人力のアノテーションとそれに伴う多数決を利用することによって，仮定の検証が可能であると考える．また我々は後者に関して，複数の LLM を用いた Zero-Shot 推論の多数決や不特定多数のレビュアが書いたレビューをランダムに参考データとした Few-Shot Prompting によって解決可能だと考える．

二つ目は，例外的なレビューケースに対する個人差の解消方法である．本稿の実験において，例外的なレビューケースに対する個人差の解消には改善の余地があることを確認した．我々は，例外的なレビューケース要素に対する重みづけを改善アプローチとして提案する．例外的なレビューケースは，基本的に他要素と離れた位置に存在している．そして，レビュアが持つ例外的なレビューケース割合は平均して 5.93% 程度である．これらのことより，我々は混同行列内の例外的なレビューケース要素がモデル推論時に重要視されていないと考える．したがって，我々は例外的なレビューケース要素に対する重みづけが有効なアプローチとなるのではないかと考えた．

三つ目は，本提案手法の汎化的な適用範囲である．本稿では，5 段階のリッカート尺度を対象として様々な実験を実施した．そのため，7 段階のリッカート尺度や満足/不満足のみの 2 段階尺度に対しては，本稿の実験結果や考察が当てはまるかが定かではない．特に例外的なレビューケースの定義という点においては，本提案手法を拡張する必要がある．したがって，我々は尺度段階が増加した場合においての例外的なレビューケースの定義や前述した例外的なレビューケース要素に対する適切な重みづけを改めて調査・分析する必要があると考える．

謝辞

前回これを書いてから 2 年の時が流れました．みなさまお元気でしょうか．私は一般的で平均的ないい感じのぼちぼちです．ですが最近は時の流れがとても早く感じます．老いがきたのでしょうか．若かったあの頃の元気はどこへ行ってしまったのでしょうね．時間に対する体感速度は歳をとるごとに速さを増していくそうです．今の状態でこれなら還暦頃には瞬き一つで空を変化させられるようになると思います．厨二病ですかね．すぐに過ぎ去ってしまう忙しない日常で，大切な人たちの姿をしっかりと焼き付ける．そのために，私はしっかりと毎日を大切に過ごしていくという意識を持とうと思いました．文字数は稼げましたかね．

閑話休題．この 2 年で，私はいろいろな経験をさせていただきました．目まぐるしく変わる環境と成長．とても充実した院生生活だったと思います．もちろん，楽しいことばかりではありませんでした．研究においてもコミュニケーションにおいてもいろいろな困難がありました．その度に，心の支えとなってくれた家族や友人，研究室の仲間の優しい言葉に何度も勇気をもらいました．家族はめげそうになっている自分の背中をしっかりと支えてくれました．高橋や城所を筆頭に，友人はみんな真剣に相談に乗ってそれぞれの想いと言葉で激励してくれました．M1 ズや先輩を筆頭に研究室の仲間は苦しみを分かち合い，前を向く元気と活力をくれました．みんな本当にありがとう．B4 の子や B3 の子とはそこまで長く一緒の時間を過ごしたわけではないけれど，話せてとても楽しかったです．また何かあったらいつでも連絡してきてください．先輩として必ず助けになるので．また，井尾さんや駒澤さん，そして鈴木先生にも本当に感謝しています．これまでありがとうございました．井尾さんや駒澤さんには雑務の面でいろいろ助けていただきました．仕事中に話しかけた際も快く応じていただきましたね．それにお二人から雑談を投げかけていただいたこともありました．とても嬉しかったです．ありがとうございました．鈴木先生にはこの研究室に所属してから 3 年半，いろいろなお言葉をいただきました．ご迷惑をおかけしましたが，先生のお陰でとても多くのことを学びました．先生からいただいた数々のお言葉は今後も忘れません．ありがとうございました．

なんだかんだあったけれど，楽しかったなぁ．

謝辞

本研究の一部は JSPS 科研費（24K03044，23K28383）の助成を受けたものです．本研究では，NII IDR データセット提供サービスにより楽天グループ株式会社から提供を受けた「楽天データセット」（https://rit.rakuten.com/data_release/）を利用しました．

参考文献

[1] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, Vol. 20, No. 3, pp. 273–297, 1995.

[2] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

[3] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, Vol. 323, No. 6088, pp. 533–536, 1986.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, Vol. 30, , 2017.

[5] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, Vol. 27, , 2014.

[6] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, Vol. 30, , 2017.

[7] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[9] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, Vol. 13, No. 4, pp. 600–612, 2004.

[10] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The thrity-seventh asilomar conference on signals, systems & computers, 2003*, Vol. 2, pp. 1398–1402. Ieee, 2003.

[11] Sabahattin Yeşilçınar and Mehmet Şata. Examining rater biases of peer assessors in different assessment environments. *International Journal of Psychology and Educational Studies*, Vol. 8, No. 4, pp. 136–151, 2021.

[12] Harry Helson. Adaptation-level as a basis for a quantitative theory of frames of reference. *Psychological review*, Vol. 55, No. 6, p. 297, 1948.

[13] Edward L Thorndike, et al. A constant error in psychological ratings. *Journal of applied psychology*, Vol. 4, No. 1, pp. 25–29, 1920.

[14] Kushal Anjaria. Knowledge derivation from likert scale using z-numbers. *Information Sciences*, Vol. 590, pp. 234–252, 2022.

[15] Lotfi A Zadeh. A note on z-numbers. *Information sciences*, Vol. 181, No. 14, pp. 2923–2932, 2011.

[16] Bingkun Wang, Shufeng Xiong, Yongfeng Huang, and Xing Li. Review rating prediction based on user context and product context. *Applied Sciences*, Vol. 8, No. 10, p. 1849, 2018.

[17] 坂本新真, 牛尼剛聡. 評価値予測タスクを用いたユーザの嗜好に適合したレビューの抽出. 第 15 回データ工学と情報マネジメントに関するフォーラム (DEIM2023), pp. 1–8. 一般社団法人電気情報通信学会, 2023.

[18] 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, 2019.

[19] 松岡天音, 山口史弥, 上田真由美, 中島伸介. 肌質や肌悩みを考慮した類似ユーザ判定に基づくコスメアイテム推薦手法. 第 16 回データ工学と情報マネジメントに関するフォーラム (DEIM2024), pp. 1–7. 一般社団法人電気情報通信学会, 2024.

[20] 宮本達矢, 北山大輔. 共通レビュアーの観点の類似性に基づく書籍推薦手法. 第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM2018), pp. 1–8. 一般社団

法人電気情報通信学会, 2018.

[21] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, Vol. 28, No. 1, pp. 11–21, 1972.

発表リスト

[1] 川上大凱，鈴木優『レビュアのパーソナリティを考慮したレビュー分類手法』第 24 回情報科学技術フォーラム（FIT2025）, 2025.

[2] 川上大凱，鈴木優『レビューと評価のギャップを用いたレビュー分類精度向上手法』東海関西データベースワークショップ 2025，2025.

[3] 川上大凱，鈴木優『異なる評価基準が混在したリッカート尺度の正規化手法』第 18 回データ工学と情報マネジメントに関するフォーラム (DEIM 2026)，2026.

付録 A　実験 1: 二つ目の分析関連

本分析に用いるレビュー数帯別詳細レビュア数を表 A.1 に示す.

表 A.1: レビュー数帯別レビュア数

データ数帯	実際のレビュー数	人数	データ数帯	実際のレビュー数	人数	データ数帯	実際のレビュー数	人数
10	10	50	> 100	466	1	> 100	606	1
20	19	1		475	1		616	1
	20	49		480	1		621	1
30	30	50		485	2		644	1
40	40	50		489	1		656	1
50	50	50		492	1		661	1
100	100	40		502	1		677	1
	101	10		505	1		692	1
> 100	436	1		506	1		704	1
	438	1		512	1		723	1
	439	1		518	1		741	1
	440	2		541	1		782	1
	442	1		546	1		791	1
	446	1		577	1		815	1
	447	1		586	2		944	1
	449	1		588	1			
	452	1		592	1			
	453	1		595	1			
	455	1		601	1			
	465	1		602	1			
計: 350 人								

付録 B　実験 1：三つ目の分析関連

三つ目の分析における詳細な精度や p 値についてを表 B.2 と表 B.4 に示す．また，同分析に用いたレビュアの混同行列を表 B.2 と付表 B.3 に示す．

表 B.1: 分析に用いた分析対象レビュアの混同行列 (Concentrated)

(a) データ数帯 20(user_2346938)

$\hat{y}\backslash y$	1	2	3	4	5
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	1	0	0
4	0	0	0	18	0
5	0	0	0	0	1

(b) データ数帯 30(user_506302)

$\hat{y}\backslash y$	1	2	3	4	5
1	0	0	0	0	0
2	0	0	0	0	0
3	0		5	0	0
4	0	0	1	22	1
5	0	0	0	0	1

(c) データ数帯 40(user_714107)

$\hat{y}\backslash y$	1	2	3	4	5
1	0	0	0	0	0
2	1	4	0	0	0
3	0	1	4	3	0
4	0	0	1	20	3
5	0	0	0	0	3

(d) データ数帯 100(user_283470)

$\hat{y}\backslash y$	1	2	3	4	5
1	0	0	0	0	0
2	0	2	1	0	0
3	0	1	8	4	1
4	0	0	6	71	6
5	0	0	0	0	0

表 B.2: 各データ数帯の要素が対角線上に集合しているレビュアにおける満足度予測の各精度

User	**Metric**	**0.0**	**0.25**	**0.5**	**0.75**	**1.0**
データ数帯 20	Accuracy	0.98	0.98	0.92	0.62	0.17
	p 値		1.000	0.033*	0.00019***	0.00014***
	RMSE	0.089	0.089	0.249	0.615	0.911
	p 値		1.000	0.042*	0.001**	0.00046***
データ数帯 30	Accuracy	0.747	0.620	0.467	0.2793	0.187
	p 値		0.068	0.005**	0.004**	0.001**
	RMSE	0.502	0.612	0.730	0.804	1.006
	p 値		0.065	0.008**	0.003**	0.001**
データ数帯 40	Accuracy	0.670	0.635	0.570	0.420	0.2700
	p 値		0.296	0.129	0.014*	0.003**
	RMSE	0.574	0.627	0.688	0.827	0.997
	p 値		0.120	0.047*	0.00028***	0.003**
データ数帯 100	Accuracy	0.746	0.638	0.528	0.2790	0.196
	p 値		0.012*	0.011*	0.006**	0.00049***
	RMSE	0.515	0.628	0.728	0.833	1.015
	p 値		0.030*	0.044*	0.015*	0.001**

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

表 B.3: 分析に用いた分析対象レビュアの混同行列 (Dispersed)

(a) データ数帯 20(user_2193022)

$\hat{y}\backslash y$	1	2	3	4	5
1	0	0	0	1	0
2	0	0	0	0	0
3	0	0	2	1	4
4	0	0	0	1	11
5	0	0	0	0	0

(b) データ数帯 30(user_1649234)

$\hat{y}\backslash y$	1	2	3	4	5
1	0	0	1	1	0
2	0	1	6	4	1
3	0	0	1	12	0
4	0	0	0	3	0
5	0	0	0	0	0

(c) データ数帯 40(user_259931)

$\hat{y}\backslash y$	1	2	3	4	5
1	0	1	0	4	0
2	0	2	7	6	1
3	0	0	6	4	1
4	0	0	0	4	3
5	0	0	0	1	0

(d) データ数帯 100(user_400408)

$\hat{y}\backslash y$	1	2	3	4	5
1	0	0	0	0	0
2	0	0	6	23	2
3	0	0	0	17	1
4	0	0	4	30	0
5	0	0	0	16	1

表 B.4: 各データ数帯の要素がばらけているレビュアにおける満足度予測の各精度

User	**Metric**	**0.0**	**0.25**	**0.5**	**0.75**	**1.0**
データ数帯 20	Accuracy	0.09	0.18	0.278	0.51	0.64
	p 値		0.037*	0.034*	0.041*	0.049*
	RMSE	1.299	1.169	1.022	0.960	0.769
	p 値		0.070	0.107	0.2757	0.103
データ数帯 30	Accuracy	0.440	0.580	0.633	0.667	0.653
	p 値		0.017*	0.120	0.2726	0.541
	RMSE	0.963	0.878	0.807	0.806	0.745
	p 値		0.165	0.223	0.992	0.2724
データ数帯 40	Accuracy	0.410	0.2765	0.2765	0.2765	0.270
	p 値		0.088	1.000	1.000	0.060
	RMSE	1.157	1.149	1.137	1.144	1.199
	p 値		0.820	0.780	0.905	0.188
データ数帯 100	Accuracy	0.2712	0.2732	0.2764	0.2708	0.290
	p 値		0.426	0.219	0.059	0.472
	RMSE	0.997	0.925	0.882	0.942	0.929
	p 値		0.028*	0.184	0.006**	0.663

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

付録 C　実験 2

参考データ選択手法によって選ばれるレビュアの例を表 C.1 から表 C.13 にかけて述べる．表 C.1 は，実際に本実験で用いた推論対象レビュアの一例である．表 C.2 から表 C.7 は，各参考データ選択手法によって計算されたレビュアのうち最も類似しているレビュアの混同行列である．表 C.8 から表 C.13 は，各参考データ選択手法によって計算されたレビュアのうち最も類似していないレビュアの混同行列である．

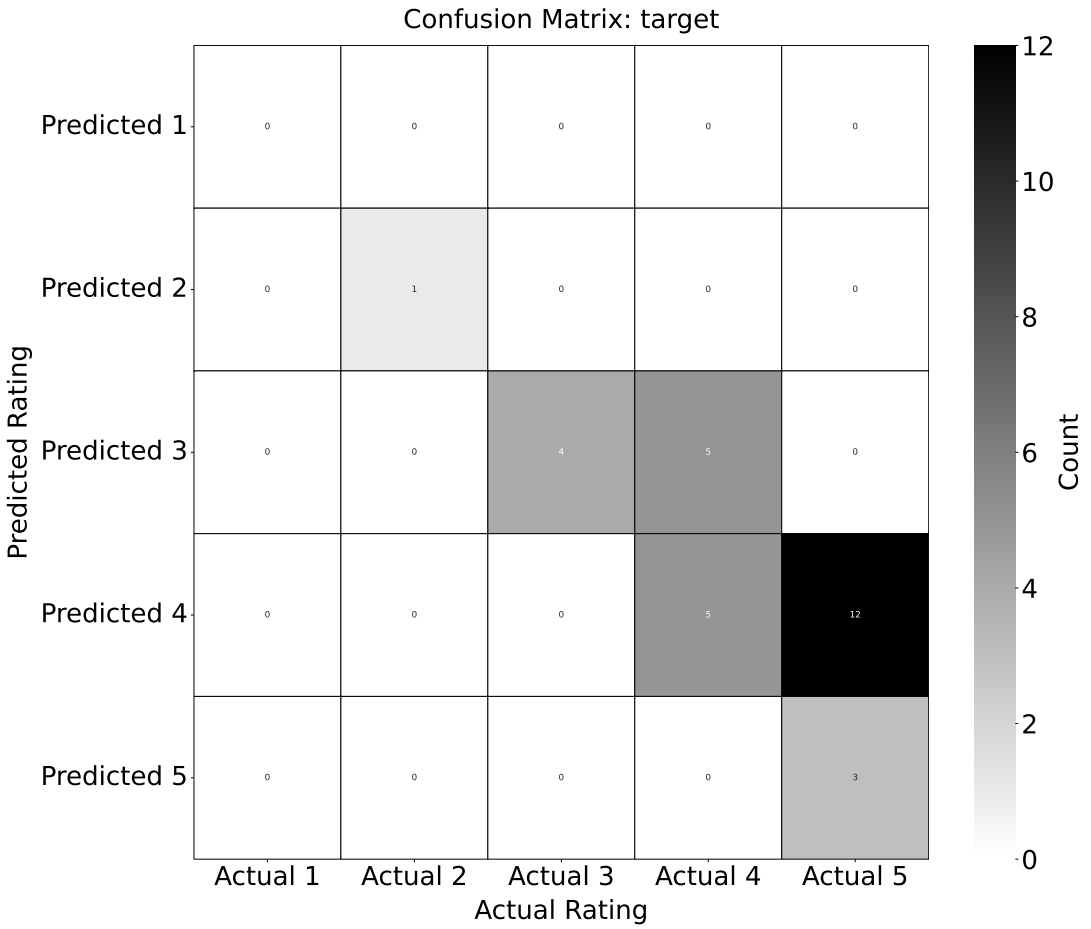

図 C.1: 推論対象レビュアの一例

また，実験結果において各データ数帯による精度を示したグラフを図 C.14 から図 C.21 にかけて示す．図 C.14 はデータ数帯が 20 のレビュア 3 人に対し，四種類の参考データ選択方針に従って満足度予測を実施した際の Accuracy を比較した結果である．図 C.15 はデータ数帯が 20 のレビュア 3 人に対し，四種類の参考データ選択方針に従って満足度予測を実施した際の RMSELoss を比較した結果である．図 C.16 はデータ数帯が 30 のレビュア 3 人に

対し，四種類の参考データ選択方針に従って満足度予測を実施した際の Accuracy を比較した結果である．図 C.17 はデータ数帯が 30 のレビュア 3 人に対し，四種類の参考データ選択方針に従って満足度予測を実施した際の RMSELoss を比較した結果である．図 C.18 はデータ数帯が 40 のレビュア 3 人に対し，四種類の参考データ選択方針に従って満足度予測を実施した際の Accuracy を比較した結果である．図 C.19 はデータ数帯が 40 のレビュア 3 人に対し，四種類の参考データ選択方針に従って満足度予測を実施した際の RMSELoss を比較した結果である．図 C.20 はデータ数帯が 100 のレビュア 3 人に対し，四種類の参考データ選択方針に従って満足度予測を実施した際の Accuracy を比較した結果である．図 C.21 はデータ数帯が 100 のレビュア 3 人に対し，四種類の参考データ選択方針に従って満足度予測を実施した際の RMSELoss を比較した結果である．

all_users_average における全一対比較結果を表 C.1 から表 C.4 にかけて示す．

表 C.1: same と他手法の統計的有意差比較 (all_users_average)

比較	Accuracy			RMSE		
	平均差	p 値	有意差	平均差	p 値	有意差
same vs. 類似度ベース手法						
same vs. cosine	0.0624	0.006509	**	0.1920	0.295353	-
same vs. pms-ssim	0.0501	0.014393	*	0.1199	0.037924	*
same vs. ssim	0.0846	0.001463	**	-0.1788	0.012833	*
same vs. block-ssim	0.0758	0.002961	**	0.1401	0.028734	*
same vs. com-ssim	0.0911	0.001044	**	0.1583	0.000316	***
same vs. com-cosine	0.0913	0.021350	*	0.1399	0.008595	**
same vs. random						
same vs. random	0.1153	0.000685	***	0.1947	0.001482	**
same vs. alter 系列						
same vs. alter cosine	0.3051	0.000003	***	0.3804	0.000003	***
same vs. alter pms-ssim	0.3088	0.000001	***	0.3823	0.000003	***
same vs. alter ssim	0.3182	0.000001	***	0.4002	0.000000	***
same vs. alter block-ssim	0.2574	0.000000	***	0.3300	0.000001	***
same vs. alter com-ssim	0.2122	0.000010	***	0.3417	0.000001	***
same vs. alter com-cosine	0.2081	0.000005	***	0.3277	0.000000	***

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, -: 有意差なし

平均差 = 各精度指標の平均差 (前者 - 後者)

表 C.2: cosine 系列と他手法の統計的有意差比較 (all_users_average)

比較	Accuracy			RMSE		
	平均差	p 値	有意差	平均差	p 値	有意差
cosine 系列 vs. ssim 系列						
cosine vs. pms-ssim	-0.0087	0.079441	-	-0.0015	0.776508	-
cosine vs. ssim	0.0109	0.091272	-	0.0096	0.016732	*
cosine vs. block-ssim	-0.0133	0.034583	*	-0.0130	0.438365	-
cosine vs. com-ssim	0.0280	<0.0001	***	-0.0081	<0.0001	***
com-cosine vs. pms-ssim	-0.0203	0.011313	*	-0.0126	0.055360	-
com-cosine vs. ssim	0.0112	0.103009	-	0.0190	0.347054	-
com-cosine vs. block-ssim	-0.0032	0.664859	-	-0.0012	0.856179	-
com-cosine vs. com-ssim	0.0253	0.010559	*	0.0258	0.005359	**
cosine 系列 vs. random						
cosine vs. random	0.0495	<0.0001	***	-0.0821	<0.0001	***
com-cosine vs. random	0.0274	0.002503	**	-0.0343	0.027109	*
cosine 系列 vs. alter 系列						
cosine vs. alter cosine	0.2274	0.000000	***	0.1506	0.000000	***
cosine vs. alter pms-ssim	0.2207	0.000000	***	0.1285	0.002293	**
cosine vs. alter ssim	0.2376	0.000000	***	0.1259	0.002293	**
cosine vs. alter block-ssim	0.1662	0.000000	***	0.0954	0.000004	***
cosine vs. alter com-ssim	0.1441	0.000000	***	0.1173	0.000001	***
cosine vs. alter com-cosine	0.1366	0.000000	***	0.0978	0.000006	***
com-cosine vs. alter cosine	0.2053	0.000000	***	0.1813	0.000000	***
com-cosine vs. alter pms-ssim	0.1986	0.000000	***	0.1876	0.000000	***
com-cosine vs. alter ssim	0.2155	0.000000	***	0.2021	0.000000	***
com-cosine vs. alter block-ssim	0.1441	0.000000	***	0.1260	0.000000	***
com-cosine vs. alter com-ssim	0.1220	0.000000	***	0.1479	0.000001	***
com-cosine vs. alter com-cosine	0.1145	0.000000	***	0.1285	0.000000	***

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, -: 有意差なし

平均差 = 各精度指標の平均差 (前者 - 後者)

表 C.3: ssim 系列と他手法の統計的有意差比較 (all_users_average)

比較	Accuracy			RMSE		
	平均差	p 値	有意差	平均差	p 値	有意差
ssim 系列同士の比較						
pms-ssim vs. ssim	0.0196	0.008415	**	-0.0081	0.040967	*
pms-ssim vs. block-ssim	0.0220	0.007684	**	0.0226	0.025479	*
pms-ssim vs. com-ssim	0.0367	0.000002	***	0.0311	0.009931	**
ssim vs. block-ssim	-0.0024	0.731862	-	-0.0033	0.435206	-
ssim vs. com-ssim	0.0171	0.008162	**	0.0482	0.032066	*
block-ssim vs. com-ssim	0.0147	0.015652	*	0.0270	0.000149	***
ssim 系列 vs. random						
pms-ssim vs. random	0.0582	<0.0001	***	-0.0359	<0.0001	***
ssim vs. random	0.0386	0.000020	***	-0.0650	<0.0001	***
block-ssim vs. random	0.0362	0.000197	***	-0.0133	0.377819	-
com-ssim vs. random	0.0216	0.002786	**	-0.0422	<0.0001	***
ssim 系列 vs. alter 系列						
pms-ssim vs. alter cosine	0.2361	0.000000	***	0.1828	0.000000	***
pms-ssim vs. alter pms-ssim	0.2294	0.000000	***	0.1892	0.000000	***
pms-ssim vs. alter ssim	0.2463	0.000000	***	0.2037	0.000000	***
pms-ssim vs. alter block-ssim	0.1749	0.000000	***	0.1276	0.000007	***
pms-ssim vs. alter com-ssim	0.1528	0.000000	***	0.1495	0.000001	***
pms-ssim vs. alter com-cosine	0.1453	0.000000	***	0.1300	0.000004	***
ssim vs. alter cosine	0.2165	0.000000	***	0.1679	0.000000	***
ssim vs. alter pms-ssim	0.2098	0.000000	***	0.1743	0.000000	***
ssim vs. alter ssim	0.2267	0.000000	***	0.1888	0.000000	***
ssim vs. alter block-ssim	0.1553	0.000000	***	0.1127	0.000000	***
ssim vs. alter com-ssim	0.1332	0.000000	***	0.1346	0.000000	***
ssim vs. alter com-cosine	0.1258	0.000000	***	0.1151	0.000000	***
block-ssim vs. alter cosine	0.2173	0.000000	***	0.2030	0.000000	***
block-ssim vs. alter pms-ssim	0.2074	0.000000	***	0.1666	0.000000	***
block-ssim vs. alter ssim	0.2243	0.000000	***	0.1811	0.000000	***
block-ssim vs. alter block-ssim	0.1529	0.000000	***	0.1050	0.000134	***
block-ssim vs. alter com-ssim	0.1308	0.000000	***	0.1269	0.000013	***
block-ssim vs. alter com-cosine	0.1233	0.000000	***	0.1074	0.000101	***
com-ssim vs. alter cosine	0.1995	0.000000	***	0.1517	0.000000	***
com-ssim vs. alter pms-ssim	0.1927	0.000000	***	0.1581	0.000000	***
com-ssim vs. alter ssim	0.2097	0.000000	***	0.1726	0.000000	***
com-ssim vs. alter block-ssim	0.1382	0.000000	***	0.0965	0.000076	***
com-ssim vs. alter com-ssim	0.1161	0.000000	***	0.1184	0.000009	***
com-ssim vs. alter com-cosine	0.1087	0.000000	***	0.0988	0.000053	***

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, -: 有意差なし

平均差 = 各精度指標の平均差 (前者 - 後者)

表 C.4: random と alter 系列の統計的有意差比較 (all_users_average)

比較	Accuracy			RMSE		
	平均差	p 値	有意差	平均差	p 値	有意差
random vs. alter 系列						
random vs. alter cosine	0.1779	<0.0001	***	-0.1469	<0.0001	***
random vs. alter pms-ssim	0.1712	<0.0001	***	-0.1533	<0.0001	***
random vs. alter ssim	0.1881	<0.0001	***	-0.1678	<0.0001	***
random vs. alter block-ssim	0.1166	<0.0001	***	-0.0917	0.000042	***
random vs. alter com-ssim	0.0945	<0.0001	***	-0.1136	0.000005	***
random vs. alter com-cosine	0.0871	<0.0001	***	-0.0941	0.000051	***

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, -: 有意差なし

平均差 = 各精度指標の平均差 (前者 - 後者)

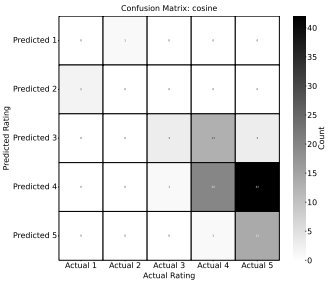

図 C.2: cosine テクニック

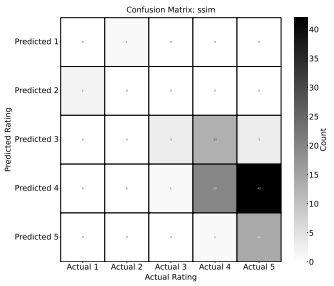

図 C.3: ssim テクニック

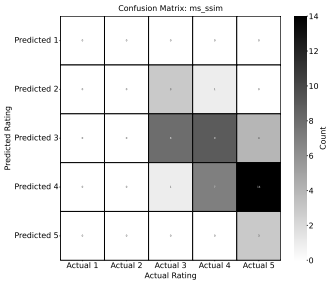

図 C.4: pms-ssim テクニック

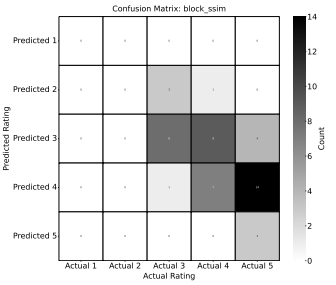

図 C.5: block-ssim テクニック

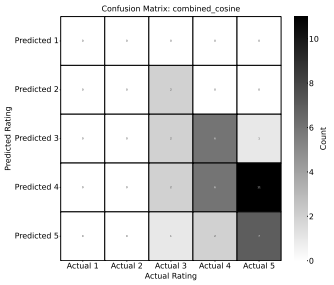

図 C.6: conbined-cosine テクニック

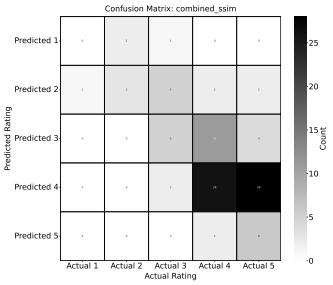

図 C.7: conbined-ssim テクニック

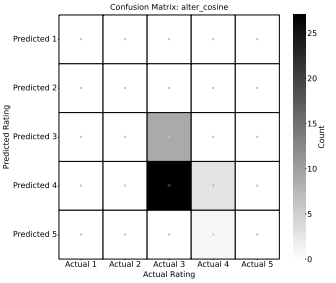

図 C.8: alter cosine テクニック

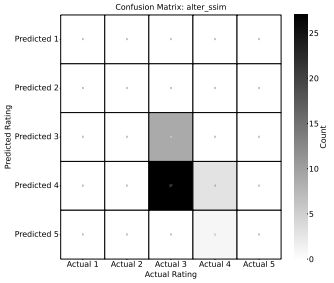

図 C.9: alter ssim テクニック

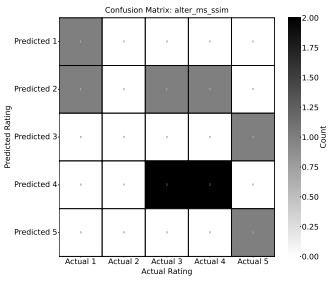

図 C.10: alter pms-ssim テクニック

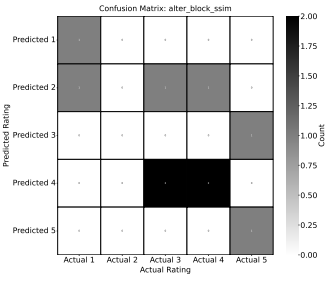

図 C.11: alter block-ssim テクニック

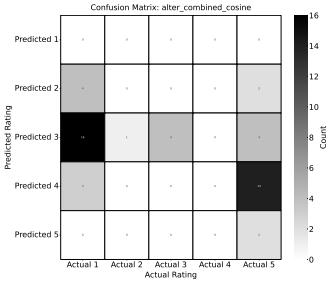

図 C.12: alter conbined-cosine テクニック

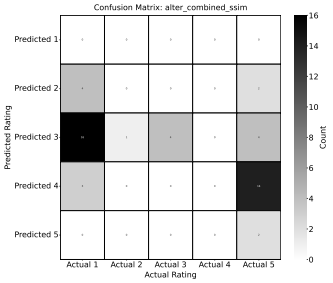

図 C.13: alter conbined-ssim テクニック

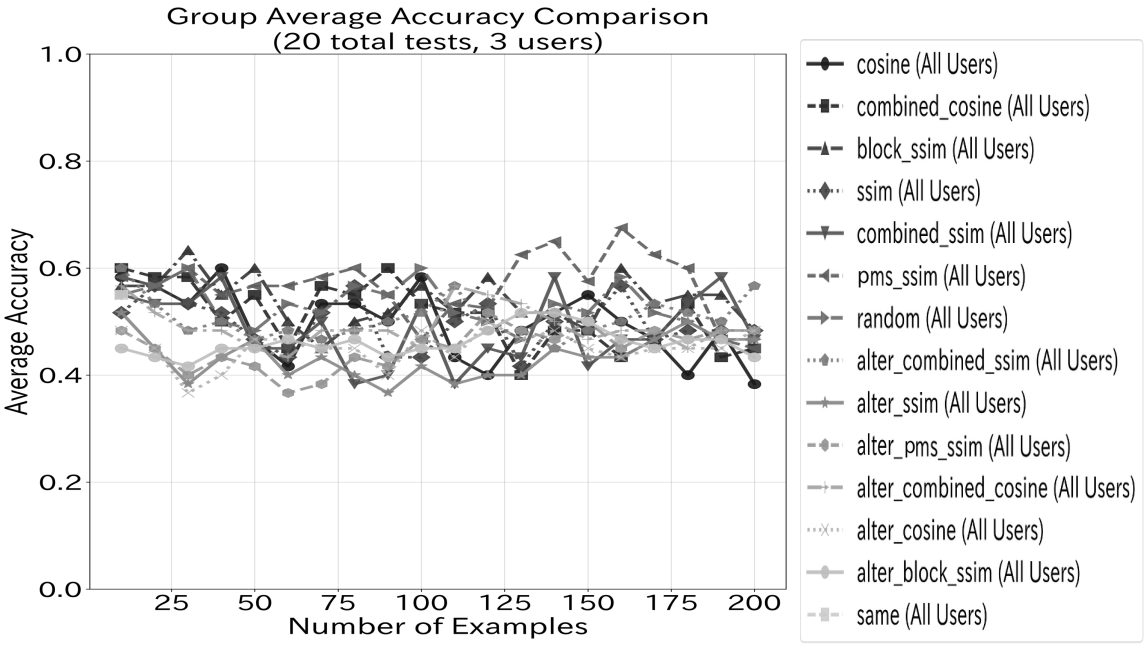

図 C.14: データ数帯 20 のレビュアが推論対象である時の各参考データ選択手法の精度比較 (Accuracy)

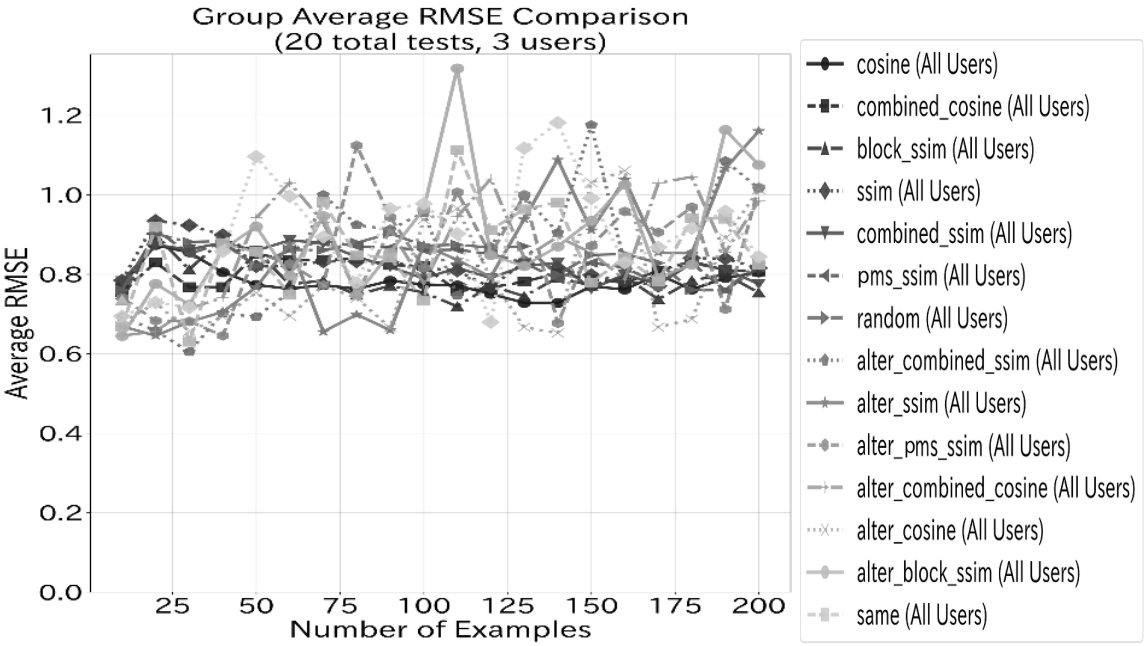

図 C.15: データ数帯 20 のレビュアが推論対象である時の各参考データ選択手法の精度比較 (RMSE)

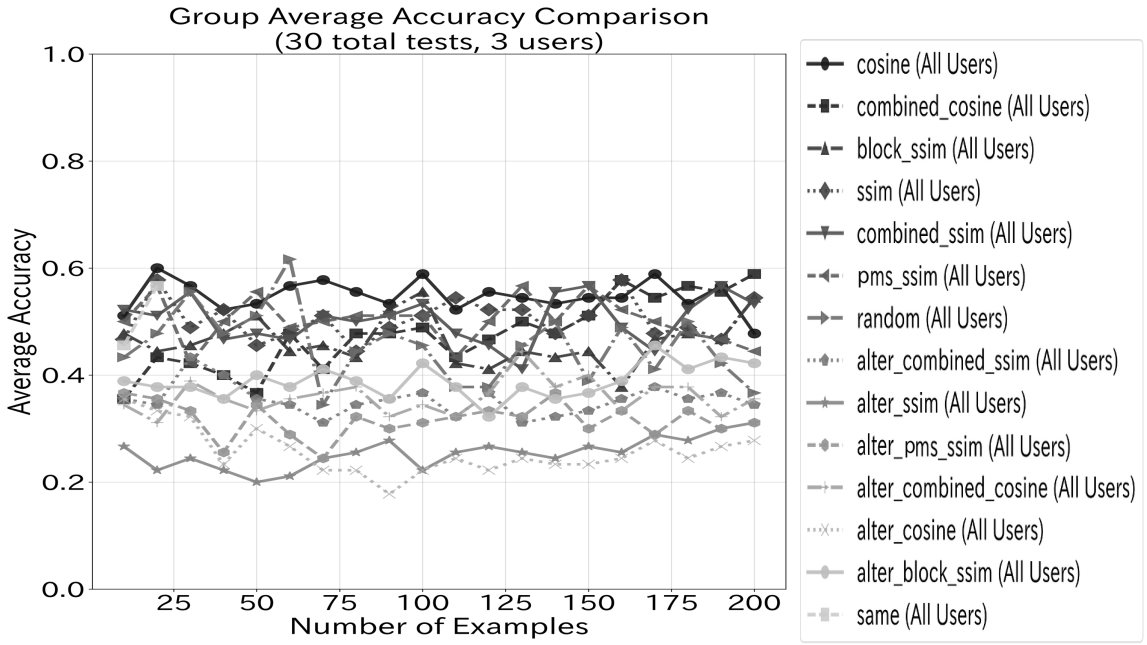

図 C.16: データ数帯 30 のレビュアが推論対象である時の各参考データ選択手法の精度比較 (Accuracy)

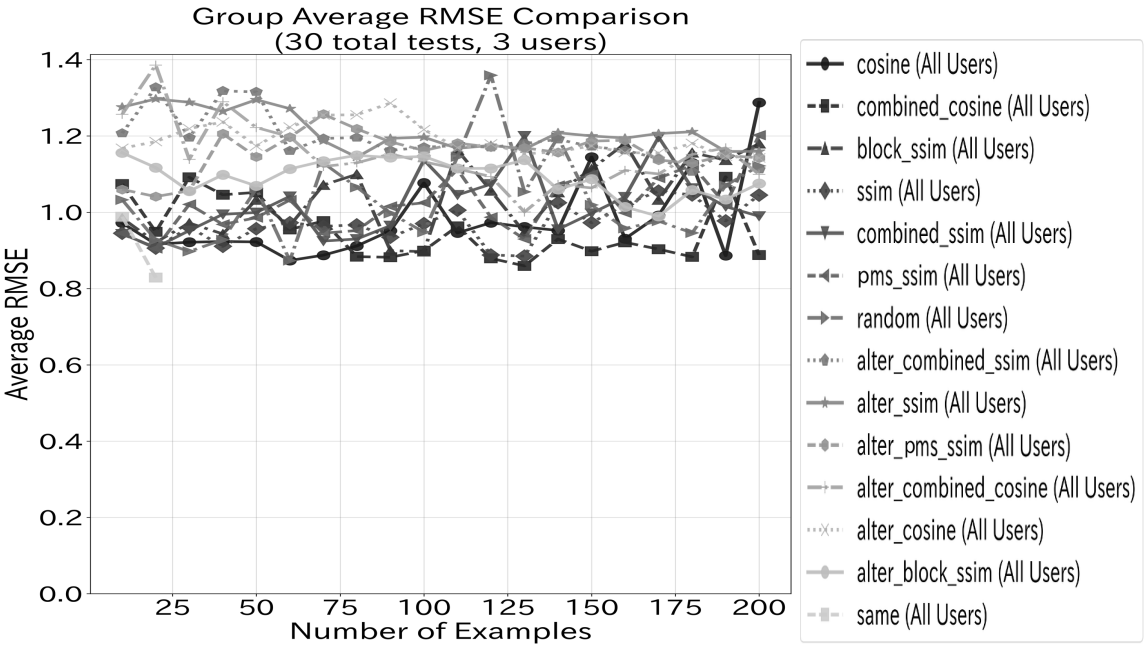

図 C.17: データ数帯 30 のレビュアが推論対象である時の各参考データ選択手法の精度比較 (RMSE)

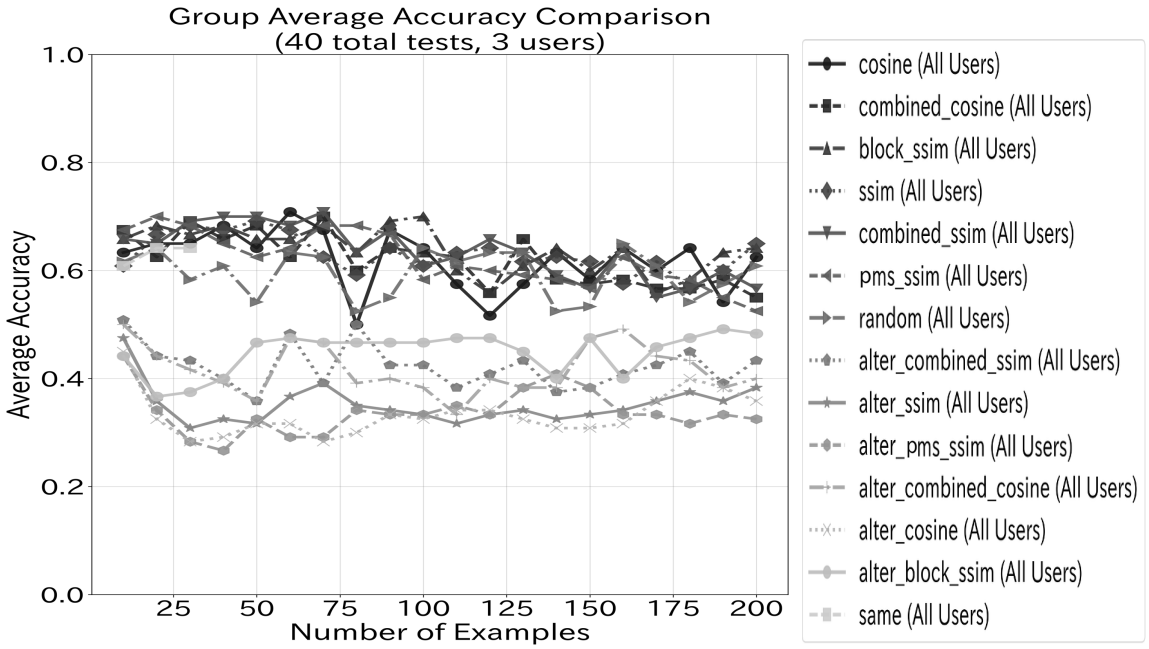

図 C.18: データ数帯 40 のレビュアが推論対象である時の各参考データ選択手法の精度比較 (Accuracy)

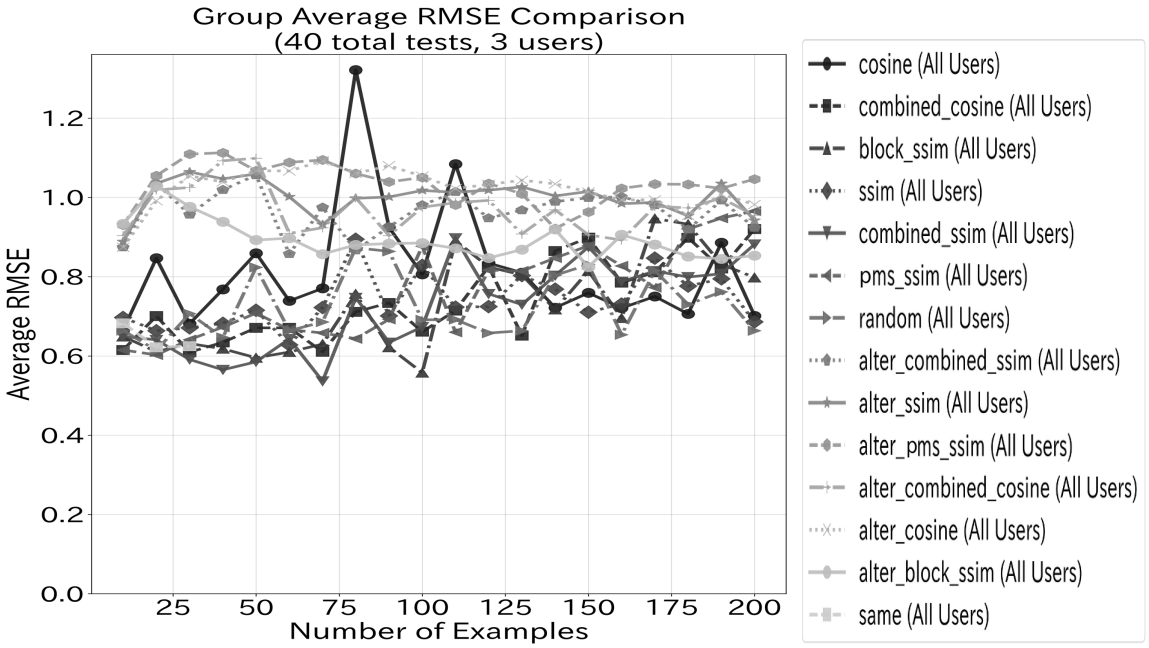

図 C.19: データ数帯 40 のレビュアが推論対象である時の各参考データ選択手法の精度比較 (RMSE)

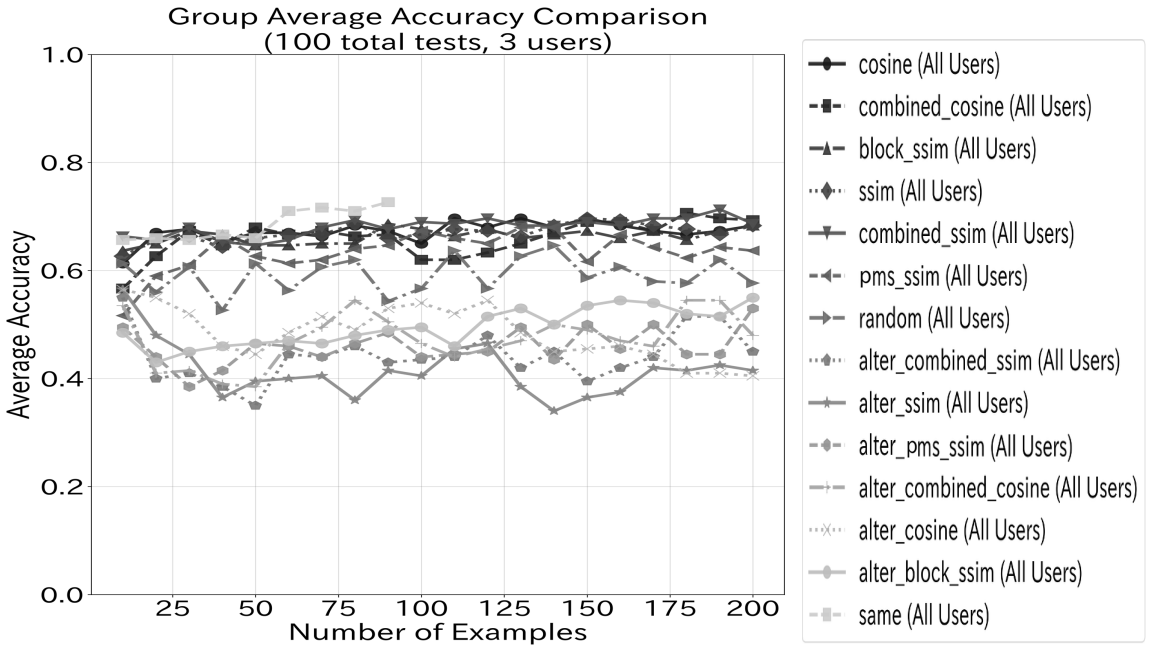

図 C.20: データ数帯 100 のレビュアが推論対象である時の各参考データ選択手法の精度比較 (Accuracy)

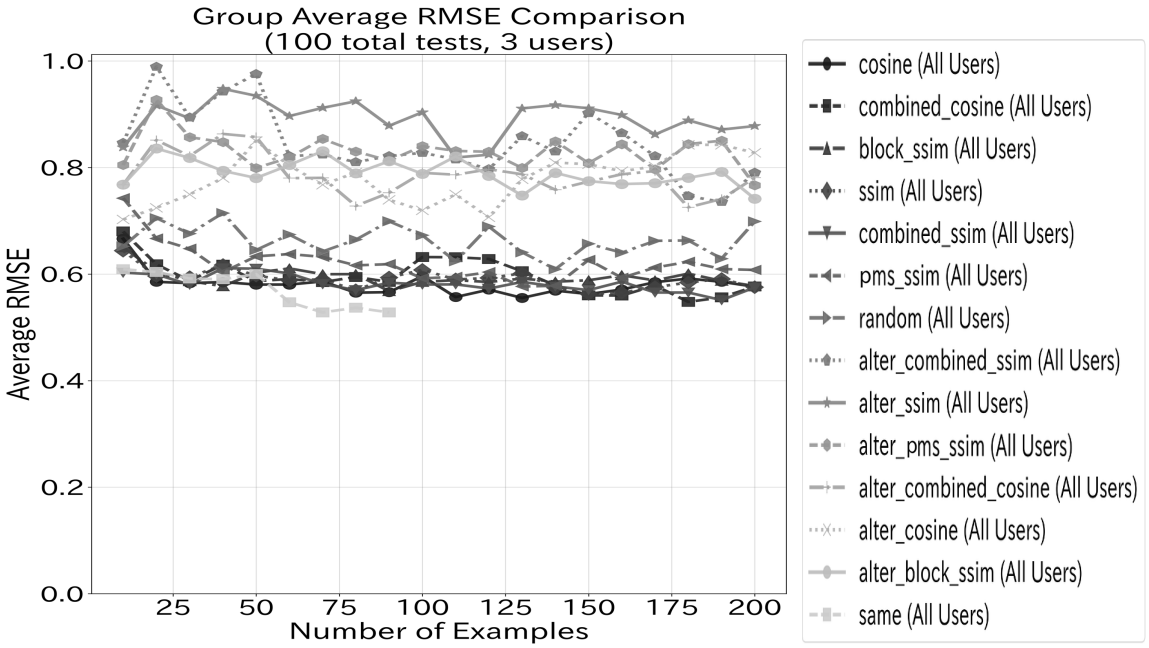

図 C.21: データ数帯 100 のレビュアが推論対象である時の各参考データ選択手法の精度比較 (RMSE)