大規模言語モデルを用いた商品比較のためのレビュー集約

中井香那子† 山本 岳洋† 大島 裕明†

† 兵庫県立大学 大学院情報科学研究科 〒 651–2197 兵庫県神戸市西区学園西町 8-2-1
E-mail: †ad24c041@guh.u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp, †††ohshima@ai.u-hyogo.ac.jp

あらまし 本研究では，大規模言語モデルを用いたレビュー集約による商品比較表の作成手法を提案する．比較表とは，2 つの商品の各観点に対してどのような評価をしているレビューが何件あるかを一覧することのできる表である．大規模言語モデルを用いて比較表を作成することで，オンデマンドで表を作成し，商品に応じて比較する観点を変更することができる．提案手法では，大規模言語モデルを用いてレビューから観点と評価を抽出し，類似している観点，評価のクラスタリングを行うことで観点ごとに評価を対応付けた表を作成する．楽天市場のレビューデータセットを用いて大規模言語モデルにより自動的にレビューを集約し，商品の比較表を作成し，商品の比較に有用であるかをユーザ実験を通して検証した．

キーワード 対照要約，LLM，文書要約

1 はじめに

商品を比較する際，レビューは重要な情報源である．しかし，商品によってはレビューが多数あり，レビューすべてを読むためには多くの時間が必要となる．また，ウェブページに表示されるレビューは様々な意見が順不同に並んでいたり，1 つのレビューの中に複数の評価が含まれていたりする．そのため，2 つの商品に共通する特徴や，片方のみに出現する特徴を見つけたり，共通する観点に対してどのような意見の違いがあるかを探したりするのは容易ではない．このように，レビューは重要な情報にもかかわらず，商品ページにあるレビューの形式は商品比較に適した形ではないという問題がある．

本研究では，商品の特徴を表す言葉を観点，観点に対して述べている文を評価とし，観点ごとに評価の集約を行う．また，複数商品間で同じ観点に対する評価を対応付けた商品の比較表を作成する．比較表の作成の際に，各評価が何件あるかという意見の分布を示すことで，どのような意見が多いかを一目で比較できるようにする．作成する比較表の例を図 1 に示す．例えば，図 1 の一番上の段は「ノイズキャンセリング」についてのレビューをまとめている．この表は，商品 A は「ノイズキャンセルが優秀」という内容のレビューが 10 件，「遮音性は微妙」という内容のレビューが 3 件あることを示している．同様に，商品 B については「性能は高くない」というレビューが 8 件，「性能に満足」という内容のレビューが 2 件あることを示している．また，各要約文の下は実際のレビューを示している．このことから，商品 A のほうが商品 B よりノイズキャンセリングについては好意的な意見が多いということがわかる．また，一番下の段は「充電残量の表示」についてのレビューをまとめている．商品 A には何も表示されず，商品 B については「充電残量の表示がわかりやすい」というレビューが 7 件あることが示されていることから，この特徴は商品 B のみの特徴であることがわかる．このような表を作成することで，商品を購入した人の意見を基に短時間で商品を比較することが可能になると考えられる．

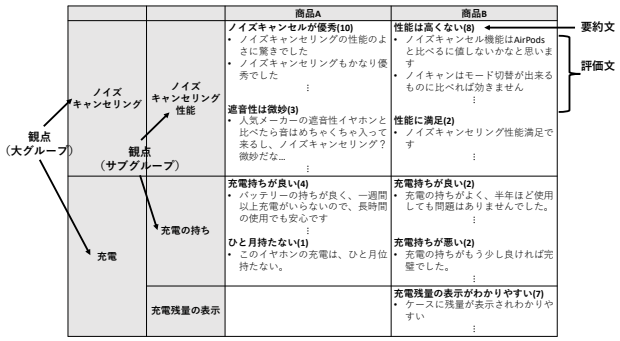

図 1 レビュー集約を行った比較表のイメージ図.

商品の比較表を作成するために，大規模言語モデルを用いてレビューを自動的に集約する．大規模言語モデルを用いるメリットは 2 点ある．1 点目はオンデマンドで表を作成できる点である．これにより製品ごとに表で比較する観点を変更することができる．例えば，同じイヤホンカテゴリの商品でも，イヤホン A と B を比較する場合は「音質，デザイン，サイズ」という観点，イヤホン C と D を比較する場合は「音質，ノイズキャンセリング，bluetooth 機能」という観点を用いるというようにレビューで言及される観点が違う場合に比較する観点を変更することが可能である．2 点目はドメインごとに学習する必要がない点である．これにより他のドメインへの応用が容易である．提案手法は，比較表を作成するためのモデルを商品ごとに学習しなおす必要がなく，複数ジャンルの商品の比較表の作成が可能である．

提案手法について，従来のレビュー表示と比較して商品の比較に有用であるかを検証するためにユーザ実験を行った．実験では「時間内に十分情報を得られるか」,「様々な観点について比較するのに役立つか」,「商品を比較するための観点を知るのに

役立つか」,「使いやすいか」の 4 つの項目でアンケートを行った．実験の結果，4 つの項目全てで提案手法が従来のレビュー表示を上回った．また，比較にかかった時間と「時間内に十分情報を得られるか」という項目では有意差が見られなかったが，他の 3 つの項目については有意差が認められた．

2 関連研究

2.1 複数文書要約

本研究は複数のレビューから比較のための表を作成する．これは，複数の文書を要約する複数文書要約と似たタスクである．Mehta らは複数文書要約において複数のシステムで要約を行い，異なるシステムの要約結果で重複する内容やパフォーマンスの高いシステムの要約内容を考慮することで要約の質を向上させる手法を提案している [9]．Iso らは 2 つのレビューセットから対比的な要約と共通の要約を生成する比較用要約フレームワーク COCOSUM を提案している [7]．Atri らは複数文書を要約する際のトピックの一貫性と文書間の関係性を強化することを目的に FABRIC というエンコーダ・デコーダモデルを提案している [2]．Yan らは大規模言語モデルの比較推論の精度を向上させるために中間的な比較表を生成する SC^2 というモデルを提案している [13]．Zhang らはユーザのニーズや好みに適合した要約を行うためにトピックに着目した TOMDS という手法を提案している [14]．Zhang らは性別や政党，感情などの社会属性に関して公平な要約を行うことを目的とし，公平性を評価する評価指標と，公平性を向上させるための方法を提案している [15]．本研究では複数のレビューを要約するが，比較に注目するという点で既存研究と異なる．

2.2 レビュー要約・感情分析

本研究は複数のレビューを観点ごとに集約する．これは，レビューを要約するタスクやレビューのポジティブ・ネガティブを判定するタスクと似たタスクである．Dave らはレビューの製品属性を特定し，肯定的なレビューと否定的なレビューを自動的に判別する手法を提案している [3]．Tripathy らは N-gram モデルと機械学習アルゴリズムを組み合わせレビューが肯定的であるか，否定的であるかを分類する手法を提案している [12]．Amplayo らはレビューと要約のペアの合成訓練データを作成し，各レビューの観点を様々な粒度で予測することで，特定の観点クエリに特化した要約を行う手法を提案している [1]．

2.3 対照要約

対照要約 (Contrastive Summarization) に関する研究が行われている．対照要約とは，2 つの文書などを比較し共通するテーマについて違いを強調した要約をすることである．

本研究は 2 つの商品のレビューを観点という共通テーマを基に，共通点や相違点に注目した要約を行うという点で対照要約の 1 種ととらえることができる．商品レビューを対象とした対照要約の研究として以下のような研究が行われている．Lerman らは消費者レビューを対象に対照要約を行い，単一商品の要約と比較して対照要約の有用性を示している [8]．Gunel らは事前学習済みの T5 モデルを使用してウェブページからアスペクトとそれに対応する値を抽出し，複数商品の比較ができるよう構造化された要約を行う STRUM という手法を提案している [5]．さらに大規模言語モデルを用いて属性ごとに複数のウェブページの内容を集約し，属性のランキングを行った商品の比較表を作成する STRUM-LLM というシステムを提案している [6]．STRUM-LLM も 2 つの商品を入力とすることで，その商品に関するレビューをウェブから取得し，比較表を作成することができるという点で，本研究と類似している．しかし，STRUM-LLM は手法の詳細が公開されておらず，再現することが難しい．また，本研究では観点を 2 段階の階層構造で表現するのに対して，STRUM-LLM では 1 段階の構造で表現している．

3 大規模言語モデルを用いたレビュー集約による比較表の作成

3.1 レビュー集約による比較表の作成

本研究では，大規模言語モデルを用いてレビューを集約し，商品の比較表を作成する．作成する比較表の例を図 1 に示す．ある商品に対してつけられたレビューを全て入力し，観点ごとに評価の内容を集約する．また，観点を基に 2 つの商品の集約結果を対応付け比較表を作成する．

レビューとはユーザが商品につけたコメントであり，加工前のデータを指す．また，**観点**をその商品の特徴を表す言葉，**評価**をレビュー中で観点に言及している部分と定義する．例えば，イヤホンの観点には「音質」,「ブルートゥース」,「ノイズキャンセリング」などがある．評価は「音質」という観点に対しての「音質が良い」という文や「ブルートゥース」という観点に対しての「ブルートゥース接続が簡単である」という文である．

比較表を作成するために，まずレビューから観点と評価を抽出する．次に，似ている観点を大きく分類した**大グループ**，各グループの中でさらに類似した観点を集めた**サブグループ**を作成する．その後，各レビューから抽出された観点と評価のペアを 1 文にまとめた**評価文**を作成する．さらに，サブグループ内で類似する評価文を集約し，集約した評価文を 1 つの文に要約する．ここで，要約した文を**要約文**と呼ぶ．最後に同じ観点サブグループに対応する要約文を対応付けて比較表を作成する．また，比較表には要約文の隣にその要約文に含まれる評価文の件数を示し，下に元の評価文を表示させる．例えば，図 1 の一番上段はノイズキャンセリングという観点大グループのうち，ノイズキャンセリング性能という観点サブグループに対する評価をまとめている．商品 A については,「ノイズキャンセルが優秀」という評価が 10 件,「遮音性は微妙」という評価が 3 件あることを示している．また,「ノイズキャンセリングが優秀」という要約文には「ノイズキャンセリングの性能の良さに驚きでした」や「ノイズキャンセリングもかなり優秀でした」といった評価文が含まれることを示している．

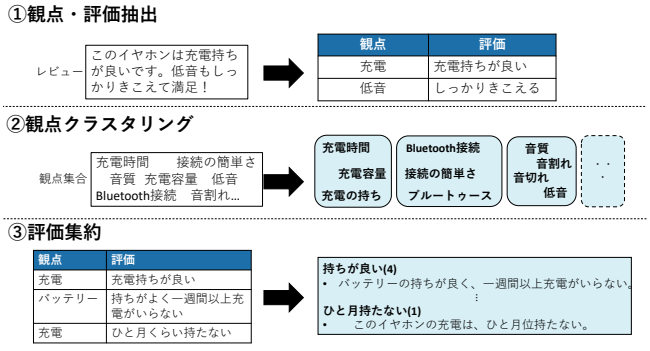

図 2 提案手法の概要.

3.2 提案手法の概要

提案手法の概要を図 2 に示し，以下で概要を説明する．

1. 観点・評価の抽出
   - レビューを大規模言語モデルに入力し，観点とそれに対する評価を抽出する．
2. 観点クラスタリング
   - 観点集合を大規模言語モデルに入力し，類似している観点をまとめる観点のクラスタリングを行う．
3. 評価集約
   - ある観点グループ内の評価文を大規模言語モデルに入力し，類似評価文を 1 文に要約する．
   - 2 つの商品に共通する観点を基に要約文を対応付ける．

以降，それぞれのステップの詳細について説明する．なお，説明にあたり 4.1 節で述べるイヤホンカテゴリの 2 商品とその商品に対するレビューを Llama-3.1-Swallow-8B-Instruct-v0.1 に入力して得られた結果を例に説明する．

3.3 観点・評価の抽出

レビューを大規模言語モデルに入力し，レビューから観点とそれに対応する評価を抽出する．以下は，「数回使用して残念ながら故障しました．値段が値段だけに仕方無いと思います.」というレビューを入力する場合のプロンプトである．

観点・評価を抽出するプロンプト

ユーザーから与えられたレビューを基に，イヤホンの特徴を表す観点と具体的にその観点に言及している評価を抜き出してください．
評価はレビューに書かれている内容をそのまま抜き出してください．
レビュー：**数回使用して残念ながら故障しました．値段が値段だけに仕方無いと思います．**

「数回使用して残念ながら故障しました．値段が値段だけに仕方無いと思います.」というレビューからは「耐久性」という観点と，「数回使用して残念ながら故障しました」という評価が抽出できた．また，イヤホンレビューを入力した際，「価格，遮断性，音遅延，防水性能」などが観点として抽出された．

3.4 観点クラスタリング

3.3 節で得られた観点を大規模言語モデルに入力し，類似している観点を 1 つのグループにまとめる観点集約を行う．ここでは観点を商品ごとに分けず，2 つの商品レビューから抽出された全ての観点を 1 つの集合として入力する．類似している観点とは「ブルートゥース性能」と「Bluetooth 機能」のように言い換えられた観点や，「音質」と「高音質」のように含意関係にある観点を指す．以下は，「遮断性，耐水性，充電時間，ペアリングの簡単さ」など入力する観点集合の一部を抜粋したプロンプトである．また，以下のプロンプトに json 形式で出力するよう指示を加えた．

観点クラスタリングを行うプロンプト

与えられた観点を基に似ている観点をグループ分けしてください．
各観点は 1 つのグループのみに入れてください．
観点：[**遮断性，耐水性，充電時間，ペアリングの簡単さ，…**]

この結果，「音質，フィット感，耐水性，操作性，充電とバッテリー，デザインと外観，接続性，音量と音量調整，防水性と耐水性，携帯用と持ち運び，価格とコスパ」の 11 グループに分けられた．例えば，「音質」グループには「高音質，低音，音遅延，重低音」などの観点が含まれる．

全ての観点を入力したが，出力された観点グループに分類されていない観点が存在する．そのため，どのグループにも分類されなかった観点については 1 つずつ入力し，再分類を行う．分類先となる観点グループは上記で得られた 11 グループのみである．以下は，「通話の品質」という観点を「音質，フィット感，耐水性，操作性，充電とバッテリー，デザインと外観，接続性，音量と音量調整，防水性と耐水性，携帯用と持ち運び，価格とコスパ」のいずれかのグループに分類する場合のプロンプトである．

観点をグループに分類するプロンプト

観点をいずれかのグループに分類したいです．
分類先は [**音質，フィット感，耐水性，操作性，充電とバッテリー，デザインと外観，接続性，音量と音量調整，防水性と耐水性，携帯用と持ち運び，価格とコスパ**] のいずれかです．分類されるグループのキーを 1 つだけ出力してください．
分類される可能性のあるグループが 2 つ以上ある場合も 1 つだけ出力してください．
出力は分類先のキーのみとしてください．
分類する観点: **通話の品質**

「通話の品質」という観点は「音質」グループに分類された．同様に，「音，重低音域，ライブ感，音作り，外の音の遮断性」などの観点が「音質」グループに追加された．ここまでで得られた観点グループを，大グループとする．

大グループでは観点を大まかに分けており，各グループの中

には詳細には異なる観点のグループが複数存在する．例えばイヤホンでは，「音質」というグループの中の「音飛び，音遅延」や「外の音の遮断性，ノイズキャンセリング，外部ノイズ遮断」というグループは「音質」に関連するがそれぞれのグループは異なる観点グループといえる．そこで，大グループの各グループの観点を対象に，再度上記と同様の手順で，類似している観点を 1 つのグループにまとめた．大グループをさらに分類することで得られたグループをサブグループとする．例えば，大グループのうち「音質」というグループは「音質，機能性，付属品，デザイン，その他」の 5 つのサブグループに分けられる．また，「音質（大グループ）-音質（サブグループ）」には「高音質，低音，重低音，中音，音遅延，音の良さ」といった観点が含まれる．

3.5 評価集約

3.4 節で得られたサブグループの各グループごとに類似評価の集約を行う．評価集約では，各観点グループの中にどのような意見があるのかを簡潔に示すために似ている評価を 1 文に要約する．しかし，評価のみを使用して要約すると情報が不足する場合がある．例えば，「音質」グループのある観点に対して「いいもの」という評価が抽出された場合，「高音」に対する「いいもの」という評価と「低音」に対する「いいもの」という評価は 1 つにまとめるべきではない．そこで評価を集約するために，観点と評価を 1 文にまとめた擬似的なレビュー文を作成した．以下は，「外音の遮断性」という観点，「とても良いためウォーキング時は周りへの注意が必要だと思います」という評価を入力する場合のプロンプトである．

観点と評価を 1 文にまとめるプロンプト

以下はイヤホンのレビューから抽出した観点と評価です.
観点: **外音の遮断性**
評価: **とても良いためウォーキング時は周りへの注意が必要だと思います**
観点と評価をつなげて日本語として自然な 1 文を作ってください.
出力は作成した 1 文のみとしてください.

上記の入力からは「外音の遮断性が高いので，ウォーキング時は周りへの注意が必要だと思います.」という出力結果が得られた．ここで得られた観点と評価をまとめた 1 文を評価文とする．

次に，同じサブグループ内の評価文を 2 商品共すべて入力し，同じ主張をしている評価文を 1 つのグループにまとめる．以下は「音質（大グループ）-音質（サブグループ）」の評価文「“動画鑑賞時の音遅延も気にならないレベルです．”，“このイヤホンの重低音は圧倒的にこちらの好みです．”，“このイヤホンは，想像以上の高音質を実現しています．”，“音の良さが想像以上だった．”，…」を入力する場合のプロンプトである．

類似評価文を集約するプロンプト

以下はあるイヤホンに対するレビューから評価を抽出したものです.
評価リスト:[**“動画鑑賞時の音遅延も気にならないレベルです，“このイヤホンの重低音は圧倒的にこちらの好みです．”，“このイヤホンは，想像以上の高音質を実現しています．”，“音の良さが想像以上だった．”，**…]
評価リスト内で同じ主張をしているレビューを 1 つのグループにまとめてください.
ポジティブネガティブが逆のレビューは別々のグループに分けてください.
また，グループ名とグループに含まれるレビューを一緒に出力してください.
類似レビューがない場合は, 各文を表す単語と元のレビューを一緒に出力してください.

その結果，「低音が素晴らしい，中音が素晴らしい，高音質，その他」の 4 グループに分けられた．また，「低音が素晴らしい」というグループには「“このイヤホンの重低音は圧倒的にこちらの好みです”，“このイヤホンは，低音域の音質にやや欠けている印象を受けました．”，“低音は多少エンジン音で消されますが，低音が良く，大変満足しました．”」など低音に関する評価文が含まれる．ここで得られた類似評価文をまとめたグループを類似評価グループとする．

次に，商品ごとに 1 つ 1 つ評価文を入力し，類似評価グループのうちどのグループに含まれるか分類を行う．以下は商品 A の「音質（大グループ）-音質（サブグループ）」に含まれる評価文「このイヤホンの重低音は圧倒的にこちらの好みです.」を「低音が素晴らしい，中音が素晴らしい，高音質，その他」のいずれかに分類する場合のプロンプトである．

評価文をグループに分類するプロンプト

レビューを以下のグループに分類したいです.
グループ：**低音が素晴らしい:** [**“このイヤホンの重低音は圧倒的にこちらの好みです．”，**…]**，中音が素晴らしい:** [**“このイヤホンの中音は文句なしに素晴らしい．”**]**， 高音質:** [**“このイヤホンは，想像以上の高音質を実現しています．”，**…]**，その他:** [**“動画鑑賞時の音遅延も気にならないレベルです．”**]
分類先は [**低音が素晴らしい，中音が素晴らしい，高音質，その他**] のいずれかです．分類されるグループのキーを 1 つだけ出力してください.
分類される可能性のあるグループが 2 つ以上ある場合も 1 つだけ出力してください.
出力は分類先のキーのみとしてください.
分類するレビュー: **このイヤホンの重低音は圧倒的にこちらの好みです．**

この結果，「このイヤホンの重低音は圧倒的にこちらの好みです.」という評価文は「音質（大グループ）-音質（サブグループ）-低音が素晴らしい（類似評価グループ）」に分類された．

また，商品 A と B の「音質（大グループ）-音質（サブグループ）」グループ内の評価文を類似評価グループに分類した結果をそれぞれ表 1，表 2 に示す．

次に，各商品の類似評価グループごとに，評価文の要約を行う．以下は商品 B の「音質（大グループ）-音質（サブグループ）-低音が素晴らしい（類似評価グループ）」内の評価文「“低音は 5800 円の神スピーカーよりも多少軽めですが，全体的な音質が良く，価格を考えると非常にコストパフォーマンスが高いと感じました．”, “このイヤホンは，低音が文句なしに素晴らしい．”, “低音が良く，大変満足しました．”」を入力する場合のプロンプトである．

類似評価文を要約するプロンプト

以下はイヤホンのレビューです．
レビュー:[**“低音は 5800 円の神スピーカーよりも多少軽めですが，全体的な音質が良く，価格を考えると非常にコストパフォーマンスが高いと感じました．”, “このイヤホンは，低音が文句なしに素晴らしい．”, “低音が良く，大変満足しました．”**]
レビューを 1 文に要約してください．
出力は要約結果のみにしてください．

この結果，「低音が良く，全体的な音質も高く，価格を考えると非常にコストパフォーマンスが高い．」という 1 文に要約された．ここで得られた，類似評価文の要約結果を要約文とする．

最後に，各要約文に結びついている観点グループを用いて，2 商品のレビューの集約結果を対応付ける．例えば，「音質（大グループ）-音質（サブグループ）-低音が素晴らしい（類似評価グループ）」というグループの要約文は，商品 A が「このイヤホンの重低音は非常に気に入りました．」であり，商品 B が「低音が良く，全体的な音質も高く，価格を考えると非常にコストパフォーマンスが高い．」である．この 2 文は，比較表の対応する評価の要約文となる．商品 A と B のレビューの集約結果の一部を表 3 に示す．

3.6 比 較 表

最後に，要約文と評価文，実レビューを観点を基に紐づけた比較表のプロトタイプシステムを作成した．ユーザ側の動作例を図 3 に示す．まず，最初の画面では各観点に対応する要約文とその要約文に含まれる評価文の件数を表示する．次に，各要約文をクリックすることで要約文の元になっている評価文を表示する．最後に各評価文にマウスポインタを合わせることで評価文の内容を含む実際のレビューを表示する．

4 実 験

実験として，観点クラスタリングと評価集約の精度を大規模言語モデルごとに測る実験と，作成した比較表が商品を比較するタスクにおいて有効であるかを検証するためのユーザ実験の 2 つを行った．

4.1 大規模言語モデルの比較実験

観点クラスタリングと評価集約について 2 つの大規模言語モデルを比較しモデルによる差があるかを検証した．

4.1.1 デ ー タ

実験に用いたデータについて説明する．商品レビューについては楽天データセットに含まれる楽天市場の商品レビューデータ [16] を利用した．このデータセットには，商品名，商品 ID，商品ジャンル ID，レビュータイトル，レビュー内容などが含まれる．比較表の対象商品は，商品ジャンル ID が同一である 2 商品とする．また，レビュー集約は各商品 100 件のレビュー内容を使用した．実験には商品ジャンルがイヤホンである商品から 2 つの商品を選択した．この 2 つの商品をそれぞれ商品 A，商品 B と呼ぶ．

4.1.2 大規模言語モデル

実験には，Llama-3.1-Swallow-8B-Instruct-v0.1(Swallow) [1] と gemma-2-9b(gemma) [11] を大規模言語モデルとして使用する．Llama-3.1-Swallow-8B-Instruct-v0.1 は Llama3.1 ベースに日本語と英語のデータセットによって継続事前学習を行ったモデルである [4] [10]．実験では，8bit 量子化したモデルを使用した．gemma-2-9b は Google が提供する Gemini モデルの作成に使用された研究と技術に基づいて構築されているモデルである．実験では，4bit 量子化したモデルを使用した．また，全てのモデルで temperature は 0 と設定した．

4.1.3 観点クラスタリングの評価方法

観点クラスタリングの評価方法について述べる．クラスタリング結果を評価するために，評価尺度として Purity を使用する．Purity とは生成されたクラスタ内で最も多いラベルの割合を測り，クラスタの純度を示す．全体で観点が N 個，生成されたクラスタが L 個，生成された i 番目のクラスタにおいて，j というクラスに割り当てられるデータが $n_{i,j}$ 個あるとすると，Purity は以下の式で求めることができる．

$$\text{Purity} = \frac{1}{N}\sum_{i=1}^{L}\max_j(n_{i,j})$$

例えば，「音質（大グループ）」に「高音質，低音，重低音，Bluetooth 機能，音の良さ，包装」という観点が含まれている場合，「高音質，低音，重低音，音飛び」，「Bluetooth 機能」，「包装」の 3 つのクラスに分けられる．この中で最も観点が多いのは「高音質，低音，重低音，音飛び」クラスであり観点は 4 つである．さらに，生成されたクラスタが 2 つあり，もう 1 つのクラスタには 10 個の観点，最もデータの多いクラスに 7 個の観点が割り当てられた場合，Purity は 0.688 である．Purity を用いた評価は，大グループ，サブグループそれぞれで同様の評価を行い，2 つの大規模言語モデルで結果の比較を行う．

4.1.4 評価集約の評価方法

評価集約の評価方法について述べる．評価集約を評価するために，要約文と要約に使用された元の評価文の一致度を測る．要約文の内容と一致する評価文を正解，一致しない評価文を不

1：https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.1

表 1　商品 A の「音質（大グループ）-音質（サブグループ）」の分類結果.

類似評価グループ	レビュー
低音が素晴らしい	このイヤホンの重低音は圧倒的にこちらの好みです.
その他	動画鑑賞時の音遅延も気にならないレベルです.

表 2　商品 B の「音質（大グループ）-音質（サブグループ）」の分類結果.

類似評価グループ	レビュー
低音が素晴らしい	低音は 5800 円の神スピーカーよりも多少軽めですが，全体的な音質が良く，価格を考えると非常にコストパフォーマンスが高いと感じました. このイヤホンは，低音が文句なしに素晴らしい. 低音が良く，大変満足しました.
中音が素晴らしい	このイヤホンの中音は文句なしに素晴らしい.
高音質	このイヤホンは，想像以上の高音質を実現しています. このイヤホンの中音は文句なしに素晴らしい. このイヤホンは，音の良さにびっくりしました. 音の良さが想像以上だった.
その他	低音は多少エンジン音で消されますが. このイヤホンは，低音域の音質にやや欠けている印象を受けました.

表 3　商品 A と商品 B の比較表.

観点	商品 A	商品 B
サイズと重量-サイズが小さくて軽い	サイズが小さくて軽いので，持ち運びに便利で，様々な機能が詰め込まれています．(4)	コンパクトなサイズで，持ち運びが非常に便利です．(8)
音量と調整-音量が大きい	-	音量が大きくて聞きやすいです．(8)
接続のしやすさ-ペアリングが簡単	ペアリングが簡単で，すぐに接続できる．(4)	ペアリングが即座に完了し，非常に簡単でした．(5)

図 3　比較表のプロトタイプシステム.

正解とする．例えば「コスパが非常に高いです.」という要約文に対して「値段が安かったので期待していなかったのだが，実際に使ってみると驚くほど高品質でした.」という評価文は正解であるが，「このイヤホンは値段が高すぎて，どぶにお金を捨

てるみたいなものです.」という評価文は不正解とする．評価文全体の正解数の割合により，2 つのモデルの比較を行う．

4.1.5 観点クラスタリングの評価結果

観点クラスタリングの大グループについて Purity は，Swallow が 0.830，gemma が 0.547 と Swallow が gemma を大きく上回った．また，観点クラスタリングのサブグループについては Swallow が 0.705，gemma が 0.824 と gemma が Swallow を上回ったが大グループほど大きな差はなかった．大グループについて gemma が Swallow を大きく下回った要因として，観点を全て入力して出力されたクラスタリング結果が正確ではなかったことが考えられる．例えば，「接続性に関する評価」というグループに「対応の速さ」や「迅速な対応」といった対応に関する観点が含まれていた．これにより，その後全ての観点を分類しなおした際に「対応」に関する観点が「接続性に関する評価」に含まれた．

4.1.6 評価集約の評価結果

評価集約については，Swallow が 0.827，gemma が 0.881 とどちらも高い正解率であった．

4.2 ユーザ実験

提案手法が，比較に有用であるかを検証するためにユーザ実験を行った．

4.2.1 デ ー タ

ユーザ実験に用いたデータについて説明する．商品レビューについては比較実験と同様に楽天データセットに含まれる楽天市場の商品レビューデータ [16] を利用した．比較表の対象商品は，商品ジャンル ID が同一であり，販売価格の近い 2 商品とする．また，1 商品につき 100 件のレビュー内容を使用した．実験には商品ジャンルがイヤホンである商品から 4 つの商品を選択した．この 4 つの商品をそれぞれ商品 C，商品 D，商品 E，商品 F と呼ぶ．この 4 つの商品のうち商品 C と D の組，商品 E と F の組の 2 組で実験を行った．

4.2.2 大規模言語モデル

実験には，4.1 節の大規模言語モデルの比較実験で安定して良い結果の出た Llama-3.1-Swallow-8B-Instruct-v0.1 [2]を大規模言語モデルとして使用する．実験では，8bit 量子化したモデルを使用した．また，全てのモデルで temperature は 0 と設定した．

4.2.3 実験参加者

実験参加者は，兵庫県立大学に所属する学生 4 名（男性 3 名，女性 1 名）である．2025 年 1 月 30 日から 2 月 4 日にかけて実験を行った．

4.2.4 比 較 手 法

比較手法としては，提案手法で用いたレビュー 100 件と同一のレビューを 2 商品並べて表示させるシステムを用いた．システムの表示例を図 4 に示す．システムでは各商品のレビューを並べて表示させ，各商品でスクロールできるようにした．

2：https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.1

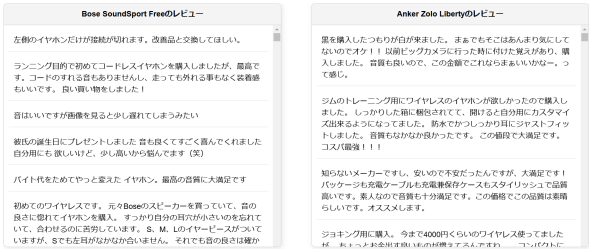

図 4 比較手法の表示例.

表 4 各被験者が取り組んだタスク.

被験者	タスク 1	タスク 2
1	比較手法/商品 CD	提案手法/商品 EF
2	提案手法/商品 CD	比較手法/商品 EF
3	比較手法/商品 EF	提案手法/商品 CD
4	提案手法/商品 EF	比較手法/商品 CD

4.2.5 実 験 手 順

実験では，まず 2 つのシステムの使い方を確認する訓練タスクを行った．次にシステムを用いて 2 つの商品を比較するタスクを行い，タスク後アンケートに回答してもらった．1 つのシステムにつき 1 タスクを行い，1 人の被験者につき 2 つのタスクを行ってもらう被験者内実験を行った．実験参加者が行ったタスクと使用したシステムの順番を表 4 に示す．全タスク終了後に最終アンケートを行い，実験を終了した．

4.2.6 実験タスク

実験参加者には，2 つのイヤホンの比較を 2 つのシステムを用いて行ってもらい，どちらの商品を購入するかを決定するタスクを行ってもらった．実際にオンラインショッピングで買い物をする状況に近づけるため，イヤホンを使用するシチュエーションを指定し，商品画像や価格，スペックを提示した．シチュエーションは「通勤・通学時に使用する」「自宅で勉強やオンライン会議をする際に使用する」の 2 つとした．「通勤・通学時に使用する」というシチュエーションについてのタスクを解く場合，以下の文章で状況を説明しタスクに取り組んでもらった．

> あなたは，現在通勤・通学時に使用するイヤホンを新しく購入しようと考えています．
> 購入する候補のイヤホンを 2 つのイヤホンに絞りました．システムを用いてレビューの情報から 2 つの商品を比較し，どちらを購入するか決めてください．
> 制限時間は 15 分です．15 分以内に決定できた場合はそこで比較を行うのをやめていただいて構いません．また，各商品の価格，スペックは以下の通りです．比較の際に参考にしてください．

タスクの制限時間は 15 分とし，購入する商品が決定できた場合は切り上げても良いこととした．

4.2.7 アンケート

各タスクの終了後にタスク後アンケートを行った．タスク後

アンケートでは，2 つの商品のうちどちらの商品を選択したか，なぜその商品を選択したか，選択にかかった時間を回答してもらった．さらに，4 つの設問について 5 段階リッカート尺度 (1：そう思わない，2：あまりそう思わない，3：どちらともいえない，4：ややそう思う，5：そう思う) で回答してもらった．質問は「時間内に十分情報を得られた」，「システムは様々な観点について比較するのに役立った」，「システムは比較するための観点を知るのに役立った」，「システムは使いやすいと思う」の 4 つである．また，全タスクの終了後，最終アンケートを行った．最終アンケートでは，どちらのシステムが比較に適しているかとその理由を回答してもらった．また，提案手法についてよかった点と改善点を回答してもらった．

4.2.8 結　　果

まず，比較にかかった時間は比較手法が平均 11.25 分，提案手法が 11.75 分とほぼ差は見られなかった．また，有意水準 5% でウェルチの t 検定を行ったところ有意差は認められなかった．次に各タスク後に行ったアンケートについて手法と項目ごとに被験者の平均と標準偏差を算出した結果を表 5 に示す．この結果からすべての項目で提案手法が比較手法を上回っていることがわかる．また，有意水準 5%でウェルチの t 検定を行ったところ，「時間内に十分情報を得られた」を除く 3 項目で有意差が認められた．

最後に，最終アンケートの結果について説明する．「どちらのシステムが比較に適していたか」という質問に対し，4 名全員が提案手法と回答した．その理由として「観点や評価で分けられていたので自分が見たい項目だけを見ることができ，さらに短くまとめてくれていたので読みやすかったから」や「観点を明示的に示してくれているかつ件数まで示してくれて数量的に比較できるため便利だと思ったから」などが挙げられた．

一方で改善点として「観点が多すぎるように感じた.」や「観点が多すぎるので，全部見ようとは思いませんでした.」など表示される観点数の多さや「観点が重複しているものがいくつかあった点.」や「観点が重複していそうなものがあり，少し探しにくく感じた.」など重複している観点の存在が指摘された．実際に作成した比較表を確認したところ，「ノイズキャンセリング」という観点は「音源」や「静粛性」などの複数のグループに存在していた．また，「表示されている内容とホバーして表示している内容が間違っているところがあった.」のように提案手法の正確性に対する課題が明らかになった．表示内容の間違いについてはレビューから観点と評価を抽出する際にレビューに書かれていない内容が抽出されたり，評価文を作成する際に抽出された評価内容だけでなく間違った情報が付け加えられたりしたことが原因だと考えられる．

5　まとめと今後の課題

本研究では，大規模言語モデルを用いてレビューを自動的に集約し，観点ごとに集約結果を対応付けた商品の比較表の作成を行った．提案手法の有用性を検証するためにユーザ実験を行い，「時間内に十分情報を得られるか」，「様々な観点について比較するのに役立つか」，「商品を比較するための観点を知るのに役立つか」，「使いやすいか」の 4 つの項目でアンケートを行った．回答者の平均は 4 つのすべての項目で比較手法を上回り，「時間内に十分情報を得られるか」を除く 3 項目では有意差が認められた．また，全タスク終了後に行ったアンケートでは，「どちらのシステムが比較に適していたか」という質問に対し，全員が提案手法と回答した．このことから，提案手法は比較に有用であるといえる．提案手法に関するアンケートより，観点や評価ごとに要約されている点，どのような意見が何件あるかを示されている点が良かった点として挙げられた．一方で，表示される観点数が多いことや，重複する観点がいくつか見られた点，表示される評価文とレビューの内容が一致していない点が改善点として指摘された．適切な観点の数の検討や集約精度の向上が必要である．

今後の課題として，評価文の生成精度の向上が挙げられる．現在，抽出した観点と評価を入力し，1 文にまとめた評価文を生成しているが，入力した情報からは得られない情報が付け加えられるという問題がある．例えば，「音質」という観点と「音切れが頻繁に起きて音楽を聴くのにストレスを感じる」という評価を入力した際に，「音質は良かったが，音切れが頻繁に起きて音楽を聴くのにストレスを感じる.」という評価文が生成された．「音切れが頻繁に起きて音楽を聴くのにストレスを感じる.」は入力で与えた情報だが，「音質は良かった」は与えられていない情報である．このようなハルシネーションを抑える方法や元のレビューと評価文が一致しているかの評価の検討を行う．

また，観点クラスタリングについて Purity という評価指標で評価を行った．そのため，各グループに対して純度の高いグループになっているかは評価ができているが，アンケートで指摘のあった重複する観点グループの有無などについては評価できていない．同様に，評価集約についても要約文と評価文の内容が一致しているかについては評価できているが，本来まとめられるべき評価文が別のグループに分けられていないかなどは評価できていない．今後，別の評価指標を用いた評価の必要がある．

謝　　辞

本研究は JSPS 科学研究費助成事業 JP24K03228, JP22H03905, JP21H03554, JP21H03775 による助成を受けたものです．ここに記して謝意を表します．また，本研究では国立情報学研究所の 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．ここに記して謝意を表します．

文　　献

[1] Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. Aspect-Controllable Opinion Summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6578–6593, 2021.

[2] Yash Kumar Atri, Arun Iyer, Tanmoy Chakraborty, and Vikram Goyal. Promoting Topic Coherence and Inter-

表 5 アンケート結果.

評価項目	比較手法 平均	比較手法 標準偏差	提案手法 平均	提案手法 標準偏差
時間内に十分情報を得られた	3.25	0.96	4.25	0.50
様々な観点について比較するのに役立った	2.25	1.26	4.75	0.50
商品を比較するための観点を知るのに役立った	2.00	0.82	5.00	0.00
システムは使いやすい	2.00	0.82	4.50	0.58

Document Consorts in Multi-Document Summarization via Simplicial Complex and Sheaf Graph. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2154–2166, 2023.

[3] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, pp. 519–528, 2003.

[4] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual Pre-Training for Cross-Lingual LLM Adaptation:Enhancing Japanese Language Capabilities. In *Proceedings of the First Conference on Language Modeling*, 2024.

[5] Beliz Gunel, Sandeep Tata, and Marc Najork. STRUM: Extractive Aspect-Based Contrastive Summarization. In *Companion Proceedings of the ACM Web Conference 2023*, pp. 28–31, 2023.

[6] Beliz Gunel, James B. Wendt, Jing Xie, Yichao Zhou, Nguyen Vo, Zachary Fisher, and Sandeep Tata. STRUM-LLM: Attributed and Structured Contrastive Summarization. *arXiv preprint arXiv:2403.19710*, 2024.

[7] Hayate Iso, Xiaolan Wang, Stefanos Angelidis, and Yoshihiko Suhara. Comparative Opinion Summarization via Collaborative Decoding. In *Findings of the Association for Computational Linguistics*, pp. 3307–3324, 2022.

[8] Kevin Lerman and Ryan McDonald. Contrastive Summarization: An Experiment with Consumer Reviews. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pp. 113–116, 2009.

[9] Parth Mehta and Prasenjit Majumder. Content Based Weighted Consensus Summarization. In *Proceedings of the 40th European Conference on IR Research*, pp. 787–793, 2018.

[10] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a Large Japanese Web Corpus for Large Language models. In *Proceedings of the First Conference on Language Modeling*, 2024.

[11] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint arXiv:2408.00118*, 2024.

[12] Abinash Tripathy, Ankit Agrawal, and Santanu Kumar Rath. Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, Vol. 57, pp. 117–126, 2016.

[13] Jing Nathan Yan, Tianqi Liu, Justin Chiu, Jiaming Shen, Zhen Qin, Yue Yu, Charumathi Lakshmanan, Yair Kurzion, Alexander Rush, Jialu Liu, and Michael Bendersky. Predicting Text Preference Via Structured Comparative Reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers*, pp. 10040–10060, 2024.

[14] Xin Zhang, Qiyi Wei, Qing Song, and Pengzhou Zhang. TOMDS (Topic-Oriented Multi-Document Summarization): Enabling Personalized Customization of Multi-Document Summaries. *Applied Sciences*, Vol. 14, No. 5, 2024.

[15] Yusen Zhang, Nan Zhang, Yixin Liu, Alexander Fabbri, Junru Liu, Ryo Kamoi, Xiaoxin Lu, Caiming Xiong, Jieyu Zhao, Dragomir Radev, Kathleen McKeown, and Rui Zhang. Fair Abstractive Summarization of Diverse Perspectives. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3404–3426, 2024.

[16] 楽天グループ株式会社. 楽天市場データ. 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), 2020. https://doi.org/10.32130/idr.2.1.