大規模言語モデルによるレビューの対照型集約にもとづく商品比較表の生成

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 (LLM) を用いてレビューを集約することで，2 つの商品に共通する評価や商品固有の評価を対照的に整理した比較表を作成する手法を提案する．比較表とは，2 つの商品の各観点に対して共通する評価が何件あるかや各商品固有の評価が何件あるかを一覧することのできる表である．LLM を用いることで，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，LLM を用いてレビューから観点と評価を抽出し，観点ごとに要約文を生成，要約文を共通と固有に分類し，1 つの比較表にまとめる．共通・固有要約文の精度評価では，比較手法である AACSum の精度を上回った．一方で，共通・固有を考慮しない比較表を比較手法として行ったユーザ実験では，2 つの比較表に大きな差は見られず，本研究で実現した LLM を用いたレビュー集約による比較表の精度においては共通・固有を考慮することはユーザの商品決定に大きな影響はないことが示唆された．

キーワード LLM，対照要約，インタラクティブシステム

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，観点ごとに評価文の集約を行う．また，複数商品間で同じ観点に対する評価文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，各評価文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「フィット感」についてのレビューをまとめた要約文が表示されている．この表では，フィット感に関して，共通する要約が「耳にしっかりフィットする」であることを示している．要約文の隣に記載されている数字がそのような評価をしている人が何人いるかを示しており，「耳にしっかりフィットする」という評価をしている人が商品 A には 10 人，商品 B には 12 人いることがわかる．また，各要約文の下は要約文に含まれるレビューからその観点に言及している部分を抽出した評価文を示している．また，その下の段にはそれぞれの商品固有の要約がまとめられている．商品 A 固有の要約では「耳が痛くなる」と評価しているレビューが 8 件，商品 B 固有の要約では「耳が痛くならない」と評価しているレビューが 5 件あることを示している．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．

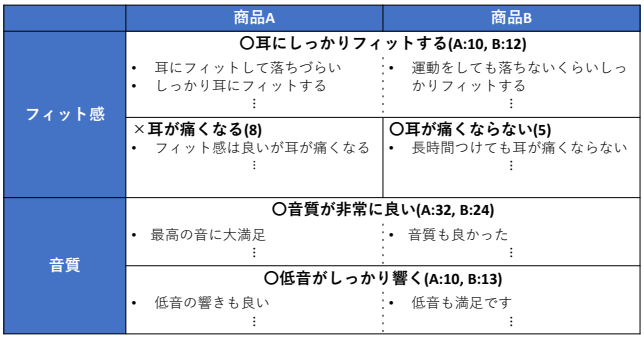


図 1 レビュー集約を行った比較表のイメージ図．

そのような比較表を作成するため，我々はこれまでに大規模言語モデル (LLM) を用いてレビューを集約し，観点ごとに対応する評価文を要約し対応付けるという手法を提案した [12]．LLM には，同じ商品カテゴリでも商品によって抽出する観点を変更できたり，他のドメインに学習せず応用することができたりするというメリットがある．しかし，観点ごとに評価文を要約し対応付けた場合，ある観点について必ずしも同じ意見がまとめられているわけではないため，2 つの商品に共通している意見なのか，固有の意見であるのかがわかりづらいという問題があった．

そこで，本研究では各商品に共通する意見，各商品のみに現れる固有の意見に注目し，共通・固有レビューのわかる対照要約による比較表の作成に取り組む．共通・固有を考慮した要約を行う手法として AACSum [8] という手法が提案されている．

しかし，この手法には「共通・固有の精度があまり良くない」，「要約文とレビューが紐づいていない」という 2 つの問題点がある．AACSum では一度に全ての評価文を入力して要約を生成しているが，本研究ではそれぞれの商品で要約を行って重要な情報を抽出し，ナゲットという細かい単位で比較を行い共通と固有に分類することで精度が上がるのではないかと考えた．本研究では，商品比較タスクにおいて共通・固有を考慮した場合と考慮しない場合の比較表に違いがあるかを検証するためにユーザ実験を行った．その結果，2 つの手法の間に回答の差は見られなかった．一方で，どちらの手法が比較に適していたかという最終アンケートの結果では，対照型比較表が共通・固有を考慮しない比較表に比べて適していたという意見が多くみられた．最終アンケートは提案手法と比較手法の二者択一であったこと，各タスク後の主観評価では差がなかったことから，共通・固有を考慮することはユーザの商品決定に大きな影響はないことが示唆された．

2 関 連 研 究

対照要約 (Contrastive Summarization) 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．

Kim らは肯定的意見集合と否定的意見集合から比較可能な文を探し，対照的な文ペアから比較要約を作成する最適化フレームワークを提案している [9]．Lerman らは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 [11]．Paul らは教師なしの確率手法とランダムウォーク定式化を組み合わせて対照要約を行う手法を提案している [14]．Huang らは要約内の比較性と両トピックに対する代表性を最大化する文を選択することで共通点と相違点を明らかにする手法を提案している [7]．Sipos らは重要な観点をカバーしながら冗長性を抑える短文レビューペアを選択するスコアリング関数を提案している [15]．Özsoy らは特定の観点に対する代表的かつ対照的な意見ペアを生成する CMSOS という手法を提案している [13]．Lacanya らは LSTM によって文脈を考慮した対照要約を行う CSLSTM という手法を提案している [10]．Hoang らは観点ベースと抽象型を組み合わせレビュー集合から共通要約と対照要約を生成する手法を提案している [6]．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 [4]．さらに LLM 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 [5]．しかし，STRUM-LLM はそれぞれの属性で集約しており，複数商品の共通・固有な意見がわかる集約にはなっていない．共通・固有を考慮した集約として，Jin らは 2 つの対象について対照的および共通の要約に加え，固有の特徴を強調する要約を行う AACSum という手法を提案している [8]．

現在対照要約を行う手法は複数提案されているが，ユーザ実験による検証は不十分である．また，本研究と同様に共通と固有の要約文を生成する AACSum は 1 節で述べたように，共通・固有の要約精度，要約文の根拠の提示の不足という 2 つの課題がある．我々は，要約を短文に分割し，短文とレビューを紐づけることで要約文の根拠となるレビューを確認できる比較表を作成する．また，要約を短文に分割し比較することで細かい単位での共通・固有の抽出に取り組む．さらに，ユーザ実験によって実際の商品比較タスクにおける有用性の検証も行う．

3 共通・固有を考慮したナゲット単位に基づく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，LLM を用いてレビューを集約し，商品の比較表を作成する．実際に本研究で提案する手法で作成した比較表の例を図 2 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，観点ごとに評価文の内容を集約する．また，各観点ごとに要約した内容から共通点を探し，共通要約文と固有要約文に分類し，観点ごとに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，**評価文**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」,「Bluetooth」,「ノイズキャンセリング」などがある．評価文は「音質」という観点に対しての「音質が良い」という文や「Bluetooth」という観点に対しての「Bluetooth 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を抽出する．次に，各観点に対応する評価文をレビューから抽出する．その後，各観点ごとに共通・固有要約文を生成する．2 つの商品に共通する内容をまとめた要約文を**共通要約文**，一方の商品のみに現れる内容をまとめた要約文を**固有要約文**と呼ぶ．比較表では，各観点ごとに共通要約文，固有要約文をまとめて表示する．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．

3.2 提案手法の概要

提案手法の概要を図 3 に示し，以下で概要を説明する．

1. 観点抽出
   - 2 つの商品のレビューを同時に LLM に入力し，共通する観点・固有の観点を抽出する．
2. 評価文抽出
   - レビューを 1 件ずつ LLM に入力し，各観点に対応する評価文を抽出する．
3. ナゲットに基づく共通要約文・固有要約文の生成
   - 各観点グループごとに評価文集合を LLM に入力し，1 文に要約する．
   - その後，生成された要約をナゲットという単位の短文に分割する．

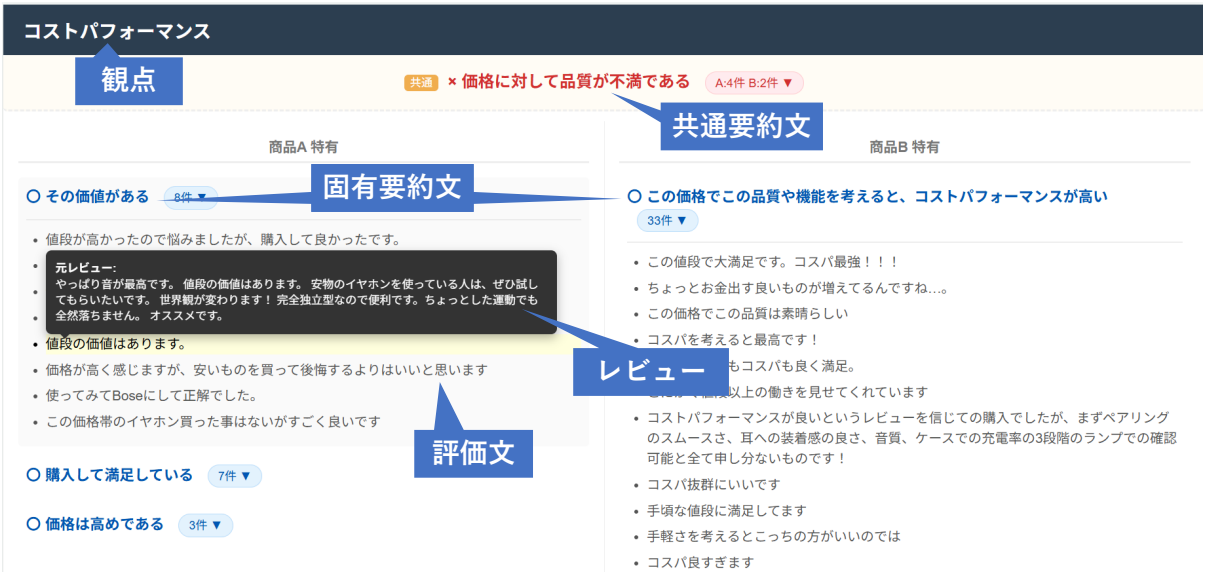


図 2　作成する比較表の例.

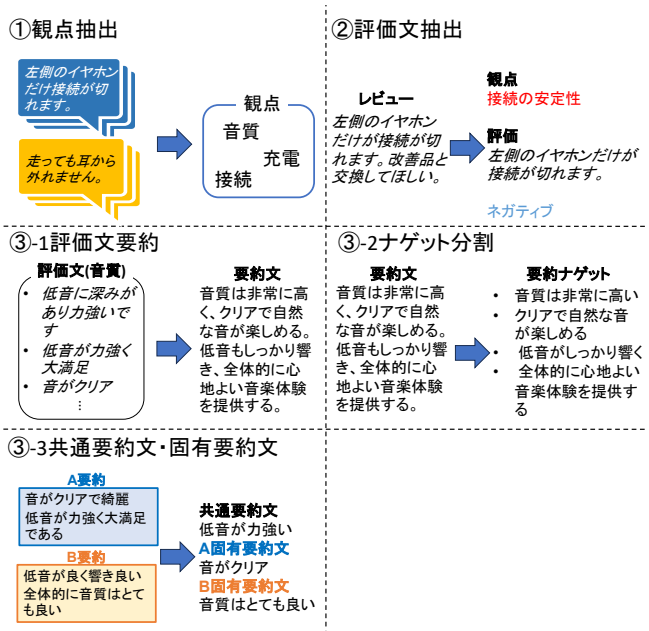


図 3　提案手法の概要.

- 作成された要約文を比較し，2 つの商品に共通する要約文と各商品固有の要約文に分類する.

以降，それぞれのステップの詳細について説明する．なお，説明にあたりイヤホンカテゴリの 2 商品に対するレビューを gpt-4o-mini に入力して得られた結果を例に説明する.

3.3　観 点 抽 出

まず，各商品のレビューを LLM に入力し，レビューから観点を抽出する．今回は各商品 100 件のレビューを使用する．以下は，「左側のイヤホンだけが接続が切れます．改善品と交換してほしい.」や「 ランニング目的で初めてコードレスイヤホンを購入しましたが，最高です．コードのすれる音もありませんし，走っても外れる事もなく装着感もいいです」といったレビューを入力する際のプロンプトである．以下のプロンプトは入力するレビューのうち 2 件のみ表示しており，他のレビューは省略している.

観点抽出を行うプロンプト

あなたは，レビュー分析者です．特に指示が無い場合は，常に日本語で回答してください．以下は商品 A と B についていたレビュー各 100 件です.
商品 A のレビュー：**左側のイヤホンだけが接続が切れます．改善品と交換してほしい．/ ...**
商品 B のレビュー: **ランニング目的で初めてコードレスイヤホンを購入しましたが，最高です．コードのすれる音もありませんし，走っても外れる事もなく装着感もいいです． / ...**
これらのレビューがあるとき，この 2 つのイヤホンを比較する際どのような観点から比較を行いますか？ A によく出てくる観点，B によく出てくる観点，A と B に共通して出てくる観点をそれぞれ 3 つずつ挙げてください.

イヤホン 2 商品のレビューを入力したところ，商品 A でよく言及されている観点「音質の良さ，フィット感，接続の安定性」，商品 B でよく言及されている観点「コストパフォーマンス，軽さと装着感，デザインの良さ」，共通する観点「音質，しやすさ，使用時の快適さ」が出力された．商品 A でよく言及されている観点，商品 B でよく言及されている観点，共通する観点の中で重複する観点が存在する場合，重複する観点を除いたリストを作成する．今回の例では重複がなかったため「音質，接続の安定性，デザインの良さ，音質の良さ，軽さと装着感，接続のしやすさ，フィット感，コストパフォーマンス，使用時の快適さ」の 9 つの観点となった.

3.4 評価文抽出

次に，比較表の観点ごとの要約を行うために，観点に対応する評価文を抽出する．また，3.5.1 節で要約を生成する際にポジティブ，ネガティブを分けて要約するためにポジティブネガティブの判定を行う．レビューを 1 件ずつ LLM に入力し，3.3 節で得られた観点に対応する評価文を抽出する．以下は，「左側のイヤホンだけが接続が切れます．改善品と交換してほしい.」というレビューを入力する場合のプロンプトである．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．

評価文抽出を行うプロンプト

ユーザーから与えられたレビューから，観点リスト内の観点に具体的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．また，評価が観点に関して肯定的か否定的かを判定し，“ポジティブ” たは “ネガティブ” のどちらかを出力してください．
観点は観点リスト内の表記と同じ表記としてください．
観点リスト: **音質，接続の安定性，デザインの良さ，音質の良さ，軽さと装着感，接続のしやすさ，フィット感，コストパフォーマンス，使用時の快適さ**
レビュー：**左側のイヤホンだけが接続が切れます．改善品と交換してほしい．**

「左側のイヤホンだけが接続が切れます．改善品と交換してほしい.」というレビューからは「接続の安定性」という観点について「左側のイヤホンだけが接続が切れます」という評価文が抽出され，この評価文はネガティブと判定された．また，ここで抽出した評価文の件数を用いて重要な観点の絞り込みを行う．抽出した結果，言及数が多い観点上位 5 件に絞った．その結果，イヤホンの例では「音質，接続の安定性，フィット感，コストパフォーマンス，接続のしやすさ」の 5 つになった．

3.5 ナゲットに基づく共通要約文・固有要約文の生成

共通要約文・固有要約文の生成にあたり，既存研究の AACSum では評価文を全て一度に入力し共通・固有要約文を生成しているが，精度がよくないという問題がある．その原因の 1 つは入力の長さであると考え，本研究ではまず各商品の要約を行う．要約を行うことにより，各商品の重要な情報を抽出しまとめ，重複する意見や不要な意見を除き共通意見と固有意見を探しやすくする．次に，各商品の要約文をナゲットに分割する．ナゲットとは，それ以上分割することのできない情報の意味的な最小単位である [17]．ここでナゲットに分割する理由は 2 つある．1 つはナゲットに分割することで細かい単位で共通しているか，片方の商品固有であるかを比較することができるからである．もう 1 つは，細かく分割することにより，どの評価文がどの要約に結びついているかをわかりやすくすることができるからである．最後に，各商品の要約ナゲットの中から 2 つの商品に共通する内容を抽出し，共通要約文と固有要約文に分類する．

3.5.1 要 約 生 成

抽出した評価文を商品，観点，感情（ネガティブ，ポジティブ）を分けて LLM に入力し，1 文の要約を生成する．以下は，音質グループのポジティブな評価文「音はいいですが」や「音も良くてすごく喜んでくれました」などを入力するプロンプトである．

要約を生成するプロンプト

イヤホンの**ポジティブ**なレビューを複数与えます．これらの複数レビューのうち**音質**に言及している**ポジティブ**な内容を 1 つに要約してください．
「〜という意見がある」「〜と評価されている」「〜との声が多い」などレビューが存在することを示すメタ表現は使わず，事実・特徴・状態を直接述べる形式にしてください．
また，言及数の多い主張を優先的にまとめてください．
ただし，出力は要約文のみとしてください．
入力：**音はいいですが/音も良くてすごく喜んでくれました/最高の音質に大満足です/ ...**

例えば，商品 A の音質に関するポジティブな評価文は「音質は非常に高く，クリアで自然な音が楽しめる．低音もしっかり響き，全体的に心地よい音楽体験を提供する.」というように要約された．

3.5.2 要約のナゲット分割

要約後，生成された要約をナゲットと呼ばれる短文に分割する．ナゲットに分割後，評価文と要約文集合を LLM に入力し，各要約文と評価文の紐づけを行う．以下は音質グループのポジティブな要約「音質は非常に高く，低音がしっかり響き，クリアで自然な音が楽しめる．臨場感もあり，まるで高級オーディオで聴いているような感覚を提供する.」を入力する際のプロンプト例である．

ナゲットに分割するプロンプト

次の文章から「意味のある情報の最小単位（ナゲット）」を抽出してください．ナゲットとは，対話システムがユーザに提供した意味のある情報の原子的，最小単位を指す．また，主に，一文，またはそれ以下の情報の断片（句や節）である．単独で再利用可能（=単独で意味が通じる）事実，主張，知識である．因果関係や主述関係がわかるように抽出してください．必ず 1 行に 1 ナゲットずつ出力してください．句点は必要ありません．
抽出のルール：
1. 明確な意味を持つ情報断片（事実・定義・因果・例示など）を抽出対象とする．
2. ユーザの情報ニーズに直接または間接的に関連する内容を含める（補足情報も OK）．
3.1 文に複数の独立した情報が含まれる場合は，ナゲッ

トとして分割する．
4. 意味が曖昧な主観語や儀礼的表現（例：「いいですね」「微妙でした」）は除外する．
5. 例示・背景・文脈情報も，意味を持つ限り抽出対象とする．
6. すべてのナゲットは，可能な限り「単独で読んで理解できる形」で記述する．
入力：**音質は非常に高く，クリアで自然な音が楽しめる．低音もしっかり響き，全体的に心地よい音楽体験を提供する．**

例えば，商品 A のポジティブな要約「音質は非常に高く，クリアで自然な音が楽しめる．低音もしっかり響き，全体的に心地よい音楽体験を提供する．」は「音質は非常に高い，クリアで自然な音が楽しめる， 低音がしっかり響く， 全体的に心地よい音楽体験を提供する」の 4 つのナゲットに分割された．

また，ナゲットに分割された各要約文に対応する評価文を紐づける．評価文を 1 つずつ入力し，その評価文の対応する観点のナゲットを分類先の候補としてどの要約文に当てはまるかを分類する．以下は「音質」という観点に対する評価文「音質も低音がきいてて BOSE らしく最高」を「音質は非常に高い，クリアで自然な音が楽しめる， 低音がしっかり響く， 全体的に心地よい音楽体験を提供する」のいずれかの要約文に分類するプロンプトである．

要約文と評価文を紐づけるプロンプト

次のレビューの内容は以下の要約文のいずれかの内容と一致していますか？
レビュー：**音質も低音がきいてて BOSE らしく最高**
要約文リスト: **音質は非常に良い，低音がしっかり響く，クリアで自然な音が楽しめる，全体的に心地よい音楽体験を提供する**
ある要約文に含まれる場合，その要約文に対してレビューを出力してください．要約文に含まれない場合は「なし」と出力してください．

例えば，商品 A のポジティブな音質の評価文「音質も低音がきいてて BOSE らしく最高」は「低音がしっかり響く」という要約文に紐づけられた．

3.5.3 ナゲット単位の要約比較に基づく共通要約文・固有要約文の抽出

2 つの商品の要約文を LLM に入力し，2 つの商品に共通する要約文を探索する．この際，共通要約文として抽出されなかった要約文を固有要約文として扱う．以下は，「フィット感」グループのポジティブな要約文を入力し，共通要約文を探すためのプロンプトである．

共通要約文を探すプロンプト

次の商品 A と商品 B の要約文リストに共通している内容がある場合，１つのグループにまとめてください．
手順:
1. 商品 A と商品 B で「内容が一致・または非常に類似している」要約文を見つけ，ペアまたはグループにしてください．
2. グループ化したものについて，その内容を端的に表す「共通要約」を作成してください．
3. どちらの商品にも共通しない要約文は，出力には含めなくて良いです．共通のみを出力してください）．
注意: 入力にある要約文を一字一句変えずに nuggets_a と nuggets_b に入れてください．
共通点がない場合は空のリストを出力してください．
【商品 A のリスト】**フィット感が非常に良い，激しく動いても外れない，装着時の安定性に満足している**
【商品 B のリスト】**イヤホンは耳にしっかりフィットする，イヤホンはランニングやジョギング中でも外れない，イヤホンは快適に使用できる**

例えば，上記の「フィット感」グループでは商品 A の「フィット感が非常に良い，激しく動いても外れない」と商品 B の「イヤホンは耳にしっかりフィットする，イヤホンはランニングやジョギング中でも外れない」が共通要約文として抽出され，「フィット感が良く，激しく動いても外れない」という要約にまとめられた．また，共通要約文として抽出されなかった，商品 A の「装着時の安定性に満足している」と商品 B の「イヤホンは快適に使用できる」という要約文は各商品の固有要約文として扱う．

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 4 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．要約文は，要約文に含まれる評価文の件数で降順に並んでいる．次に，各要約文に含まれる評価文の件数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．最後に各評価文にカーソル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 共通・固有要約文の精度評価

共通・固有要約文の抽出について各観点の評価文を参照し，どの程度の精度でそれぞれが抽出できているかを評価した．また，観点とそれに対応する評価文を入力し，1 ステップで観点ごとに共通・固有の要約を生成する AACSum との精度の比較を行った．

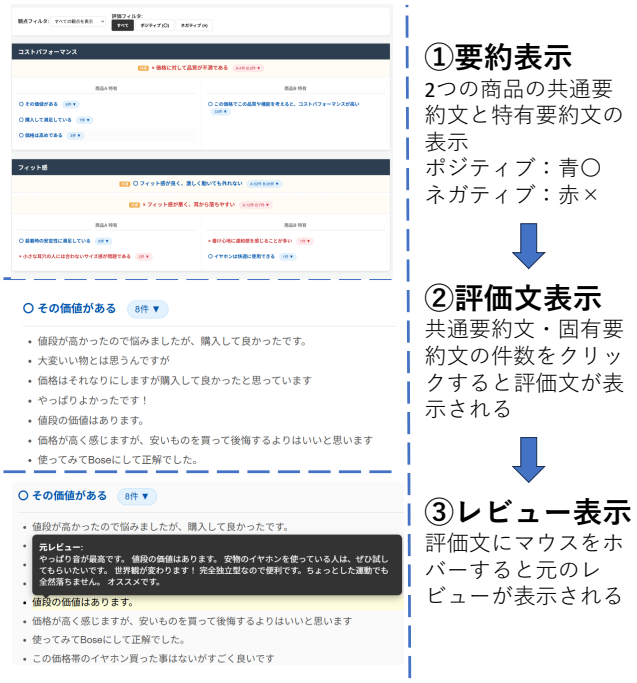


図 4　比較表のプロトタイプシステム．

4.1　デ　ー　タ

実験では楽天データセットに含まれる楽天市場のレビューデータを用いた [16]．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．比較表を作成するために，各商品 100 件のレビューを用いる．比較表の対象商品は，商品ジャンル ID が同一であり，レビューが 100 件以上ついている販売価格の近い 2 商品とした．複数のドメインでの有効性を検証するため，テレビとマウスという 2 つのドメインを実験に用いた．2 つのドメインでそれぞれ 2 つの商品を選択した．選択したテレビとマウスをそれぞれテレビ A，B，マウス A，B と呼ぶ．それぞれ同じ商品ジャンル内の 2 商品の組み合わせ 2 組で実験を行った．

4.2　LLM

実験では LLM として，gpt-4o-mini-2024-07-18 (gpt-4o-mini) [1]を使用した．また，全ての過程で temperature は 0.0 と設定した．

4.3　比 較 手 法

比較手法として，AACSum [8] を用いた．AACSum は，観点とそれに対応する評価文を入力し，各観点ごとに共通要約文と固有要約文を生成する手法である．評価の条件を揃えるため，入力には提案手法で抽出された観点及び評価文を使用し，ポジティブとネガティブを分けて生成を行った．また，生成された要約文は 3.5.1 節のナゲットに分割するプロンプトを使用し，ナゲットに分割したうえで評価を行った．

4.4　評 価 方 法

共通・固有要約文の評価を行うために，著者を含む 3 名によりアノテーションを行った．アノテーション作業では各観点ごとに，評価文集合を参照し提案手法及び比較手法で生成された要約文を共通，A 固有，B 固有の 3 種類に分類した．アノテーターの評価結果がどれくらい一致していたかを示す Fleiss の κ 係数は 0.74 であり，かなりの一致 (Substantial agreement) であった [3]．評価結果は，3 人の多数決の結果を，3 人が全員別の評価を行った場合のみ著者の結果を正解として使用した．評価指標として，正解率，適合率，再現率，F_1 値を用いた．それぞれの手法の評価において，適合率，再現率，F_1 値については共通，A 固有，B 固有のマクロ平均を用いた．

1：https://platform.openai.com/docs/models/gpt-4o-mini

表 1　共通・固有要約文の評価結果．適合率，再現率，F_1 値はマクロ平均を示す．

手法	正解率	適合率	再現率	F_1 値
提案手法	0.78	0.79	0.81	0.78
比較手法	0.72	0.69	0.69	0.68

4.5　評 価 結 果

まず，評価対象となった要約文の数は，テレビタスクでは提案手法が 26 文，比較手法が 61 文であった．同様に，マウスタスクでは，提案手法が 31 文，比較手法が 83 文であった．提案手法，比較手法における共通・固有要約文の精度評価結果を表 1 に示す．

4.6　考　　察

表 1 より，正解率，適合率，再現率，F_1 値全てで提案手法が比較手法を上回っていることがわかる．特に，2 つの手法の大きな差は共通要約文にあり，提案手法の再現率，適合率がそれぞれ 0.54，0.87 であったのに対し，比較手法は 0.35，0.56 と大きく下回った．特に再現率が低いことから，それぞれ固有として生成された要約文が人手の評価では共通と判断されたケースが多いことがわかる．AACSum の生成結果には共通・固有要約文の中に同じ意味を持つ文章が含まれることがあった．例えば，テレビの「コストパフォーマンス」という観点に対する AACSum の要約は共通要約文が「コストパフォーマンスが高い」と「価格に対して満足できる商品である」，商品 A の固有要約文が「コストパフォーマンスは最高である」と「32 型でこのお値段はコスパ最強である」，商品 B の固有要約文が「コストパフォーマンスが高い商品である」，「半年使ってみたところ，値段相応である」，「値段を考えたら十分である」であった．このように共通要約文である「コストパフォーマンスが高い」と，A 固有要約文の「コストパフォーマンスは最高である」，B 固有要約文の「コストパフォーマンスが高い商品である」は内容が一致していた．アノテーター 3 人はこの 3 文を全員が「共通要約文」であると判断した．一方で提案手法では要約文を生成後に共通と固有に分けており，共通と固有に同じ意味を持つ文章が重複するケースをなくすことができたと考えられる．例えば，AACSum 同様にテレビの「コストパフォーマンス」に対する要約結果を見ると，共通要約文が「コストパフォーマンスが高い」と「価格に対する満足度が高い」，A 固有要約文が「価格

の割には作りがしっかりしている」，B 固有要約文が「半年使用した結果，コストパフォーマンスは良くないと感じる」であった．このように「コストパフォーマンスが高い」という内容は固有要約文には含まれず，共通要約文にのみまとめられた．

また，提案手法と比較手法では同じ観点と評価文を使用しているが，生成された要約文をナゲットに分割した結果，提案手法が 57 文，比較手法が 144 文と大きく異なる結果となった．要約文の数が大きく異なる 1 つの要因として，AACSum においては評価文が要約されず，そのまま出力されるケースが複数確認された．例えば，テレビの「画質」という観点について商品 A は「画質もそこそこです」などの 4 件の評価文を，商品 B は「残像が残ります．特にアマゾンプライム動画など，テレビ放送以外は更に観れません．」など 11 件の評価文を入力したが，AACSum はそれらすべてをそのまま固有の要約文として出力した．その結果，入力した評価文中の「残像が残ります．」や「映像に残像のようなものが見えます．」というように共通する意見も 1 つにまとめられなかった．一方で提案手法についても，出力された結果が評価文の必要な情報を全て含んでいるかについては評価を行えていない．今後，比較のために必要な情報が失われていないかという観点についても評価が必要である．

5 ユーザ実験

本研究では，商品比較タスクにおいて共通・固有を考慮した場合と考慮しない場合の比較表に違いがあるかを検証するためにユーザ実験を行った．ユーザ実験では， 4 節で使用したデータと LLM を使用した．

5.1 実験参加者

実験参加者は，兵庫県立大学に所属する学生 12 名（男性 7 名，女性 5 名）である．参加者全員が情報学を専攻する学生であり，そのうち 4 名が学部生，8 名が大学院生であった．実験参加者の大学生には 1,220 円，大学院生には 1,410 円を支払った．2025 年 12 月 17 日から 2026 年 1 月 8 日にかけて実験を行った．

5.2 比較手法

比較手法としては，提案手法の共通要約文を探索する前までに得られた要約文を商品ごとに表示するシステムを用いた．システムの表示例を図 5 に示す．システムでは観点ごとに要約文を横並びに表示させた．また，要約文に含まれる評価文の件数が多い順に並べる点や，件数をクリックすると評価文，レビューを見ることができるという点は提案手法と同様である．

5.3 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 2 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．



図 5 比較手法の表示例.

表 2 各被験者が取り組んだタスク.

被験者	タスク 1	タスク 2
1,5,9	提案手法/テレビ	比較手法/マウス
2,6,10	比較手法/テレビ	提案手法/マウス
3,7,11	比較手法/マウス	提案手法/テレビ
4,8,12	提案手法/マウス	比較手法/テレビ

5.4 実験タスク

実験参加者には，テレビとマウス 2 商品の比較を 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．実験参加者には，1 つ目のタスクではテレビまたはマウスどちらかの商品を提案手法またはベースラインのシステムを用いて比較してもらい，どちらの商品を購入するかを決定してもらった．その後 2 つ目のタスクは，1 つ目で比較しなかった商品を使用しなかったシステムを用いて同様の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，テレビやマウス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「一人暮らしを始めることになって購入するテレビ」「大学の課題や研究に取り組む際に使用するマウス」の 2 つとした．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．

5.5 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」の 3 つである．さらに，SUS を用いてシステムの使いやすさについて評価を行った [2]．SUS は 100 点を最高点とし，10 個の質問によって標準化された評価が可能なユーザビリティを評価する指標である．SUS の評価は 10 個の質問に 5 段階リッカート尺度で回答してもらう．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．

5.6 結　　果

この節では実験結果を報告する．提案手法と比較手法の分析には有意水準 5%のウェルチの t 検定を用いた．まず，タスクにかかった時間は比較手法が平均 9.78 分，提案手法が平均 7.92 分であった．また，検定の結果有意差は認められなかった．

次に，比較表が比較の手助けになっていたか客観的な評価を行うために，各タスク後に行ったアンケートの自由記述の結果を分析した．具体的には，自由記述のうち「2 つの商品の共通点と相違点を思いつく限り挙げてください」という設問について分析を行った．共通点や相違点を 1 つの観点につき 1 つとカウントし，比較手法と比較を行った．2 つの商品の共通点について挙げられた個数の平均は比較手法が 2.59 個，提案手法が 2.92 個であり提案手法が比較手法を上回った．一方で，2 つの商品の相違点については比較手法が 2.25 個，提案手法が 2.17 個であり比較手法が提案手法を上回っている．どちらも有意差は認められなかった．比較表から得られる情報がどれだけあったかを分析するために，スペック情報から見つけることのできる共通点や相違点を分析対象から除き，もう一度分析を行った．その結果，2 つの商品の共通点は比較手法が 1.67 個，提案手法が 2.17 個と提案手法が比較手法を上回った．同様に 2 つの商品の相違点は比較手法が 1.83 個，提案手法が 1.50 個と比較手法が提案手法を上回った．どちらも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 3 に示す．この結果，「時間内に十分情報を得られたか」という質問については，提案手法が比較手法を上回っている．一方で，「システムは，様々な観点について比較するのに役立ったか」という質問では 2 つの手法の結果が等しく，「システムは商品を比較するための観点を知るのに役立ったか」という質問では比較手法が提案手法を上回る結果となった．また，3 つすべての質問で有意差は認められなかった．

また，SUS のスコアは比較手法が 85.21，提案手法が 86.46 であり，提案手法が比較手法を上回ったが，有意差は認められなかった．Bangor らが報告する約 3,500 件のシステムの SUS 平均スコアは 69.5 であり [1]，どちらも平均を上回っている．

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，12 人中 11 人が提案手法と回答した．その理由として，「共通項がまとめられていて，比較する労力が減ったから」や「比較すべき観点・レビューがわかりやすいから」などが挙げられた．一方で改善点として，「観点が抽象的な場合があり，自分が比較したい点について比較表では比較することができなかった」や「マウスでの電池の持ち，バッテリーの持ちを調べたかったが，比較表の観点に電池の持ちはなかったため，述べられているレビューがあるのかないのかがわからなかった」のように比較できる観点が限られていることが指摘された．また，「同じ観点内に 1 つのレビューから複数の評価が抽出されており，件数が実際より多く感じる」といった同一レビューから抽出された評価文が複数表示されることに対する指摘があった．

5.7 考　　察

ユーザ実験の結果，共通・固有を考慮した場合，考慮しなかった場合に大きな差は見られなかった．このことから，今回の実験からは共通・固有をまとめることは商品の比較のしやすさに大きな効果はないと言える．一方で，最終アンケートでは，12 人中 11 人が提案手法が比較手法よりも比較に適していたという回答した．その理由として，「共通項がまとめられていて，比較する労力が減ったから」や「比較すべき観点・レビューがわかりやすいから」などが挙げられた．また，有意差は認められなかったが，タスクにかかった時間は共通・固有を考慮した場合のほうが短く，「時間内に十分情報を得られたか」という質問も共通・固有を考慮した場合のほうが考慮しない場合よりもスコアが高くなっている．この結果から，共通部分を抽出することによりユーザの負担の削減ができたと考えられる．また，各タスク後のアンケートで質問した各商品の共通点について有意差は認められないが提案手法が比較手法を上回っており，共通する内容をまとめることで実際に共通点が見つけやすくなったと考えられる．本研究では要約文を共通・固有に分類したが，今後共通だけでなくある観点に対する良いと悪いといった対比を考慮することで相違点についても探しやすくなる可能性がある．

最終アンケートの質問である比較表の改善点として「自分の比較したい観点が比較表上になかった」という観点の不足が複数人に指摘された．本研究では，LLM によって代表的な観点を抽出した後，観点に紐づく評価文の数が多い 5 件を選択した．しかしながら，LLM が抽出した代表的な観点が比較において妥当であったか，観点数が 5 件で十分であったかという点は評価を行っていない．今後，観点の抽出の仕方や，比較における適切な観点数の選択などが必要であると考えられる．また，「ネガティブな評価が複数あっても，その元のレビューが同じことがあり，実際よりネガティブの比率が大きく見える」や「同じ観点内に 1 つのレビューから複数回抽出されており，件数が実際より多く感じる」のように同一レビューから同じ観点に対する評価文が複数抽出され，表示される評価文の件数がレビュー件数より多くなっている点が指摘された．今後，1 つの観点に対して 1 レビューから 1 評価文となるような抽出の仕方の検討をすることや同じ要約文に紐づく同一レビューから抽出された

表 3 ユーザ実験のアンケート結果.

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	4.58	0.51	4.83	0.39
様々な観点について比較するのに役立った	4.42	0.51	4.42	0.79
商品を比較するための観点を知るのに役立った	4.67	0.49	4.42	0.79

評価文は 1 文にまとめ，要約文の隣に表示する件数がレビュー数に一致するように変更することが解決策として考えられる．

6 まとめと今後の課題

本研究では，LLM を用いてレビューを自動的に集約し，観点ごとにナゲット単位での共通・固有要約文の抽出を行い，共通・固有それぞれをまとめた比較表の作成に取り組んだ．共通・固有要約文の精度評価では，比較手法を上回る結果になった．また，共通・固有に着目した比較表の商品比較タスクにおける有用性の検証のために，共通・固有を考慮しない比較表をベースラインとしてユーザ実験を行った．実験の結果,「どちらのシステムがタスクに適していたか」という最終アンケートの質問については多くが提案手法と回答したが，各タスク後のアンケートでは大きな差が見られなかった．最終アンケートは二者択一であったこと，主観評価では差がなかったことから，共通・固有を考慮することはユーザの商品決定に大きな影響はないことが示唆された．また，LLM の観点抽出における不足点や 1 つのレビューから評価文が重複して抽出されている点が指摘された．これらの点については，今後検討する必要がある．

謝　　辞

本研究は，JSPS 科研費 JP24K03228，JP25K03229，JP25K03228 の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] Aaron Bangor, Philip T. Kortum, and James T. Miller. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, Vol. 4, No. 3, pp. 114–123, 2009.

[2] John Brooke. SUS: A Quick and Dirty Usability Scale. *Usability Evaluation In Industry*, Vol. 189, No. 194, pp. 4–7, 1996.

[3] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, Vol. 76, No. 5, p. 378, 1971.

[4] 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.

[5] 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.

[6] Hung-Manh Hoang, Duc-Loc Vu, Huong Nguyen-Thi-Thuy, Duy-Cat Can, and Hoang-Quynh Le. Contrastive summarization of user reviews: An aspect-based abstractive approach. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pp. 488–496, 2024.

[7] Xiaojiang Huang, Xiaojun Wan, and Jianguo Xiao. Comparative news summarization using linear programming. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 648–653, 2011.

[8] Hyeon Jin, Chaewon Yoon, Yurim Oh, and Hyun-Je Song. Abstractive aspect-based comparative summarization. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 1043–1047, 2025.

[9] Hyun Duk Kim and ChengXiang Zhai. Generating comparative summaries of contradictory opinions in text. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 385–394, 2009.

[10] S. K. Lavanya and B. Parvathavarthini. Context Aware Contrastive Opinion Summarization. In *IFIP Advances in Information and Communication Technology*, pp. 16–29, 2020.

[11] 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.

[12] Kanako Nakai, Takehiro Yamamoto, and Hiroaki Ohshima. Generating comparative table by LLM-based product review summarization. In *Proceedings of the 27th International Conference on Information Integration and Web Intelligencet*, pp. 31–46, 2025.

[13] Makbule Gülçin Özsoy and Ruket Çakıcı. Contrastive max-sum opinion summarization. In *Information Retrieval Technology*, pp. 256–267, 2014.

[14] Michael Paul, ChengXiang Zhai, and Roxana Girju. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 66–76, 2010.

[15] Ruben Sipos and Thorsten Joachims. Generating comparative summaries from reviews. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pp. 1853–1856, 2013.

[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.

[17] 酒井哲也. 情報アクセス評価方法論–検索エンジンの進歩のために–. コロナ社, 2015.