LLMを用いた多段階クエリ変換によるより具体的な商品レビューの検索

福井　智矢†　　山本　岳洋†　　湯本　高行†

† 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: †ad24a050@guh.u-hyogo.ac.jp, ††{t.yamamoto,yumoto}@sis.u-hyogo.ac.jp

あらまし　本研究では，ある商品に対する具体的なレビューを抽象的なレビューをもとに商品レビュー集合から検索する手法を提案する．具体的なレビューは商品に対する理解を深めるのに有益である．しかし，単純な疎検索や密検索では具体性を考慮してレビューを検索することができず，抽象的なレビューから具体的なレビューを検索することが難しいと考えられる．そこで本研究では大規模言語モデル（LLM）を用いて多段階に生成された仮想の具体的なレビューをクエリとしてレビューの検索を行う．これにより，具体的なレビューが持つ特徴を含むレビューを検索できると考えられる．実験では，楽天レビューデータを活用し，提案手法が実際に存在する抽象的なレビューをもとにより具体的なレビューを検索することができるか評価を行った．実験の結果，密検索の実験設定においてLLMを用いた多段階クエリ変換によって平均nDCGが向上した．

キーワード　情報検索，大規模言語モデル，クエリ変換

1　はじめに

近年，楽天やAmazonといったECサイトに購入者によるレビューが数多く掲載されている．レビュー閲覧者はそれらのレビューから商品に対する理解を深め，次の意思決定の参考にすることができる．そのレビューの中でも，具体的なレビューは商品に対する理解を深めるのに大変有益である．例えば，あるワイヤレスインターフォンについて理解を深めたいと考えている際に，「受信器と送信器との距離を3mまで近づけても反応しない.」といったレビューや「1日に1回人が来るか来ないかという程度なのに，2週間で電池切れになった.」といったレビューがあれば，ワイヤレスインターフォンが持つ電波強度や省電力性などの性能に対して理解を深めることができる．

しかし，楽天やAmazonなどのECサイトに掲載されているレビューには，ワイヤレスインターフォンに対して「設置したけど反応しない.」といった，抽象的で商品に対する理解を深めるのに不十分なレビューも含まれている．たとえば，このレビューからワイヤレスインターフォンが持つ電波強度について理解を深めることは難しい．なぜならば，受信器や送信器を設置した場所や受信器と送信器との距離について述べられていないためである．そのため，これらについて述べられている具体的なレビューを探し出す必要があるが，レビューを1つ1つ人が確認していくのは手間と時間がかかる．

この問題を解決するための1つの方法として，ある商品に対する抽象的なレビューから，より具体的なレビューを検索するというものが考えられる．しかし，既存の単純な疎検索，密検索では抽象的なレビューから具体的なレビューを検索することが難しいと考えられる．なぜならば，疎検索，密検索には抽象的なレビューから具体的なレビューを検索する仕組みが組み込まれていないからである．単純な疎検索は用語の一致度をもとに検索するため，抽象的なレビューに含まれている用語を含むレビューを取得することはできるが，これらが具体的である保証はない．一方で，単純な密検索は意味的な類似度をもとに検索するため，抽象的なレビューと意味的に類似しているレビューを取得することはできるが，これらも具体的である保証はない．

そこで，本論文では抽象的なレビューからより具体的なレビューを検索すること，つまり，具体性を考慮した検索手法を提案する．具体性を考慮した検索が可能になれば，単純な疎検索，密検索で得られる検索結果よりさらに具体的なレビューを含んだ検索結果を提示することができ，レビュー閲覧者がより簡単に，早く商品に対して理解を深めることができる．

既存研究では，一般的な文書検索においてLLMを用いたクエリ変換は効果があるとされている[4] [6]．しかし，具体性という文書の認知的特性を考慮した検索においてLLMを用いたクエリ変換が有効かは示されていない．そのため，本研究では以下のリサーチクエスチョン（RQ）の検証を行う．

- RQ：LLMを用いた具体性を考慮したクエリ変換はより具体的なレビューの検索において有効か

本研究のクエリ変換では，抽象的なレビューをもとに仮想の具体的なレビューを生成する．しかし，LLMを用いたクエリ変換では，クエリ変換の過程で実在する文書には含まれない語彙を含む内容が生成され，検索精度に影響を与える可能性が指摘されている[4]．この問題に対応するために，本研究ではクエリ変換の際に実レビューを与える．まず，実レビューを得るために，実レビューを与えずにLLMによるクエリ変換（1段階目のクエリ変換）を行い，検索を行う．次に得られたレビューランキングの上位t件の実レビューを与え，再度LLMによるクエリ変換（2段階目のクエリ変換）を行う．このようにして得られたクエリを実際の検索に用いる．

提案手法の有効性を検証するため，楽天レビューデータセットをもとに，ある抽象的なレビューに対するレビューの適合性や具体性をスコア化した評価データを用意した．実験では，こ

の評価データを用い，クエリ変換なしの検索，擬似適合性フィードバックの 2 つのベースライン手法と 1 段階のクエリ変換を用いた検索，2 段階のクエリ変換を用いた検索の 2 つの提案手法を比較し，提案手法の有効性を検証した．

実験を行った結果，次のことがわかった．

- LLM を用いた具体性を考慮したクエリ変換がより具体的なレビューの検索において有効であること
- 密検索において実レビューを用いた 2 段階のクエリ変換が有効であること

本論文の構成は以下のとおりである．2 節では文書や単語の具体性の予測，レビューの有用性の予測と LLM を用いたクエリ変換の関連研究について述べる．3 節では問題設定と LLM を用いたクエリ変換手法について述べる．4 節では評価用データの作成方法，実験設定やその結果について述べる．最後に，5 節で本論文のまとめと今後の課題について述べる．

2 関 連 研 究

本節では，文書や単語の具体性の予測，レビューの有用性の予測，クエリ変換の関連研究について述べる．

2.1 文書や単語の具体性の予測

文書や単語の具体性の予測に関する研究はさまざまな手法を用いて取り組まれている．

たとえば，Tanaka ら [8] は，サポートベクタ回帰を用いた単語の具体性を予測する方法と，それを文書レベルに拡張する方法を提案した．彼らは，単語の具体性を，単語の知覚可能性やイメージのしやすさをもとに定義した．単語のイメージのしやすさを表す特徴量には，画像共有サイトにおける単語の頻度や単語の極性などが用いられている．本研究では Tanaka らにならい，具体性をイメージのしやすさと定義した．また，Turney ら [11] は代表的な抽象的な語，具体な語と単語との類似度を用いた単語の具体性の予測が有効であることを示した．Charbonnier ら [1] は，単語の具体度を予測するモデルが英語だけでなくドイツ語でも有効であることを示した．Martínez ら [7] は GPT-4o のフレーズの具体性の予測が人間の具体性評価と高い相関（$r = 0.8$）を持つことを示した．彼らは具体性が高い語と低い語をプロンプトを通してモデルに渡すことによって，具体性の基準を明確にしている．

既存の研究と本研究の違いについて述べる．本研究では抽象的なレビューがクエリとして与えられる．既存の研究では，このクエリなどの基準を設けない，絶対的な具体性の予測を行っているが，本研究ではクエリという基準を考慮した，相対的な具体性に着目して研究に取り組む．絶対的な具体性では，クエリに関わらず同じ具体性のスコアを計算するため，検索意図を考慮したランキングを生成することが難しい．そこで本研究では，クエリとなった抽象的なレビューと比べてより具体的なレビューを取得することで検索意図を考慮したランキングを求めることを目指す．

2.2 レビューの有用性の予測

EC サイトやレビューサイトで誰もが自由にレビューを記載できるようになった現代では，多くの人が商品レビューを元に購入するかなどの意思決定を行っている．しかし，閲覧者にとって有用なレビューを書くことは難しい上に，手間がかかるため，すべてのレビューが有用であるわけではない．レビューの有用性の予測を行うことは，有用なレビューを提示することで意思決定の支援に繋がるものと考えられる．レビューの有用性の予測に関する研究はさまざまなものがある．

たとえば，Fan ら [3] は，レビューの有用性の予測に，製品のメタデータ（タイトル，ブランド，カテゴリ，説明など）を適用することで予測精度が向上することを示した．Tang ら [9] は，レビューの書き手とレビュー評価者との関係性を考慮することでレビューの有用性の予測精度が向上することを示した．Chen ら [2] は，CNN を用いた商品カテゴリーに依存しないレビューの有用性予測モデルを提案した．

既存の研究と本研究の違いについて述べる．本研究では，具体性に着目しており，有用性とは異なる．しかし，具体度が高いレビューは商品理解を深めることに役立つと考えられるため，具体性と有用性との間には関連性がある可能性がある．そのため，本研究でより具体的なレビューが検索できると，有用なレビューの発見に役立つと考えている．

2.3 LLM を用いたクエリ変換

クエリ変換とは，ユーザが入力したクエリをより関連した文書を検索できるように書き換えることである．近年では，LLM を利用したクエリ変換が注目を集めている．

たとえば，Gao ら [4] は，クエリに対する仮想の回答文書を LLM で生成し，検索に活用する HyDE という手法を提案し，検索精度が向上することを示した．しかし，検索精度は LLM で生成される仮の回答文書に依存しているため，LLM の学習時に十分な量の学習データを用意できなかったドメインや特定のタスクでは検索精度が低下することが指摘されている．

そこで Lei ら [6] は入力されたクエリから，検索対象コーパス中のクエリと関連する文を抽出し，抽出した文を利用したクエリ拡張手法を提案した．このように，実際の文を利用してクエリ拡張することにより，LLM のみを用いてクエリ拡張をしたときと比べて検索精度が向上することがわかっている．

本研究でも，Lei らの研究と同様の考えを用いる．Lei ら [6] は，BM25 による検索結果を LLM によって処理することでクエリと関連する文を抽出しているが，本研究では，LLM を用いてクエリ変換を行い，得られた検索結果からクエリと関連するレビューを取得する．

3 多段階のクエリ変換によるレビューランキング

本節では，抽象的なレビューからより具体的なレビューを検索するタスクの問題設定について述べる．次に，提案手法である LLM を用いた多段階のクエリ変換手法について述べる．

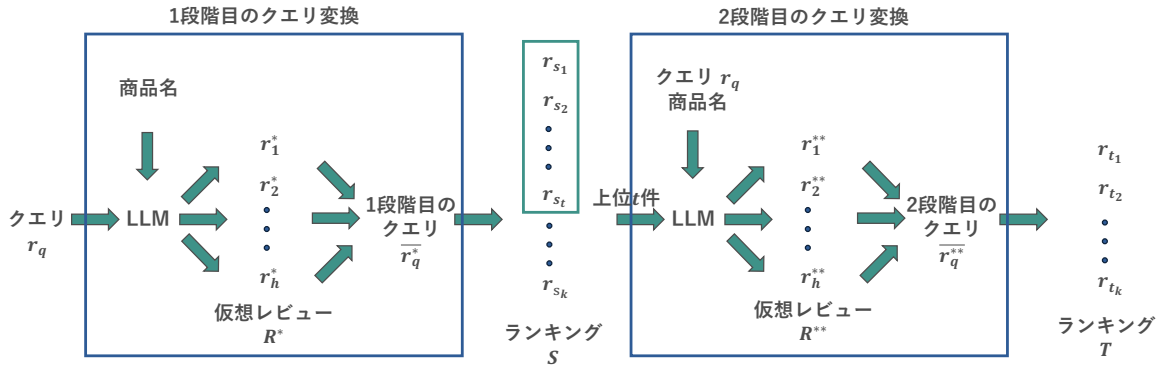

図 1　多段階のクエリ変換によるレビューランキング.

3.1　問題設定

本研究では，抽象的なレビューからより具体的なレビューを検索するタスクを定義する．ここでの具体的とは，2.1 節にて述べたように，商品の特徴について明確，かつ容易にイメージできることを指す.

ある商品に対するレビュー集合 $R = \{r_1, r_2, ..., r_n\}$ 及び，抽象的なレビュー（クエリ）$r_q \in R$ が与えられる．本研究では，このクエリ r_q からより具体的なレビューを含むランキング $T = [r_{t_1}, r_{t_2}, ..., r_{t_k}]$ を求めることを目指す．ここで，$t_i (1 \leqq i \leqq k)$ は $t_i \in [1, k]$ であり，それぞれの順位に対応するレビューの添字を表す．ランキング T はクエリ r_q の検索意図に合致しており，かつレビュー対象商品に対するイメージが明確な順に並んでいる.

3.2　手法の流れ

本節では，LLM を用いた多段階のクエリ変換手法について説明する．まず，図 1 にて提案手法の全体像を示す．抽象的なレビューからより具体的なレビューを検索するために，抽象的なレビューから仮想の具体的なレビューを LLM を用いて生成する．提案手法では 2 段階の手順を踏む．1 段階目では，実レビューを用いずに仮想の具体的なレビューを生成し，検索を行う．2 段階目では，1 段階目の検索で得られた実レビューを用いて仮想の具体的なレビューを生成し，検索を行う.

3.3　1 段階目のクエリ変換によるレビューランキング

1 段階目のクエリ変換によるレビューランキングでは，入力としてクエリ r_q，商品名を LLM にプロンプトを通して与える．以下に使用したプロンプトを示す.

> 次の商品に対するレビューをより具体的にして下さい．レビューにはレビュー本文のみ書いてください．ここでの具体的とは，明確にイメージできることを指します.
> ###商品
> ＜商品名＞
> ###レビュー
> ＜クエリ r_q ＞
> ###出力

なお，出力形式を json 形式とする文章もプロンプトに含めている．このプロンプトをもとに h 回 1 件ずつ仮想の具体的なレビューを生成することで，仮想の具体的なレビュー集合 $R^* = \{r_1^*, r_2^*, ..., r_h^*\}$ を求める．次に生成した h 件の仮想の具体的なレビューをもとに検索に使用するクエリ $\overline{r_q^*}$ を求める．$\overline{r_q^*}$ の求め方は検索方法が密検索か疎検索かによって異なる．詳細は 3.5 節にて記載する．最後に求めたクエリ $\overline{r_q^*}$ を用いてレビューランキング $S = [r_{s_1}, r_{s_2}, ..., r_{s_k}]$ を求める．ここで，$s_i (1 \leqq i \leqq k)$ は $s_i \in [1, k]$ であり，それぞれの順位に対応するレビューの添字を表す.

3.4　2 段階目のクエリ変換によるレビューランキング

2 段階目のクエリ変換によるレビューランキングでは，入力としてクエリ r_q，商品名，3.3 節で求めたレビューランキング S の上位 t 件のレビューを LLM にプロンプトを通して与える．以下に使用したプロンプトを示す.

> 次の商品に対するレビューをより具体的にして下さい．レビューにはレビュー本文のみ書いてください．ここでの具体的とは，明確にイメージできることを指します．必要であれば以下のレビューを参考にして下さい.
> ###参考レビュー
> sample_review 1: ＜ $r_{S_1} \in S$ ＞
> ⋮
> sample_review t: ＜ $r_{S_t} \in S$ ＞
> ###商品
> ＜商品名＞
> ###レビュー
> ＜クエリ r_q ＞
> ###出力

なお，出力形式を json 形式とする文章もプロンプトに含めている．これらの入力をもとに，3.3 節と同様にして仮想の具体的なレビュー集合 $R^{**} = \{r_1^{**}, r_2^{**}, ..., r_h^{**}\}$，クエリ $\overline{r_q^{**}}$ を求める．最後に求めたクエリ $\overline{r_q^{**}}$ を用いてレビューランキング $T = [r_{t_1}, r_{t_2}, ..., r_{t_k}]$ を求める.

3.5　仮想レビュー集合からクエリを求める方法

本節では，仮想のレビュー集合 R'（R^* または R^{**}）からクエリ q'（$\overline{r_q^*}$ または $\overline{r_q^{**}}$）を求める方法について述べる．クエリ

表 1　評価に用いたクエリの一覧.

ID	クエリ	検索意図	レビュー件数
ワイヤレスインターフォン 1	受信器をすぐ近くに置かないと反応しません 失敗	送受信可能，不可能な受信器と送信器との距離についてより具体的に知りたい	106
ワイヤレスインターフォン 2	対応は速く二日で届きましたが，受信機，送信機共に裏蓋で苦労したので☆ 3 つで.	裏蓋の特徴についてより具体的に知りたい	106
除湿機 1	除湿能力は想像以上に低いです． もう少し予算を高くして除湿能力が高いタイプをおすすめします.	除湿能力についてより具体的に知りたい	101
除湿機 2	タンクがすぐに満水になって面倒です．デザインは気に入ってますが，買って失敗しました.	どれくらいで満水になるのかより具体的に知りたい	101
ヘアドライヤー 1	ちょっと音はうるさいけど，早く乾いて便利です.	音の大きさについてより具体的に知りたい	44
ヘアドライヤー 2	予想以上に大きかったです… 風が冷たくちょっと残念だったかも…	ドライヤー本体の大きさについてより具体的に知りたい	44
電気ケトル 1	お値段以上の使い易さでデザインもオシャレでした！	デザインについてより具体的に知りたい	37
電気ケトル 2	デザインがとてもよく気分が上がる．注ぎ口の感じが実に良い．ハンドドリップ用に購入．持った感じ重みがあるけど取っ手の握り心地もいい.	どんな感じの注ぎ口なのかより具体的に知りたい	37

表 2　密検索 (ruri-large) における各手法での nDCG@10．太字の数値は手法の中で最も精度が高いことを表す．() の中の値は標準偏差を表す.

	ベースライン		提案手法	
クエリ ID	クエリ変換なし	擬似適合性フィードバック	1 段階のクエリ変換	2 段階のクエリ変換
ワイヤレスインターフォン 1	0.286	0.341	**0.450**(0.035)	0.442(0.048)
ワイヤレスインターフォン 2	0.292	0.323	0.391(0.081)	**0.491**(0.055)
除湿機 1	0.153	0.153	0.433(0.036)	**0.591**(0.034)
除湿機 2	0.063	0.058	0.161(0.061)	**0.174**(0.069)
ヘアドライヤー 1	0.264	0.215	**0.397**(0.061)	0.281(0.058)
ヘアドライヤー 2	0.096	0.096	**0.153**(0.018)	0.144(0.029)
電気ケトル 1	0.396	0.306	0.517(0.076)	**0.530**(0.064)
電気ケトル 2	0.436	0.394	0.423(0.049)	**0.439**(0.012)
平均	0.248	0.236	0.366(0.138)	**0.386**(0.164)

q' の求め方は検索方法が密検索か疎検索かによって異なる．

検索方法が密検索の際は，HyDE [4] と同様に仮想のレビュー R' の平均ベクトルをクエリ q' とする．以下に式を示す.

$$q' = \frac{1}{h} \sum_{r'_i \in R'} \boldsymbol{r'_i} \tag{1}$$

なお，$\boldsymbol{r'_i}$ は r'_i の埋め込み表現を表す.

検索方法が疎検索のときは，検索対象とする語彙を増やすために仮想のレビュー R' を結合したものをクエリ q' とする.

4 実　　験

本節では，まず，評価データについて述べる．次に，ベースライン手法を含む実験設定について紹介し，実験結果を示す．最後に実験結果に対する考察を述べる.

4.1 評価データの作成

本節では，評価データの作成方法について述べる．評価データ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 [12] を利用した．本研究ではその中でも 4 つの家電製品に対するレビューを用いている．それぞれの商品でクエリとなる抽象的なレビュー r_q と検索意図の組を 2 件ずつ計 8 件用意した．表 1 にクエリの一覧を示す．レビュー件数は検索対象のレビューの数を表す.

本研究では手法の評価のために nDCG を用いた．nDCG の計算のために必要な各レビューの利得は，クエリの検索意図，クエリ r_q と比べたときのイメージのしやすさをもとに以下の

表 3　疎検索 (BM25) における各手法での nDCG@10．太字の数値は手法の中で最も精度が高いことを表す．() の中の値は標準偏差を表す．

	ベースライン		提案手法	
クエリ ID	クエリ変換なし	擬似適合性フィードバック	1 段階のクエリ変換	2 段階のクエリ変換
ワイヤレスインターフォン 1	0.369	0.293	0.445(0.055)	**0.448**(0.059)
ワイヤレスインターフォン 2	0.454	0.367	**0.509**(0.024)	0.488(0.012)
除湿機 1	0.334	0.382	0.416(0.047)	**0.458**(0.038)
除湿機 2	0.144	**0.354**	0.320(0.073)	0.290(0.041)
ヘアドライヤー 1	0.202	0.317	0.507(0.029)	**0.515**(0.032)
ヘアドライヤー 2	0.135	**0.271**	0.229(0.059)	0.189(0.064)
電気ケトル 1	**0.505**	0.433	0.476(0.068)	0.443(0.038)
電気ケトル 2	0.440	**0.444**	0.411(0.046)	0.382(0.062)
平均	0.323	0.358	**0.414**(0.105)	0.402(0.113)

ように設定した．

- 0: クエリの検索意図とレビュー内容が合致していない．もしくはクエリの検索意図とレビュー内容は合致しているが，クエリと比べて使用状況がイメージしにくい．
- 1: クエリの検索意図とレビュー内容が合致しており，クエリと比べて使用状況がイメージしやすい．

上記の基準に従って利得の設定を行うために，各レビューの検索意図に対する適合性，クエリと比べたときのイメージのしやすさを 3 人の評価者によって評価した．実験に用いた評価データの検索意図に対する適合性とイメージのしやすさは評価者の多数決によって設定した．各レビューの検索意図に対する適合性では，クエリ r_q に対する適合性の有無を 2 値で評価した．各レビューのクエリと比べたときのイメージのしやすさでは，各レビューがクエリよりイメージしやすいかを 2 値で評価した．ここでは検索意図に適合していたレビューに対してのみ評価を行い，適合していなかったレビューに対しては無条件に 0 とした．このクエリと比べたときのイメージのしやすさを nDCG の計算に用いる．

以上の内容で評価を行った結果，適合性の評価に関しては Fleiss の κ 係数は 0.807 となり，ほぼ完全な一致（Almost）となった [5]．イメージのしやすさの評価に関しては Fleiss の κ 係数は 0.614 となり，かなりの一致（Substantial）となったが，3 人の適合性に対する評価が 1 で共通したものに限定した場合の Fleiss の κ 係数は 0.415 となり，適度な一致（Moderate）となった．

4.2 実験設定

提案手法の有効性を示すために，次の手法を用いて実験を行った．

(1) クエリ変換なしのレビュー検索

これは本実験のベースラインとなる手法の 1 つである．抽象的なレビュー r_q をクエリとし，検索を行う．

(2) 擬似適合性フィードバックによるレビュー検索

これは本研究のベースラインとなる手法の 1 つである．まず，抽象的なレビュー r_q をクエリとし，検索を行う．検索結果の上位 5 件を適合レビュー集合 R_+ だとみなし，r_q のクエリ拡張に用いる．クエリの拡張には R_+ と r_q の和集合を R' として 3.5 節の方法を用いる．最後に，得られたクエリ q_s を用いて検索を行う．

(3)1 段階のクエリ変換によるレビュー検索

3.3 節で得られたレビューランキング S をクエリ r_q に対する出力 T とする手法である．生成する仮想レビューの件数 h は 5 件に設定した．

(4)2 段階のクエリ変換によるレビュー検索

3.4 節で得られたレビューランキング T をクエリ r_q に対する出力 T とする手法である．生成する仮想レビューの件数 h は 5 件に，2 段階目のクエリ変換に利用する実レビューの件数 t は 2 件に設定した．

各手法において疎検索と密検索の 2 種類の検索方法を試す．疎検索には Python ライブラリである rank_bm25 の BM25 を，密検索には BERT ベースの，日本語に特化したテキスト埋め込みモデルである ruri-large [10] を用いる．手法 (3)，(4) においてクエリ変換を行う LLM には elyza/Llama-3-ELYZA-JP-8B[1]を用いる．LLM のパラメータは HyDE [4] を参考に，temperature を 0.7 に，top_p を 1 に設定した．

構築したデータセットを用いた実験の手順を以下で述べる．クエリとなった抽象的なレビューをもとにそれぞれの手法で検索を行う．各手法の性能評価は nDCG@10 で行う．なお，手法 (3)，(4) は実行する度に異なるランキング結果が得られるため，10 回実行した結果の平均 nDCG で評価を行う．

4.3 実験結果

まず表 2，3 に，本研究で作成した評価データにおける各手法の検索精度を示す．表 2 は密検索での結果，表 3 は疎検索での結果を示している．平均の行の数値は，手法ごとの nDCG の平均値を表す．

密検索では，すべてのクエリにおいて提案手法である LLM を用いた具体性を考慮したクエリ変換による検索がベースライン手法を上回った．提案手法の中でも 2 段階のクエリ変換によるレビュー検索の平均精度が一番高かった．しかし，すべての

1：https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B

表 4　除湿 1 において 1 段階のクエリ変換，2 段階のクエリ変換で生成された仮想のレビュー．

No	1 段階	2 段階
1	この除湿機の除湿能力は，個人的には想像以上に低いと感じました．実際，梅雨の時期に使用しましたが，思った程湿度が下がらず，除湿の効果を実感することができませんでした．予算を高くして，除湿能力が高いタイプを購入することをおすすめします．	乾燥能力がイマイチで，予算を高くして除湿能力の高いタイプを勧めます．想像以上に除湿能力が低く，十二分に除湿する能力を発揮できません．例えば，タオル 12 枚をバーガーに掛けて下に除湿器を置いて 8 時間かけて動作してみたが，タオルは全く乾きませんでした．水はタンクに少しだけ貯まりました．除湿能力の低さにがっかりしました．
2	除湿能力は想像以上に低く，想像より早く湿度が上がってしまいました．特に夏場は効果をあまり感じられず，部屋がジメジメした状態でした．除湿機を置く場所も工夫して，風を当ててみたりしましたが，思うほどの除湿効果はありませんでした．もう少し予算を高くして除湿能力が高いタイプを購入することをおすすめします．	この除湿機の除湿能力は想像以上に低く，部屋干しで使用した洗濯物が水分を含んだままです．私は 8 時間使用しましたが，タオル 12 枚が全く乾きませんでした．水はタンクに少し貯まっていますが，除湿能力が低いせいでしょうか？また，機械音がして送風は出ていますが，音が聞こえます．静音性能も期待外れです．除湿能力が高くて予算が許すなら，もう少し高額な商品を選ぶことをおすすめします．
3	除湿能力は想像以上に低く，部屋干しで洗濯物がなかなか乾かなかった．除湿機が小型なため，風が弱く，湿度が高めの環境では効果が半減する印象．梅雨時期の結露対策として使用するには，除湿能力が低く，不向きである．	この除湿機は思った以上に除湿能力が低く，正直がっかりしました．具体的には，12 枚のタオルを 8 時間かけて除湿を行ったのですが，水は少し貯まっていたものの，タオルは全く乾いていませんでした．機械音は静かで，デザインも良かったのですが，除湿能力が低い点が大きなマイナスポイントです．予算に余裕があれば，除湿能力が高いタイプを選ぶことをおすすめします．
4	除湿能力は想像以上に低く，乾燥させたい洗濯物が思うように乾かず，不満が残る．特に梅雨時期の湿気はさほど除去できなかった．静音と省エネの点は良かったが，除湿能力の低さが全てを台無しにした．	除湿能力は想像以上に低く，実際に使用してみると，タオル 12 枚をバーに掛けて下に除湿器を置いて 8 時間動作させても，乾くどころか水がタンクに少し貯まってしまうほどです．最低限の除湿能力しかないため，衣類を乾燥させることは困難でした．
5	除湿能力は想像以上に低く，湿った空気を吸い込むのではなく，機内を循環している空気の水分を除去する感じです．梅雨時期の除湿には向いていないと思います．	この除湿機は予想以上に除湿能力が低いと感じた．具体的には，タオル 12 枚をハンガーに掛けて 8 時間運転してみたが，全く乾いておらず，水は少し溜まっていた．梅雨の時期に大活躍してくれるのを期待して購入したが，除湿能力の低さにがっかりしている．

クエリにおいて 2 段階のクエリ変換によるレビュー検索の精度が他の手法の検索精度を上回るという結果は得られなかった．たとえば，ヘアドライヤー 1 において 2 段階のクエリ変換によるレビュー検索の nDCG は 0.281 であったが，1 段階のクエリ変換によるレビュー検索の nDCG は 0.397 であった．

疎検索では，平均精度において提案手法である LLM を用いた具体性を考慮したクエリ変換による検索がベースライン手法を上回った．提案手法の中でも 1 段階のクエリ変換によるレビュー検索の平均精度が一番高かった．しかし，一部のクエリでは 1 段階のクエリ変換によるレビュー検索の精度より 2 段階のクエリ変換によるレビュー検索の精度の方が高かったり，ベースライン手法の精度の方が 1 段階のクエリ変換によるレビュー検索の精度より高かったりした．たとえば，除湿機 1 において 1 段階のクエリ変換によるレビュー検索の nDCG は 0.416 であったが，2 段階のクエリ変換によるレビュー検索の nDCG は 0.458 であった．また，ヘアドライヤー 2 において 1 段階のクエリ変換によるレビュー検索の nDCG は 0.229 であったが，擬似適合性フィードバックによるレビュー検索の nDCG は 0.271 であった．

4.4 考　　察

本節では前節で述べた実験結果に対する考察を行う．検索の平均精度において，提案手法である LLM を用いた具体性を考慮したクエリ変換による検索が密検索，疎検索ともにベースライン手法を上回ったことから，一般に，LLM を用いた具体性を考慮したクエリ変換が具体的なレビューの検索において有効であることがわかった．

これは LLM 内部の具体性に関する知識を検索クエリに反映することができたためであると考える．たとえば，除湿機 1 では，1 段階のクエリ変換において生成された仮想のレビュー集合 R^* は表 4 のようになった．このときのクエリ r_q は「除湿能力は想像以上に低いです． もう少し予算を高くして除湿能力が高いタイプをおすすめします.」であり，どんな除湿能力があるかについては記述されていないが，生成された仮想のレビューでは梅雨という時期に関する語彙や洗濯物という利用シーンに関する語彙を補うことができている．その結果，クエリ変換なしの密検索において 1 位にランクインしたレビューは「コスパ

表 5　ヘアドライヤー 1 において 2 段階のクエリ変換で生成された仮想のレビュー.

No	仮想レビュー
1	速乾性はとても便利で，最大風量で使うと特に早く乾きます．ただ，音はやや大きく，うるさいと感じるレベルです．例えば，集合住宅で窓を開けて使用する場合，音の大きさは注意が必要なレベルです．音の大きさは，最大風量にすると特にうるさく，中くらいの風量にすると，より使いやすくなります.
2	使ってみると，音は少しうるさいものの，乾燥時間が短くてとても便利です．風量を最大にすると特に早く乾きますが，その分音がかなり大きく，集合住宅では窓を閉めて使用するか，時間を指定して使用する必要がありそうです．ただ，風量を中程度に抑えれば，音はややマシになり，乾燥時間も短くて助かります．音がもう少し静かだったら，文句なしの商品です.
3	使ってみると，音はうるさいです．最大風量で使うと，特に音が大きくてびっくりします．集合住宅で窓を開けて使うのは，注意が必要なレベルの音量です．ただ，早く乾くのはとても便利です．多少の音のうるささを我慢すれば，快適に使うことができます.
4	乾燥時間はやや短くて便利ですが，音はうるさいです．特に最大風量で使用すると，音がものすごく大きくて，集合住宅で窓を開けて使う時は注意が必要なレベルです．中間の風量で使用することで，音は多少小さくなりますが，乾きは早く，十分に使えます．ただし，音がもう少し静かだったら，より使いやすくなるという点が残念です.
5	音はうるさいが，最大風量で使用するとあっという間に乾くため，便利に使えている．音の大きさは，集合住宅で窓を開けて使用する場合，要注意レベルのうるささだ.

最高です．少し重いですが，操作も簡単で除湿力も凄いです！」であったが，1 段階のクエリ変換を用いた密検索では「デザインも良く 静かで，価格が良く 乾かす能力が今まで買った除湿器の中でも 最低でした． 機械音がして，送風は出ています． しかし，タオル 12 枚バーガーに掛けて下に 除湿器を置いて 8 時間かけて動作して，乾いていません．水はタンクに少し貯まっています． 故障でしょうか？ ただ，除湿能力が低いのでしょくか？ 疑問です.」というレビューが 1 位にランキングされた.

また，密検索において 2 段階のクエリ変換によるレビュー検索の平均精度が 1 段階のクエリ変換によるレビュー検索の平均精度を上回ったことから，一般に，密検索において，具体性を考慮したクエリ変換の際に実レビューを与えることは有効であることがわかった．これは，実レビューの内容から実レビュー中の語彙や表現に関する知識を得ることができたためであると考える．たとえば，除湿機 1 では，2 段階のクエリ変換において生成された仮想のレビュー集合 R^{**} は表 4 のようになった．1 段階のクエリ変換において生成されたレビューには，「湿度が上がってしまいました.」や「機内」など除湿能力について調べている際に出現しないであろう表現が含まれていた．一方で，2 段階のクエリ変換において生成されたレビューには「タオル」や「8 時間」などの時間といった除湿能力について調べている際に出現しそうな表現が多く含まれていた.

しかし，ヘアドライヤー 1 のように 1 段階のクエリ変換によるレビュー検索の精度の方が 2 段階のクエリ変換によるレビュー検索の精度より高い場合もあった．これは，生成される仮想のレビューの内容が実レビューの語彙に大きく影響を受けるためであると考える．ヘアドライヤー 1 での 2 段階のクエリ変換によって生成された仮想レビュー集合 R^{**} を表 5 に示す．クエリ変換の際に渡されたレビューの 1 つに「集合住宅で窓を開けて使うときは，要注意レベルの音の大きさです.」という記述が含まれていたため，2 段階のクエリ変換にて，集合住宅についての内容が生成されていた．その結果，1 段階のクエリ変換によるレビュー検索ではランクインしていた「店頭でこの大音量を聞いていたら購入しませんでした（笑）おまけにスタンドが付属されているのが納得の重量． しかしながら，音量に伴って風量も凄まじく，子供に逃げられる前にロングヘアを乾かしきることができるので大活躍しています！ 前髪を乾かす時には呼吸が苦しくなる程の風量．周りの音を完全にかき消す大音量．二の腕痩せに効果がありそうなダンベル的ドライヤーを想像した上で，子沢山でお風呂タイムが戦争！なご家庭にはぜひオススメします！ (^^)」というレビューが 2 段階のクエリ変換によるレビュー検索ではランクインしなかった.

疎検索において一般に，2 段階のクエリ変換によるレビュー検索の精度より 1 段階のクエリ変換によるレビュー検索の精度が高かったのは，検索クエリのレビュー全体に対する適合性を保つことができたからだと考える．2 段階のクエリ変換では実レビューの内容を踏まえて仮想のレビューを生成したため，表 4， 5 でにある通り，実レビューの内容に生成内容が大きく影響を受けた．結果，用語の一致で検索を行う疎検索では密検索と比べて特に，与えられた実レビューを検索するのに特化したクエリ変換が行われ，2 段階のクエリ変換によるレビュー検索では検索精度が下がってしまったと考える.

5　まとめと今後の課題

本論文では，抽象的なレビューをもとにより具体的なレビューを提示し，商品に対する理解を深める支援を行う問題に取り組んだ．この問題では，抽象的なレビューだけでは，商品に対して不明瞭な点があるため，それらを解決する，より具体的なレビューを提示する必要があった．そこで，不明瞭な点を解決するより具体的なレビューを，抽象的なレビューをもとに検索するタスクを定義し，LLM を用いた多段階のクエリ変換によるより具体的な商品レビューの検索手法を評価した.

実験では，ベースライン手法としてクエリ変換なしのレビュー検索，擬似適合性フィードバックによるレビュー検索を，提案

手法として 1 段階のクエリ変換によるレビュー検索と 2 段階のクエリ変換によるレビュー検索を用いた．実験を通じて，LLM を用いた具体性を考慮したクエリ変換がより具体的なレビューを検索する上で有効であること，中でも密検索の実験設定で実レビューを用いた 2 段階のクエリ変換が有効であること，クエリ変換時に入力される実レビューの語彙に強く影響を受けることがわかった．今後の課題としては，仮想のレビューの生成件数や，2 段階のクエリ変換にて与える実レビューの件数などのハイパーパラメータの検証を行うこと，一部のレビューに対してではなく，対象商品のレビュー全体に適合したクエリ変換を可能にすることがある．

謝　　辞

本研究は JSPS 科学研究費助成事業 JP24K03228, JP21H03775 による助成を受けたものです．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] Jean Charbonnier and Christian Wartena. Predicting the concreteness of german words. In *Proceedings of the 5th Swiss text analytics conference & 16th conference on natural language processing, Vol. 2624*, 2020.

[2] Cen Chen, Yinfei Yang, Jun Zhou, Xiaolong Li, and Forrest Sheng Bao. Cross-domain review helpfulness prediction based on convolutional neural networks with auxiliary domain discriminators. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 602–607, 2018.

[3] Miao Fan, Chao Feng, Lin Guo, Mingming Sun, and Ping Li. Product-aware helpfulness prediction of online reviews. In *The World Wide Web Conference*, pp. 2715–2721, 2019.

[4] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1762–1777, 2023.

[5] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, Vol. 33, No. 1, pp. 159–174, 1977.

[6] Yibin Lei, Yu Cao, Tianyi Zhou, Tao Shen, and Andrew Yates. Corpus-steered query expansion with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 393–401, 2024.

[7] Gonzalo Martínez, Juan Diego Molero, Sandra González, Javier Conde, Marc Brysbaert, and Pedro Reviriego. Using large language models to estimate features of multi-word expressions: Concreteness, valence, arousal. *Behavior Research Methods*, Vol. 57, No. 1, pp. 1–11, 2025.

[8] Shinya Tanaka, Adam Jatowt, Makoto P. Kato, and Katsumi Tanaka. Estimating content concreteness for finding comprehensible documents. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pp. 475–484, 2013.

[9] Jiliang Tang, Huiji Gao, Xia Hu, and Huan Liu. Context-aware review helpfulness rating prediction. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pp. 1–8, 2013.

[10] Hayato Tsukagoshi and Ryohei Sasano. Ruri: Japanese general text embeddings. *arXiv preprint arXiv:2409.07737*, 2024.

[11] Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 680–690, 2011.

[12] 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.