対話型検索における応答長が ナゲットレベルの賛否分布に与える影響の分析

熊野　桃† 飛岡　憲†† 山本　岳洋††

† 兵庫県立大学 社会情報科学部　〒 651–2197 兵庫県神戸市西区学園西町 8–2–1
†† 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: †{fa22n028,af24h009}@guh.u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp

あらまし　本研究では，対話型検索におけるシステムの応答長に着目し，応答に含まれる情報の賛否分布が応答の長短によってどのように変化するかをナゲットレベルで分析した．システムの応答に含まれる賛否情報の分布は，ユーザの意見形成に影響すると考えられる．しかし，実運用における応答長の調整がその分布に与える影響は明らかではない．そこで本研究では，「SNS は私たちの社会にとって良いものか」のような賛否両論ある 5 つのトピックを対象に，大規模言語モデル（LLM）の応答長と応答に含まれる情報の賛否分布の関係を分析した．まず，賛成，中立，反対の立場をとる質問クエリを各 10 件作成し，LLM に長い応答（長応答）と短い応答（短応答）の 2 種類を生成させた．その後，得られた応答をナゲットに分割し，各ナゲットがトピックに対してどの立場を支持する情報かをラベル付けすることで賛否分布を分析した．分析の結果，短応答では長応答より情報分布に偏りがあり，特に反対クエリにおいてその傾向が顕著であることが明らかになった．また，このような傾向はトピックの性質により程度が異なることも明らかとなった．

キーワード　情報検索，対話型検索，生成 AI，ナゲット

1 はじめに

近年 ChatGPT や Gemini などの大規模言語モデル（LLM）を利用した対話型サービスが広く普及し，それらを利用した情報収集である対話型検索は一般的に行われるようになった．総務省の令和 7 年度情報通信白書 [1] でも，個人の生成 AI サービス利用の拡大や，企業における業務での生成 AI 利用が報告されている．

対話型検索における LLM の出力結果は，ユーザの意見形成に少なからず影響を与えると考えられる．たとえば，「SNS は私たちの社会にとって良いものか？」というトピックについて調べるとき，システムの出力がトピックに対して賛成をサポートする情報が多い場合と，反対をサポートする情報が多い場合とで，検索後のユーザの意見は変わる可能性がある．ユーザからのクエリに対してシステムが検索結果をランキングしたリストを返す従来型検索では，検索結果がユーザの意見に影響を及ぼすことについての研究は多く存在し [2] [3] [4] [5]，対話型検索においても同様の問題に関する研究が行われている [6]．このことから，対話型検索において出力結果がユーザの意見形成に影響を及ぼす可能性を調査することは社会的に重要である．

ユーザの意見形成に影響する要素として，検索結果に含まれる情報の賛否分布が挙げられる．Epstein らは，特定の立場を支持する Web 検索結果がユーザの意見形成に影響する可能性があるとしており [4]，Leung らはユーザの既存の意見による偏ったクエリが検索結果を偏らせ，ユーザの意見の更新が制限されると報告している [7]．このように，検索結果の賛否分布はユーザの意見形成に影響を与えることがある．

対話型検索における情報の賛否分布には，応答文の長さが影響を及ぼす可能性がある．一般的に LLM による応答は具体例や状況説明を含み，長い文章であることが特徴として挙げられる．また，生成 AI サービスの 1 つである ChatGPT のデフォルトの応答スタイルは中立的であるとされており [8]，偏った立場のクエリを入力するとクエリとは反対の立場の情報を応答に含めることがある．しかし応答長を短くすると，文章の長さを抑えたうえでクエリの情報要求に簡潔に答えるため，クエリとは異なる立場の情報を削減することが考えられる．このように，応答文の長さによって対話型検索の情報の賛否分布に影響が出る可能性がある．

本研究では，生成 AI を利用した対話型検索システムの応答長を調整したとき，ユーザの意見形成に影響があるかを調査することを目的としている．その前段階として，対話型検索システムの応答長を調整することによって，出力文に含まれる情報の賛否分布がどのように変化するかを調査した．具体的には，以下のリサーチクエスチョンを明らかにした．

- **RQ1**：長応答と短応答で LLM 応答に含まれる情報の賛否分布は変化するか
- **RQ2**：LLM のモデルによって応答に含まれる情報の賛否分布は変化するか

これらのリサーチクエスチョンを明らかにするため，賛否両論あるトピックに関する質問クエリに対し，2 種類の LLM で長応答と短応答を収集し，情報の賛否分布を分析した．賛否両論あるトピックに対して，賛成，中立，反対の 3 つの立場のクエリを LLM で生成し，それらのクエリに対して長応答と短応答を収集した．得られた LLM の応答をナゲットという情報の単

位に分割し，それらがトピックに対してどの立場を支持する情報かを表すラベルを付与することで，LLM の応答に含まれる情報の賛否分布を量的に分析した．

RQ1 を明らかにするために，各トピックで得られた LLM の長応答，短応答の賛否分布を分析した．その結果，多くのトピックで短応答ではクエリのスタンスと一致するスタンスのナゲットの割合が長応答よりも高くなる傾向がみられ，特に反対クエリに対する反対ナゲットの割合が多くなる傾向があった．また，トピックの性質によって応答に含まれる情報の賛否分布の特徴が異なることも確認された．

RQ2 を明らかにするために，2 つの LLM で長応答，短応答の賛否分布を分析した．本研究で比較した GPT-4o と Gemini 2.5 Flash の 2 つの LLM では，賛否分布に大きな差は見られなかった．

2 関 連 研 究

2.1 検索行動と意見形成に関する研究

文書をクエリへの関連度順にランキングしたリストを結果として表示する Web 検索において，ユーザの検索行動を分析する研究が行われている．Suzuki らは，ウェブ検索におけるユーザの検索行動を調査し，強い確証バイアスを有するユーザは検索結果の上位のみを頻繁に閲覧し，検索を短時間で終了する傾向があることを報告した [2]．また，検索結果を操作するとユーザの意思決定にどう影響するかを調査する研究が行われている．Draws らは，検索結果の表示順序が検索エンジン操作効果（SEME）の要因であるかを調査し，検索結果の表示順序は SEME の直接的な要因ではないが，閲覧した情報の割合が態度変容に寄与する可能性が高いことを明らかにした [5]．Bink らは，Web 検索結果ページ（SERP）の上部に表示され，ユーザがリンクをクリックしなくても内容を把握できるようにするスニペットが，強い事前意見を持たないユーザの意見形成に影響することを明らかにした [3]．Epstein らは，特定の立場を支持するページの多い検索結果と立場の混在する検索結果がユーザの意見に与える影響を比較し，特定の立場を支持する検索結果はユーザの意見形成に影響する可能性があるとした [4]．このように，Web 検索において検索結果の情報分布がユーザの意見形成に影響することが研究されている．ユーザの意見形成への影響の調査は，LLM を利用した対話型検索においても行われている．Sharma らは，LLM を利用した対話型検索システムがユーザの情報検索や意見形成に与える影響を調査し，対話型検索では従来型検索より偏った情報検索が行われやすいことを報告した [6]．また，ユーザは道徳的な話題での意思決定においても影響を受けることが報告されている [9] [10]．

Web 検索と対話型検索において，検索結果の情報分布がユーザに与える影響について研究されている．しかし，対話型検索における LLM の応答については研究がまだ少なく，特に応答長が LLM の応答に含まれる情報の賛否分布に与える影響については研究されていない．本研究では，LLM の応答に含まれる情報の賛否分布が応答長に影響を受けるかを調査し，ユーザの意見形成への影響の分析に繋げることを目指す．

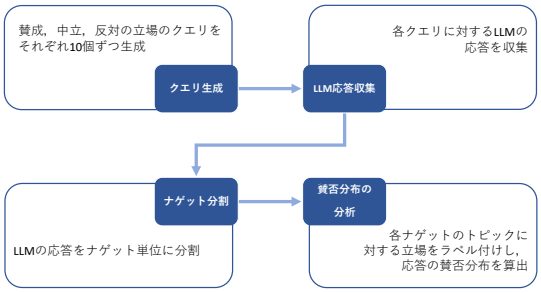


図 1: 本研究の実験概要．

2.2 対話型検索における検索行動の特徴

LLM を利用した対話型検索の普及に伴い，対話型検索におけるユーザの検索行動の特徴について調査する研究が行われている．Huang らは，チャットボット利用時の対話ターン数がユーザの満足度や対話の質に与える影響を調査し，応答ターン数が多いほどユーザの満足度は高まるが，ターン数が必ずしも対話の質の向上には直結しないことを明らかにした [11]．Xu らは，対話型検索と Web 検索を比較し，ユーザは対話型検索に対して従来の Web 検索よりも高い有用性，楽しさ，満足度を感じることを報告した [12]．

以上のように，対話型検索では Web 検索とは異なる検索行動がみられることが研究されている．本研究では，Web 検索ではあまり焦点を当てられてこなかった応答長という概念が，ユーザの検索行動に影響を及ぼすかを分析することを目指している．

3 応答長と賛否分布の関係分析

3.1 実 験 概 要

1 節で述べたリサーチクエスチョンを明らかにするため，本研究では図 1 に示す手順で実験を行った．まず，賛否両論あるトピックについて，賛成，中立，反対の 3 つの立場（スタンス）をとるクエリを生成した．その後，LLM の応答を長応答と短応答の 2 種類収集した．収集した応答をナゲットという情報の単位に分割した．最後に，各ナゲットがトピックに対してどのようなスタンスをとる情報であるかを表すラベルを付与し，賛否分布を分析した．以下で各手順について詳しく述べる．

3.2 トピック

本研究では表 1 に示す 5 つのトピックを使用した．これらのトピックは，賛否両論あるトピックにおけるユーザのバイアスと検索システムのバイアスの相互作用を分析した Draws らの研究 [5] で用いられたトピックを日本語に翻訳したものである．これらのトピックは，既に特定の強い意見を持っている人が少ないトピックとされており [5]，今後のユーザ実験を踏まえてこれらのトピックを採用した．

表 1: 本研究で使用したトピック.

ID	トピック
SNS	SNS は私たちの社会にとって良いものか
携帯	携帯電話の電波は安全か
肥満	肥満は病気か
ボトル	ボトル入り飲料水は禁止されるべきか
動物園	動物園は存続するべきか

表 2: トピック「SNS は私たちの社会にとって良いものか」に対して実際に得られたクエリの例.

スタンス	クエリ
賛成	SNS が私たちの社会に良い影響を与える理由は何ですか？ 災害時に SNS がどのように役立つか例を教えてください。
中立	SNS がコミュニケーションにおいてもたらす効果にはどんなものがありますか？ SNS はどのように社会に貢献し、また社会に害を及ぼしていますか？
反対	SNS を使うことの社会的なデメリットは何ですか？ SNS が個人情報やプライバシーの問題を悪化させている例を教えてください。

3.3 クエリ収集

本研究では，賛否両論あるトピックに対し，様々な考えを持つユーザの情報要求を想定したクエリを扱うため，賛成，中立，反対の 3 つのスタンスからのクエリを使用した．実際のユーザが入力したクエリを収集するのは人的コストがかかるため，表 1 の各トピックに対し，LLM で 3 つのスタンスのクエリをそれぞれ 10 個ずつ生成した．LLM は，OpenAI 社の GPT-4o (2024-11-20)[1] を使用した．実際に得られたクエリを一部抜粋して表 2 に示す．また，以下にクエリ生成のプロンプトを示す．

> あなたは一般的な日本人で、ごく日常的な検索において「{**トピック**}」という話題についての質問を AI エージェントに質問しようとしています。質問クエリを作成してください。賛成の立場から情報を求めるクエリ、中立的な立場から情報を求めるクエリ、反対の立場から情報を求めるクエリそれぞれ 10 個ずつ作成してください。
> 以下にある 2 つの話題でのクエリ例を示します。
> 例 1：ヴィーガンの健康効果
> 賛成：ヴィーガンになると体にはどのような好影響がありますか？
> 中立：ヴィーガンになると健康にはどのようなメリットとリスクがありますか？
> 反対：ヴィーガンの体へのデメリットを教えてください。
> 例 2：シナモンの糖尿病への効果
> 賛成：シナモンには血糖値を下げる効果がありますか？
> 中立：シナモンの体への効果は何ですか？
> 反対：シナモンの健康への効果は限定的ですよね？

実際に使用したプロンプトには，出力形式を指定する指示も含めている．プロンプト内の “**トピック**」” には，表 1 に示したトピックが入力される．実際にユーザが入力するクエリに近づけるため，LLM に一般的な日本人という役割を与え，few-shot でクエリを生成した．few-shot に使用した例は，賛否両論ある 2 つのトピックに対し，著者らがクエリ例を作成したものである．

クエリを LLM で生成して収集するにあたり，内容の重複したクエリが生成されることが懸念された．例として，“災害時に SNS がどのように役立つか教えてください。” というクエリと “SNS は災害時にどのように役立ちますか？” のような，質問している内容が同じであるようなクエリが生成される可能性があった．そこで各トピック，各スタンスのクエリを確認したところ，表 2 のように作成したクエリの内容に重複は見られず，様々な考えを持つユーザの情報要求を想定したクエリ集合を得ることができた．

長応答:
SNS（ソーシャルネットワーキングサービス）が私たちの社会に良い影響を与える理由は、さまざまな面でポジティブな役割を果たしているからです。
以下にその主な理由を挙げます。
1. 情報発信・共有の容易さ（略）
2. コミュニケーションの促進（略）
3. 社会問題への意識向上（略）
4. 多様性の推進と文化の共有（略）
5. ビジネスやキャリアの発展（略）
6. 自己表現の場を提供（略）
7. 迅速な支援や災害時の情報提供（略）
8. コミュニティの形成とつながりの強化（略）
結論
SNS は、私たちの日常生活や社会全体において、情報流通、コミュニケーション、多様性の尊重、迅速な支援など、さまざまな面で良い影響をもたらしています。ただし、使い方次第では逆の影響も及ぼす可能性があるため、慎重で責任ある利用が重要です。

短応答:
SNS は、情報の共有、コミュニケーションの促進、教育や意識向上、コミュニティ形成を助け、知識や意見の交換を容易にします。

図 2: GPT-4o による長応答と短応答.

3.4 応答の収集

本研究では，応答長が LLM の応答に含まれる情報の賛否分布にどのような影響を与えるかを調べるため，各トピックのクエリに対して応答長の異なる 2 種類の応答（長応答，短応答）を収集した．ここで，**長応答**とは LLM に応答の長さを指定せずにそのまま出力させた文章のことを指し，**短応答**とはプロンプトにより短い文章を出力するよう提示して出力させた文章のことを指す．LLM のプロンプトは，長応答では “あなたは優秀な日本語 AI アシスタントです。丁寧に回答してください。”，短応答では “あなたは優秀な日本語 AI アシスタントです。短く答えてください。” とすることで，応答文の長さが異なる LLM の応答を収集した．また，応答長による賛否分布への影響がモデルによって違いがあるかを調べるため，2 つの LLM で応答を収集した．LLM は対話型検索でよく利用される GPT-4o と Google の Gemini 2.5 Flash[2] の 2 つの LLM を利用した．クエリ “SNS が私たちの社会に良い影響を与える理由は何ですか？” に対し GPT-4o で実際に得られた長応答と短応答を図 2 に示す．

1：https://platform.openai.com/docs/models/gpt-4o?snapshot=gpt-4o-2024-11-20

2：https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash?hl=ja

表 3: “SNS は、情報の共有、コミュニケーションの促進、教育や意識向上、コミュニティ形成を助け、知識や意見の交換を容易にします。” という文章のナゲット分割結果.

ナゲット ID	抽出ナゲット
1-1	SNS は情報の共有を助ける
1-2	SNS はコミュニケーションの促進を助ける
1-3	SNS は教育を助ける
1-4	SNS は意識向上を助ける
1-5	SNS はコミュニティ形成を助ける
1-6	SNS は知識の交換を容易にする
1-7	SNS は意見の交換を容易にする

3.5 ナゲット抽出

本研究では応答に含まれる情報の賛否分布を量的に分析するために，収集した応答をナゲット単位に分割して分析を行った．ナゲットとは，意味のある情報の最小単位である [13]．LLM の応答をこのような単位に分割することで，定量的に賛否分布を測定できると考えた．ナゲット分割の例として，クエリ “SNS が私たちの社会に良い影響を与える理由は何ですか？” に対して GPT-4o での短応答で出力された，“SNS は、情報の共有、コミュニケーションの促進、教育や意識向上、コミュニティ形成を助け、知識や意見の交換を容易にします。” という文章のナゲット分割結果を表 3 に示す．

ナゲット分割方法について，人手で全ての応答をナゲット単位に分割することは困難であるため，LLM を用いてナゲット分割を行った．ナゲット分割を行う LLM は，性能とコストのバランスを考慮して GPT-4o を用いた．以下にそのプロンプトを示す．

以下で指定する文章から「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。また、必ず 1 行に 1 ナゲットずつ出力してください．抽出できるナゲットがない場合は何も出力しないでください。

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」「ありがとうございます」）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．

{**トピックごとの分割例**}

ナゲットを抽出する文章：
{**応答文**}

実際に使用したプロンプトには，出力形式を指定する指示とナゲットの定義の説明も含めている．プロンプト内の “**応答文**” には，各クエリに対して LLM が出力した応答文が入力される．また，“**トピックごとの分割例**” は，トピックごとにナゲットの分割例を作成したものが入力される．LLM によるナゲット分割の精度向上のため，トピックごとに例文を用意し，その例文からのナゲット分割例を人手で作成して few-shot に利用した．

ナゲット分割のプロンプトは，音声のみでシステムとユーザが検索を行う音声対話型検索におけるシステムの応答の長さとユーザの検索行動の関係についてナゲットレベルで調査した飛岡らの研究を参考にした [14] [15]．しかし，このプロンプトのみでは 2 つのナゲットに分割されるべきものが 1 つのナゲットとして抽出される誤りが多かった．具体的には，“SNS は知識の普及や意識啓発を促進する” のような，“SNS は知識の普及を促進する” と “SNS は意識啓発を促進する” の 2 つに分割されるべきナゲットが 1 つのナゲットとされる誤りである．飛岡らはこのような誤りを全て人手で再分割していた [14] [15] が，本研究ではコスト削減の観点から LLM でナゲット分割の修正を行った．以下にそのプロンプトを示す．

次の文章は、ある文章をナゲット単位に分割したものです。:
{**元のナゲット**}
この文章のナゲット分割には、次のような誤りが含まれている可能性がありますので指示のように修正してください。間違いがない場合は何も出力しないでください。
・「○○や××は～である」このような記述は「○○は～である」と「××は～である」に分けてください。
・「～は○○や××である」このような記述は「～は○○である」と「～は××である」に分けてください。
3 つ以上のナゲットへの分割が必要な場合もあります。

実際に使用したプロンプトには，出力形式を指定する指示とナゲットの定義の説明も含めている．プロンプト内の “**元のナゲット**” には 1 つ目のプロンプトにて分割された，誤りを含む抽出ナゲットが入力される．このプロンプトで，“SNS は知識の普及や意識啓発を促進する” のような誤ったナゲットを，“SNS は知識の普及を促進する” と “SNS は意識啓発を促進する” の 2 つのナゲットに再分割することができた．以上の 2 つのプロンプトを用いて，全ての応答データを LLM によってナゲット単位に分割した．

3.6 ナゲットに対する賛否ラベルの付与

賛否分布を分析するため，抽出した各ナゲットがトピックに対して賛成，中立，反対のどのスタンスであるかのラベルを付与した．コスト削減の観点から，本研究では賛否ラベルの付与に対しても LLM を用いた．賛否ラベルの付与方法について，ナゲットがトピックに対し賛成のスタンスである情報の場合は “pros”，中立的である場合は “neut”，反対である情報の場合は “cons” のラベルを付与した．例として，“SNS は知識の交換を容易にする” というナゲットは “SNS は私たちの社会にとって良いものである” という主張をサポートする情報であるため，トピック “SNS は私たちの社会にとって良いものか” と

表 4: LLM によるナゲットに対する賛否ラベルと正解データの混同行列.

	正解データ		
LLM による判定	賛成	中立	反対
賛成	531	180	4
中立	48	653	44
反対	1	46	250

表 5: LLM によるナゲットに対する賛否ラベルのスタンス別適合率，再現率，F_1 値，正解率.

スタンス	正解率	適合率	再現率	F_1 値
賛成	0.867	0.743	0.916	0.820
中立	0.819	0.877	0.743	0.804
反対	0.946	0.842	0.839	0.840
マクロ平均	0.877	0.820	0.832	0.822

いうトピックに賛成するナゲットである．そのため，この場合は “pros” のラベルが付与される．以下にナゲットへの賛否ラベル付与に用いたプロンプトを示す.

「{**ナゲット**}」
この文章は、「{**トピック**}」という話題についての文章をナゲット単位に分割したものです。このナゲットがトピックに対して賛成、反対のどちらの立場の主張をサポートする情報か、あるいはどちらの立場もサポートしない中立的な立場の情報なのかを判定してください。

実際に使用したプロンプトには，出力形式を指定する指示とナゲットの定義の説明も含めている．プロンプト内の**ナゲット**は賛否ラベルを付与するナゲットが，**トピック**にはナゲットに対応するトピックが入力される．ナゲットの定義の説明，ナゲット，ナゲットに対応するトピックを提示し，ナゲットがトピックに対してどの立場をとるかを判定させた.

LLM による賛否ラベル付与の妥当性については，著者を含まない 3 名の評価者がプロンプトと同様の基準においてアノテーションを行った．対象データとして，各トピックのクエリから賛成，中立，反対の各スタンスのクエリを無作為に 1 つずつ抽出した．これらのクエリに対する GPT-4o，Gemini 2.5 Flash の長応答と短応答に含まれていたナゲットに対し，対応するトピックに対してナゲットが賛成，中立，反対のどのスタンスを支持する情報かという基準でアノテーションを行った．Fleiss の κ は 0.612（95% CI: [0.587, 0.637]）であり，かなりの一致（Substantial agreement）であった [16]．また，3 名の評価者による評価を多数決した結果を正解データとして，LLM によって付与した賛否ラベルの正解率を算出した．正解データに対する LLM のラベルの正解率は 0.816 であった．LLM による賛否ラベルと正解データの混同行列を表 4 に示す．また，LLM により付与した賛否ラベルのスタンス別適合率，再現率，F_1 値，正解率を表 5 に示す．各ラベルに対する評価指標の数値は，賛成ラベルに対する適合率と中立ラベルに対する再現率を除いて 8 割を超えており，おおむね高い数値を示している.

3.7 分布の算出

3.6 節でナゲットに付与した賛否ラベルを利用し，各トピック，LLM，応答長での応答に含まれる情報の賛否分布をナゲットレベルで算出した．スタンスを $s \in \{$ 賛成, 中立, 反対 $\}$ とし，応答に含まれる情報の賛否分布 P は，応答中に含まれるナゲット数の合計 N とスタンス別のナゲット数 N^s を用いて以下のように算出する.

$$P = \left[p^{\text{賛成}}, p^{\text{中立}}, p^{\text{反対}}\right] = \left[\frac{N^{\text{賛成}}}{N}, \frac{N^{\text{中立}}}{N}, \frac{N^{\text{反対}}}{N}\right] \quad (1)$$

例として，$p^{\text{賛成}}$ は，1 つのスタンスのクエリ集合に対する各応答に含まれるナゲットのうち賛成ナゲットの占める割合を表す．これは賛成ナゲット数を合計ナゲット数で割って算出する．このような計算を 3 つのスタンスのクエリ集合に対して行い，各トピック，LLM，応答長，クエリのスタンスごとにナゲットレベルの賛否分布を算出した.

4 実験結果

本節では，リサーチクエスチョンに基づいて収集したデータとその分析の結果を報告する．最初に，データの概要を把握するために基礎統計量を整理し，全体的な傾向を確認する．その後，詳細な分析結果を提示し，結果から読み取れる傾向を整理する.

4.1 基礎統計量

プロンプトによって応答長の制御が適切に行われたか確認するため，LLM ごとに長応答と短応答の平均応答長を確認した．応答長の平均値は，GPT-4o の長応答で 1,222.21 文字，短応答で 93.81 文字，Gemini-2.5-Flash の長応答で 1,648.71 文字，短応答で 92.59 文字であった．長応答の文字数が短応答の文字数より長いことから，応答長の制御が適切に行われていたことが確認できた.

抽出したナゲット数について大まかな傾向をつかむため，LLM，応答長ごとの平均抽出ナゲット数を表 6 に示す．いずれのモデルにおいても，反対クエリに対する長応答に対して平均抽出ナゲット数が最も多いのは中立ナゲットであるのに対し，短応答では反対ナゲットが最も多くなっている．反対クエリに対する GPT-4o の長応答において，最も平均抽出数が多いのは中立ナゲットで 246.6 であるのに対し，GPT-4o の短応答で最も平均抽出数が多いのは反対ナゲットの 26.2 となっている．このことから，反対クエリにおける短応答では賛否分布が偏る傾向が読み取れる．また，いずれのモデル，応答長でもクエリのスタンスが中立の場合は他のスタンスのクエリに比べて合計抽出ナゲット数が多いことが確認できる．これは，クエリのスタンスが中立の場合，応答は賛成と反対の両方の情報を含

表 6: 各モデル・応答長における平均抽出ナゲット数.

	ナゲットのスタンス			
クエリのスタンス	賛成	中立	反対	合計
GPT-4o（長応答）				
賛成	281.4	193.0	9.0	483.4
中立	172.6	252.8	111.8	537.2
反対	75.8	246.6	143.0	465.4
GPT-4o（短応答）				
賛成	41.0	14.2	1.2	56.4
中立	29.0	28.8	21.0	78.8
反対	9.4	20.8	26.2	56.4
Gemini 2.5 Flash（長応答）				
賛成	286.8	206.2	11.2	504.2
中立	238.0	334.8	113.2	686.0
反対	100.0	264.6	147.8	512.4
Gemini 2.5 Flash（短応答）				
賛成	29.8	17.0	0.0	46.8
中立	40.6	29.8	23.4	93.8
反対	4.0	18.4	30.4	52.8

むことになり，結果として応答に含まれる情報が多くなったためだと考えられる．また，賛成クエリに対する応答に含まれる反対ナゲット数に比べ，反対クエリに対する応答に含まれる賛成ナゲット数が多いことが確認できる．GPT-4o の長応答において，賛成クエリに対する反対ナゲットの平均抽出数は 9.0 であるのに対し，反対クエリに対する賛成ナゲット数は 75.8 となっている．このことから，賛成クエリに対する反対ナゲットと反対クエリに対する賛成ナゲットでは，反対クエリに対する賛成ナゲットの方が出力されやすい傾向が読み取れる．

4.2 節で，トピックごとに賛否分布の特徴について述べる．

4.2 結　　果

GPT-4o の応答の賛否分布を表 7 に，Gemini 2.5 Flash の応答の賛否分布を表 8 に示す．また，長応答と短応答の分布に差があるかを有意水準 5%でカイ二乗検定した結果を表 9 に示す．

まず表 7 の GPT-4o の結果を見ると，トピック “肥満は病気か” に対する中立クエリへの応答を除き，短応答では中立ナゲットの割合が長応答よりも低下していることが分かる．例えばトピック “ボトル入り飲料水は禁止されるべきか” に対する反対クエリへの応答は，長応答では中立ナゲットが 55.9%と半数以上であるのに対し，短応答では 26.8%まで低下している．一方で，短応答ではクエリと同じスタンスのナゲットの割合が増加する傾向が見られる．特に反対クエリでは，短応答における反対ナゲットの割合が長応答に比べ高くなっており，トピック “SNS は私たちの社会にとって良いものか” では長応答で反対クエリに対する反対ナゲットの割合が 70.2%であるのに対し，短応答では 91.3%を占めている．この傾向は他のトピックでも概ね共通している．トピック “動物園は存続するべきか” の反対クエリに対する応答では，長応答で反対ナゲットが 35.1%であるのに対し，短応答では 54.8%まで増加している．また，トピック “ボトル入り飲料水は禁止されるべきか” の反対クエリに対する応答においても，短応答では反対ナゲットの割合が 39.0%となり，長応答（15.9%）と比較して高くなっている．

次に表 8 の Gemini 2.5 Flash の結果を見ると，GPT-4o と同様に短応答では中立ナゲットの割合が低下し，クエリと同一のスタンスのナゲットの割合が高くなる傾向が確認できる．トピック “SNS は私たちの社会にとって良いものか” では，賛成クエリに対する短応答の賛成ナゲットの割合が 98.2%と，長応答（88.8%）よりも高くなっている．また，反対クエリにおいて反対ナゲットの割合が短応答で高くなるという傾向も同様に見られ，トピック “動物園は存続するべきか” では短応答における反対ナゲットの割合が 64.4%となり，長応答での割合（24.2%）を大きく上回っている．

また，他のトピックとは傾向の異なる賛否分布となるトピックがあることが明らかになった．トピック “携帯電話の発する電波は安全か” では，クエリのスタンスにかかわらず中立ナゲットの割合が高いという特徴が見られる．長応答の中立ナゲットの割合がすべてのクエリのスタンスにおいて最も高い割合を占めており，長応答より中立ナゲットの低くなる傾向のある短応答においても，各スタンスのクエリに対して中立ナゲットが半数以上を占めている．また，トピック “肥満は病気か” では，長応答，短応答のいずれにおいても反対クエリに対する応答の反対ナゲットの割合が低く，最も反対ナゲットの割合が多い反対クエリに対する短応答でも 20%前後となった．

長応答と短応答の賛否分布についてカイ二乗検定を行った表 9 を見ると，多くのトピックおよびクエリのスタンスにおいて有意差が確認されている．特に反対クエリに対する応答の賛否分布では，“携帯電話の発する電波は安全か” を除く 4 つのトピックにおいて両モデルで長応答と短応答の賛否分布の間に有意差が認められた．このことから，反対クエリでは応答長により賛否分布の偏りが大きくなるという影響が出ると考えられる．

5 議　　論

分析結果から得られた各リサーチクエスチョンに対する考察をまとめ，対話型検索システムにおける応答長の賛否分布への影響について，システムの使用や設計に対する示唆を述べる．その後，本研究の限界点について論じる．

5.1 分析のまとめと示唆

RQ1：長応答と短応答で LLM 応答に含まれる情報の賛否分布は変化するか

GPT-4o では，トピック “肥満は病気か” の中立クエリに対する応答を除き短応答では中立ナゲットの割合が長応答より少ない結果となっていた．また，Gemini 2.5 Flash でも同様に中立ナゲットの割合が短応答で低下する傾向がみられた．これは

表 7: GPT-4o による応答のナゲットレベル賛否分布．値は，応答長，クエリのスタンス，ナゲットのスタンスごとのナゲット数とその割合（%）を示す．

SNS は私たちの社会にとって良いものか

	ナゲットのスタンス					
	長応答			短応答		
クエリのスタンス	賛成	中立	反対	賛成	中立	反対
賛成	507 (80.5%)	95 (15.1%)	28 (4.4%)	59 (90.8%)	3 (4.6%)	3 (4.6%)
中立	271 (40.3%)	91 (13.5%)	310 (46.1%)	38 (40.9%)	2 (2.2%)	53 (57.0%)
反対	28 (4.9%)	141 (24.9%)	398 (70.2%)	0 (0.0%)	6 (8.7%)	63 (91.3%)

携帯電話の発する電波は安全か

	ナゲットのスタンス					
	長応答			短応答		
クエリのスタンス	賛成	中立	反対	賛成	中立	反対
賛成	81 (20.8%)	302 (77.6%)	6 (1.5%)	21 (39.6%)	30 (56.6%)	2 (3.8%)
中立	38 (8.2%)	389 (83.5%)	39 (8.4%)	9 (13.2%)	48 (70.6%)	11 (16.2%)
反対	26 (6.2%)	364 (86.9%)	29 (6.9%)	4 (7.7%)	42 (80.8%)	6 (11.5%)

肥満は病気か

	ナゲットのスタンス					
	長応答			短応答		
クエリのスタンス	賛成	中立	反対	賛成	中立	反対
賛成	215 (40.2%)	320 (59.8%)	0 (0.0%)	44 (59.5%)	29 (39.2%)	1 (1.4%)
中立	148 (27.3%)	373 (68.7%)	22 (4.1%)	19 (23.2%)	59 (72.0%)	4 (4.9%)
反対	90 (19.9%)	312 (68.9%)	51 (11.3%)	15 (25.9%)	31 (53.4%)	12 (20.7%)

ボトル入り飲料水は禁止されるべきか

	ナゲットのスタンス					
	長応答			短応答		
クエリのスタンス	賛成	中立	反対	賛成	中立	反対
賛成	229 (55.6%)	176 (42.7%)	7 (1.7%)	22 (73.3%)	8 (26.7%)	0 (0.0%)
中立	163 (34.8%)	229 (48.8%)	77 (16.4%)	32 (47.1%)	24 (35.3%)	12 (17.6%)
反対	110 (28.2%)	218 (55.9%)	62 (15.9%)	14 (34.1%)	11 (26.8%)	16 (39.0%)

動物園は存続するべきか

	ナゲットのスタンス					
	長応答			短応答		
クエリのスタンス	賛成	中立	反対	賛成	中立	反対
賛成	375 (83.1%)	72 (16.0%)	4 (0.9%)	59 (98.3%)	1 (1.7%)	0 (0.0%)
中立	243 (45.3%)	182 (34.0%)	111 (20.7%)	47 (56.6%)	11 (13.3%)	25 (30.1%)
反対	125 (25.1%)	198 (39.8%)	175 (35.1%)	14 (22.6%)	14 (22.6%)	34 (54.8%)

表 8: Gemini 2.5 Flash による応答のナゲットレベル賛否分布．値は，応答長，クエリのスタンス，ナゲットのスタンスごとのナゲット数とその割合（%）を示す．

SNS は私たちの社会にとって良いものか

	ナゲットのスタンス					
	長応答			短応答		
クエリのスタンス	賛成	中立	反対	賛成	中立	反対
賛成	525 (88.8%)	43 (7.3%)	23 (3.9%)	55 (98.2%)	1 (1.8%)	0 (0.0%)
中立	332 (44.7%)	116 (15.6%)	295 (39.7%)	54 (48.6%)	11 (9.9%)	46 (41.4%)
反対	30 (4.9%)	140 (22.8%)	443 (72.3%)	1 (1.1%)	10 (11.5%)	76 (87.4%)

携帯電話の発する電波は安全か

	ナゲットのスタンス					
	長応答			短応答		
クエリのスタンス	賛成	中立	反対	賛成	中立	反対
賛成	71 (19.0%)	295 (79.1%)	7 (1.9%)	19 (46.2%)	22 (53.7%)	0 (0.0%)
中立	42 (6.5%)	568 (87.4%)	40 (6.2%)	5 (6.0%)	60 (71.4%)	19 (22.6%)
反対	27 (5.8%)	411 (88.0%)	29 (6.2%)	0 (0.0%)	31 (83.8%)	6 (16.2%)

肥満は病気か

	ナゲットのスタンス					
	長応答			短応答		
クエリのスタンス	賛成	中立	反対	賛成	中立	反対
賛成	230 (36.9%)	392 (62.8%)	2 (0.3%)	19 (31.7%)	41 (68.3%)	0 (0.0%)
中立	233 (37.0%)	367 (58.3%)	29 (4.6%)	39 (49.4%)	32 (40.5%)	8 (10.1%)
反対	127 (28.2%)	261 (57.9%)	63 (14.0%)	4 (9.1%)	28 (63.6%)	12 (27.3%)

ボトル入り飲料水は禁止されるべきか

	ナゲットのスタンス					
	長応答			短応答		
クエリのスタンス	賛成	中立	反対	賛成	中立	反対
賛成	219 (48.2%)	211 (46.5%)	24 (5.3%)	21 (63.6%)	12 (36.4%)	0 (0.0%)
中立	230 (32.3%)	374 (52.6%)	107 (15.0%)	48 (53.9%)	24 (27.0%)	17 (19.1%)
反対	122 (23.9%)	311 (60.9%)	78 (15.3%)	6 (16.2%)	11 (29.7%)	20 (54.1%)

動物園は存続するべきか

	ナゲットのスタンス					
	長応答			短応答		
クエリのスタンス	賛成	中立	反対	賛成	中立	反対
賛成	389 (81.2%)	90 (18.8%)	0 (0.0%)	35 (79.5%)	9 (20.5%)	0 (0.0%)
中立	353 (50.6%)	249 (35.7%)	95 (13.6%)	57 (53.8%)	22 (20.8%)	27 (25.5%)
反対	194 (37.3%)	200 (38.5%)	126 (24.2%)	9 (15.3%)	12 (20.3%)	38 (64.4%)

短応答では応答を短縮するため，事実説明や具体例などの情報が出力されにくいためである可能性がある．また，短応答ではクエリに一致するスタンスのナゲットの割合が高まり，特に反対クエリに対する反対ナゲットの割合に対する影響が大きいことが明らかとなった．加えて，賛成クエリでは反対ナゲットの割合が低いのに対し，反対クエリでは賛成ナゲットは出力されにくいことも明らかとなった．この結果から，対話型検索システム使用時は応答長による賛否分布への影響に留意する必要性が示唆される．

RQ2：LLM のモデルによって応答に含まれる情報の賛否分布は変化するか

本研究で比較した 2 つの LLM では，応答に含まれる情報の賛否分布に大きな差は見られなかった．両モデルにおいて短応答ではクエリと同一のスタンスの情報が強調されやすいという傾向は共通して見られたことから，応答長による賛否分布への影響は特定の LLM のモデルのみではなく多くのモデルで起こりうることが明らかとなった．この結果から，使用する LLM のモデルによらず LLM の応答長による賛否分布への影響に留意する必要性が示唆される．

また，賛否分布にはトピック特有の特徴も見られた．トピック “携帯電話の発する電波は安全か” では，他のトピックに比べ中立ナゲットの割合が高い傾向にある．原因としては，携帯電話の発する電波という科学分野にも絡むトピックであることから，事実説明のナゲットが多かった可能性がある．また，トピック “肥満は病気か” では他のトピックに比べ賛成ナゲットの割合が高いことが表 7，表 8 から読み取れる．これは，肥満は世界保健機構（WHO）によって病気に分類されていることから [17]，ネット上の情報を学習した LLM は「肥満は病気ではない」という主張をサポートするような情報を出力しにくいことが原因だと考えられる．このように，トピックの性質が応答の賛否分布に影響する場合があることが明らかとなった．

5.2 限 界 点

本研究の限界点について述べる．本研究では，人的コスト削減の観点から LLM で作成したクエリを利用した．LLM でのクエリ作成はコストを低く抑えられるという利点があるが，実際に対話型検索を行うユーザが入力するクエリとの乖離が懸念される．そのため，本研究で収集した LLM の応答と実際のユー

表 9: 各トピック，LLM，クエリのスタンスごとの長応答と短応答の賛否分布のカイ二乗検定に対する p 値．*は有意差（$p < 0.05$）が認められたことを示す．

トピック	クエリのスタンス	p 値	
		GPT-4o	**Gemini 2.5 flash**
SNS	賛成	6.91×10^{-2}	8.42×10^{-2}
	中立	4.61×10^{-3}*	2.84×10^{-1}
	反対	$8.46 \cdot 10^{-4}$*	9.43×10^{-3}*
携帯	賛成	3.74×10^{-3}*	2.53×10^{-4}*
	中立	3.32×10^{-2}*	1.16×10^{-6}*
	反対	4.30×10^{-1}	2.80×10^{-2}*
肥満	賛成	1.41×10^{-4}*	6.50×10^{-1}
	中立	7.13×10^{-1}	4.46×10^{-3}*
	反対	4.01×10^{-2}*	5.46×10^{-3}*
ボトル	賛成	1.50×10^{-1}	1.41×10^{-1}
	中立	8.90×10^{-2}	1.76×10^{-5}*
	反対	1.73×10^{-4}*	1.87×10^{-8}*
動物園	賛成	8.45×10^{-3}*	9.4×10^{-1}
	中立	6.32×10^{-4}*	6.28×10^{-4}*
	反対	6.19×10^{-3}*	6.53×10^{-10}*

ザに対する LLM の応答では情報の賛否分布が異なる可能性がある．この点に関しては，今後ユーザ実験を行う際にユーザの入力するクエリを収集することで解決できると考える．

また，ナゲットへの賛否ラベル付与において，そのナゲット自体がトピックに対しどのようなスタンスをとっているかという付与方法を行った．そのため，ナゲットがトピックに対し中立的である場合に “neut” のラベルを付与する想定だった．しかし，実際にはトピックに無関係なナゲットにも “neut” ラベルが付与されてしまっており，賛否分布に影響した可能性がある．この問題を解決するため，賛成，中立，反対の 3 つに加え，トピックに無関係なナゲットであるという判定も加えた，より詳細な賛否分布の分析が必要である．また，長応答では中立的もしくは単なる事実となる情報が増えるために中立ナゲットが応答に多く含まれ，それが長応答と短応答の賛否分布の違いの要因となっている可能性がある．この問題は，トピックに無関係なナゲットが中立ナゲットとして数えられていることが原因でもある．そのため，前述のように詳細な賛否分布の分析を行うとともに，長応答で中立ナゲットが多くなることが応答長による賛否分布に影響していることに留意する必要がある．

6 まとめと今後の課題

本研究では，LLM を用いた対話型検索において，応答長（長応答，短応答）が応答に含まれる情報の賛否分布に与える影響を分析した．賛否が分かれる複数のトピックを対象に賛成，中立，反対のスタンスを想定したクエリを作成し，それらに対する 2 種類の LLM による応答（長応答，短応答）を収集した．収集した応答をナゲットレベルに分割し，ナゲットがトピックに対してどのスタンスであるかを判定してラベルを付与することで，応答に含まれる情報の賛否分布を定量的に分析した．結果，短応答では長応答と比較して中立ナゲットの割合が低下し，クエリと同じスタンスの情報が相対的に多く含まれる傾向が確認された．特に，反対クエリに対する短応答ではこの傾向が顕著であり，応答長が情報の賛否分布に影響を及ぼすことが明らかとなった．また，これらの傾向は本研究で用いた 2 つの LLM において共通して観測され，応答長による影響は特定の LLM のモデルに依存しない可能性が明らかとなった．

以上の結果から，対話型検索システムにおいて応答長は，応答の可読性や簡潔さだけでなく，応答に含まれる情報の偏りにも影響を与える可能性があることが明らかになった．このことから，LLM のモデルによらず短応答では賛否分布が偏ることに留意する必要性が示唆される．本研究は，対話型検索において応答長が応答に含まれる情報の賛否分布に影響を及ぼすかを分析し，LLM のモデルによらず短応答で情報の賛否分布が偏ることを明らかにした．今後はユーザ実験を行うことで，応答長による賛否分布の違いがユーザの意見形成に影響するかを明らかにしたいと考えている．

謝辞 本研究は，JSPS 科研費 JP24K03228，JP25K03229 の助成を受けたものです．ここに記して謝意を表します．

文　献

[1] 総務省. 令和 7 年版情報通信白書. https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r07/pdf/00zentai.pdf. 2025 年 10 月 6 日閲覧.

[2] Masaki Suzuki and Yusuke Yamamoto. Analysis of relationship between confirmation bias and Web search behavior. In *Proceedings of the 22nd International Conference on Information Integration and Web-Based Applications & Services*, pp. 184–191, 2021.

[3] Markus Bink, Sebastian Schwarz, Tim Draws, and David Elsweiler. Investigating the influence of featured snippets on user attitudes. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pp. 211–220, 2023.

[4] Robert Epstein and Ji Li. Can biased search results change people's opinions about anything at all? A close replication of the Search Engine Manipulation Effect (SEME). *PLoS ONE Vol.19, No.3, e0300727*, 2024.

[5] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. This is not what we ordered: Exploring why biased search result rankings affect user attitudes on debated topics. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 295–305, 2021.

[6] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. Generative echo chamber? Effect of LLM–powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2024.

[7] Eugina Leung and Oleg Urminsky. The narrow search effect and how broadening search promotes belief updating. *Proceedings of the National Academy of Sciences*, Vol. 122, No. 13, e2408175122, 2025.

[8] OpenAI. Customizing your ChatGPT personality. https://help.openai.com/en/articles/11899719-customizing-your-chatgpt-personality. 2025 年 10 月 8 日閲覧.

[9] Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. ChatGPT's inconsistent moral advice influences users' judg-

ment. *Nature Scientific Reports*, Vol. 13, No. 4569, 2023.
[10] Shinnosuke Ikeda. Inconsistent advice by ChatGPT influences decision making in various areas. *Nature Scientific Reports*, Vol. 14, No. 15876, 2024.
[11] Shih-Hong Huang, Ya-Fang Lin, Zeyu He, Chieh-Yang Huang, and Ting-Hao Kenneth Huang. How does conversation length impact user's satisfaction? A case study of length-controlled conversations with LLM–powered chatbots. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, No.188*, pp. 1–13, 2024.
[12] Ruiyun Xu, Yue Feng, and Hailiang Chen. ChatGPT vs. Google: A comparative study of search performance and user experience. *arXiv preprint arXiv:2307.01135*, 2023.
[13] 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.
[14] 飛岡憲, 山本岳洋, 大島裕明. 音声対話型検索における応答の長さとタスク複雑性が情報獲得と検索行動に与える影響. 情報処理学会論文誌：データベース, Vol. 19, No. 1, pp. 1–13, 2026.
[15] Ken Tobioka, Takehiro Yamamoto, and Hiroaki Ohshima. Effects of response length on user search experience in spoken conversational search. In *Proceedings of the 27th International Conference on Big Data Analytics and Knowledge Discovery*, pp. 231–236, 2025.
[16] Joseph Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, Vol. 76, pp. 378–382, 1971.
[17] World Health Organization. Obesity and overweight. `https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight`. 2025 年 12 月 30 日閲覧.