引用文脈分析における大規模言語モデル (LLM) の応用可能性

西川 開 [A*], 小柴 等 [A]

Exploring the applicability of Large Language Models to citation context analysis

Kai NISHIKAWA and Hitoshi KOSHIBA

投稿区分: 研究論文

A 文部科学省 科学技術・学術政策研究所

* 文部科学省 科学技術・学術政策研究所 科学技術予測・政策基盤調査研究センター
100-0013 東京都千代田区霞ヶ関 3 丁目 2-2
E-mail: k.nishikawa@nistep.go.jp
corresponding author (代表著者, 問合せ先著者)

Abstruct

Exploring the applicability of Large Language Models to citation context analysis

Keywords: Scientometrics, Citation Context Analysis, Annotation, Large Language Model(LLM), ChatGPT

In contrast to conventional quantitative citation analysis, a method called citation context analysis has been proposed that takes into account the contextual information of individual citations. Although citation context analysis is expected to provide complementary findings to citation analysis, it requires the creation of a large dataset through annotation work, which is costly.

On the other hand, some attempts have been made to have LLM (Large Language Model), which is rapidly becoming popular these days, do the annotation work. However, most of these previous studies were conducted on general texts, and it is not necessarily clear how well they perform when applied to texts with special vocabulary and formatting, such as research papers. This study aims to explore the applicability of LLM to citation context analysis by referring to a publicly available citation context analysis dataset and a manual for the annotation work used to create it. More specifically, we will examine the following issues: 1. Whether LLM can replace humans for annotation tasks in citation context analysis? 2. How can LLM be effectively utilized in citation context analysis?

The results show that LLM annotation performance is comparable to or better than human annotation in terms of consistency, but not in terms of accuracy. However, the accuracy of LLM annotation is not as high as that of human annotation. Therefore, it is not appropriate at this time to have LLM immediately replace human annotators in citation context analysis. However, if it is difficult to prepare a sufficient number of human annotators, it is possible to use LLM as one of the annotators. This study provides the above basic findings that are important for the future development of citation context analysis.

要旨

引用文脈分析における大規模言語モデル (LLM) の応用可能性

キーワード：科学計量学，計量書誌学，引用文脈分析，アノテーション, 大規模言語モデル（LLM），ChatGPT

従来の計量書誌的な引用分析に対して，個々の引用が持つ文脈的な情報を考慮に入れて分析を行う，引用文脈分析と呼ばれる手法が提案されている。引用文脈分析は引用分析に対して相補的な知見をもたらすことが期待されるが，分析のためにはアノテーション作業により大規模なデータセットを作成する必要があり，そのためのコストが大きいことが課題となっている。

一方で，昨今急速に普及しつつある大規模言語モデル (LLM, Large Language Model) にアノテーション作業を代行させようとする試みも見られるようになっている。ただし，こうした先行研究の多くは一般的なテキストを対象とするものであり，論文のような特殊な語彙・フォーマットをもつテキストに適用した場合にどのような性能を発揮するかは必ずしも明らかではない。本研究では，公開されている引用文脈分析のデータセットとその作成に用いたアノテーション作業のためのマニュアルを参照して，LLM の引用文脈分析への応用可能性を探ることを目的とする。より具体的には，1. 引用文脈分析におけるアノテーション作業について LLM は人間を代替できるか 2. 引用文脈分析において LLM をどのように活用することが有効であるかといった点について検討を行う。

本研究の結果から，LLM によるアノテーションのパフォーマンスは一貫性という観点からは人間に匹敵もしくは上回るものの，精度においては高いパフォーマンスを発揮しているとはいえないことがわかった。このため，引用文脈分析に伴う人間によるアノテーション作業をただちに LLM に代行させることは現時点では適切ではない。しかし，人間のアノテーターの人数を確保することが難しい場合，LLM をアノテーターの一人として用いることは可能である。本研究は，引用文脈分析の今後の発展のために重要となる，以上のような基礎的な知見を提供するものである。

目次

図目次

表目次

本文

1　はじめに

論文間の引用関係に着目した計量的な分析は，基本的・暗黙的にすべての引用が等価であるという前提のもとに行われる [ZDM13, TB19]。これに対して，個々の引用が本来持っている引用の位置情報や当該引用の周囲にある文章の意味内容といった文脈的な情報も考慮に入れて分析を行う，引用文脈分析という手法が提案されている。引用文脈分析は従来の計量的な引用分析に対して相補的な知見をもたらすことが期待されるが，分析に必要なデータを作成するコストが多大になるという課題もある。そのため，例えば「複数分野間での引用文脈の傾向の違いを分析する」といった，多くのデータを必要とする研究を実施することは難しい [Cen23, Nis23] 。

引用文脈分析では，引用論文の本文中において任意の被引用論文が言及されている箇所の周囲のテキスト等を用い，個々の引用の文脈的な特徴を判定・付与することで分析のためのデータを作成する。データを作成する方法としては，人間がデータを作成する方法と，機械学習等を用いて自動的にデータを作成する方法の 2 種類が存在する [TB19]。しかし，後者の自動生成についても計算機に教師データを与えて学習させる「教師あり学習」を用いることが多いことから [IHA+21]，いずれの場合でも人間によるデータ作成（アノテーション作業）により大規模なデータを作成することが必要となる。このアノテーション作業に多大なコストを要する点が，引用文脈分析の発展の妨げとなっているといえる。

関連して，昨今の GPT[BMR+20] に代表される大規模言語モデル (LLM, Large Language Model) の発展により，論文に限らない一般的なアノテーション作業について LLM に代行させようとする試みが見られている (e.g., [DGRÁTS22, PWF23, RSA+23, HLG+23, Rei23])。これら先行研究の結果として，LLM はクラウドソーシングにより雇用された人間のアノテーターを上回るパフォーマンスを発揮することもあることが明らかとなっている。また，LLM を用いることにより，人間のアノテーターを雇用する場合よりも時間的・金銭的に安価なコストで多くのデータを作成できることが示されている。ただし，同じテキスト分類作業であっても具体的な作業内容によって LLM によるアノテーションのパフォーマンスは異なることなども報告されており，LLM を用いることでアノテーション作業を直ちに自動化できるかは必ずしも明らかではない。

このように，LLM を用いたアノテーション作業の自動化・支援には期待できるものの，筆者らの知る限り，引用文脈分析ひいては論文を対象とするアノテーションに LLM を適用した研究は行われていない。論文は，独自のフォーマットや文章のスタイルを有し，専門用語を多分に含む，特殊なテキストであると考えられる。したがって，論文に対するアノテーションは先行研究で用いられているような，クラウドソーシングが容易な一般的なアノテーションとは異なる性質を有すると考えられる。実際に，引用文脈分析では研究者や RA(Research Assistant) として雇用された大学院生など，論文を読むことに慣れたものがアノテーションを行うことが多い。また，作業に際しては事前に定義されたスキーマやマニュアルに即してアノテーションを行うという手順を踏むことから，アノテーターをそれらに習熟させるために一定のトレーニングを要することも多くみられる。

これらのことからも，論文へのアノテーションとクラウドソーシングで行えるような種別のアノテーションとは性質が異なることが示唆されている。そのため，先行研究の知見が引用文脈分析ひいては論文を対象とするアノテーションにも適用できるものかは明らかではない。

そこで本研究では，LLM の引用文脈分析への応用可能性を探ることを目的とする。具体的には，引用文脈分析に関する先行研究 [Nis23] において人間が実施したアノテーション作業に基づき，同様の作業を LLM に行わせることで，以下の点について検討を行う。

1. 引用文脈分析におけるアノテーション作業について LLM は人間を代替できるか
2. 引用文脈分析において LLM をどのように活用することが有効であるか

本研究の結果から，LLM によるアノテーションのパフォーマンスは一貫性という観点からは人間に匹敵もしくは上回るものの，精度においては高いパフォーマンスを発揮しているとはいえないことがわかった。このため，引用文脈分析に伴う人間によるアノテーション作業をただちに LLM に代行させることは現時点では適切ではない。しかし，人間のアノテーターの人数を確保することが難しい場合，LLM をアノテーターの一人として用いることは可能である。本研究は，引用文脈分析の今後の発展のために重要となる，以上のような基礎的な知見を提供するものである。

2 関連研究

2.1 引用文脈分析

引用文脈分析は引用内容分析やテキスト内引用分析などと呼ばれることもある。引用文脈分析と引用内容分析は区別されることもあるが [TB19]，本稿では総称として引用文脈分析という言葉を用いる。

引用文脈分析の手順としては，1. カテゴリの設定，2. 引用論文のテキストを用いた各カテゴリの値の判定（アノテーション），3. データの分析，といった形をとる。

各カテゴリの値の判定であるアノテーションはコーディングとも呼ばれ，すでに述べたとおり，人間の手作業によりデータを作成するもの (manual coding) と機械学習等を用いて自動的にデータを作成するもの (automatic data processing) の 2 通りの方法がある。このとき，引用の目的といった意味論的なカテゴリ [ZDM13] については，機械学習を用いる場合でも教師データが必要となることが多いため，人間によるアノテーションが求められる。

また，アノテーション作業一般に該当するが，作業が時間的に高コストな点が課題となる。さらに，引用文脈分析では各カテゴリの値の分布は偏りが大きいことから，一般的なアノテーションに比べてより大規模なデータが必要となる。結果として，分析対象となる分野・ジャーナルが限定的とならざるを得ず，複数分野を比較するような研究の数は文献 [Nis23] など限られている。さらに，引用文脈分析用の公開されているアノテーション結果データセットも少ないながら存在するが，それらデータセットも特定の分野の論文のみを対象としたものに限られており，一般に分析が難しい・分析コストが高い状況にある。

2.2 LLM によるアノテーション作業

LLM のアノテーション作業への活用については，主に社会科学分野において複数，人間のアノテーターとの比較を行う研究が行われている (e.g., [DGRÁTS22, PWF23, RSA+23, HLG+23, Rei23])。

ここで対象となるテキストは SNS への投稿など一般的なものが主流であり，タスクとしては，SNS の投稿に対するヘイトスピーチか否かの分類，新聞記事に対するジャンルの分類，等が挙げられる。テキスト，タスクの種別からして作業の難易度は高くなく，読解に専門知識がいるようなものは余り見られない。こうした特徴を受け，クラウドソーシングのアノテーターが雇用されることが多い。

タスクの評価指標としては主に一貫性・精度・コストが使われる。複数のレポートで共通して述べられている知見としては，LLM によるアノテーションが一貫性・正答率・コストともに（クラウドソーシングの）人間を上回るという点，とり得る値の数が増える等により LLM の性能が大きく変わるという点，等が挙げられる。

このように，アノテーション作業一般について LLM を活用した事例の報告は様々存在するものの，論文に対する LLM の適用例は少なく，引用文脈分析についてはみあたらない。

3 方法

3.1 タスクとデータ

引用文脈分析では多数のカテゴリが使用されており，カテゴリによってアノテーション作業の内容は大きく異なる。本研究では，先行研究 [Nis23] において整理されるカテゴリのうち，引用目的と引用感情の 2 カテゴリに関するアノテーション作業を LLM に実行させる。データセット [NM23] における「引用目的」と「引用感情」の定義は表 1 の通りである。

表 1　引用目的と引用感情の定義

種別	説明	値
引用目的	当該の被引用論文を引用する目的	背景・比較・批判・エビデンス・利用の 5 値をとる
引用感情	著者が当該の被引用論文を引用する際の心的態度	好意的，否定的，中立的の 3 値をとる

文献 [Nis23] では引用文脈に引用箇所をはじめ全 6 種のカテゴリが定義されているが，本稿において，そのうちの引用目的と引用感情にのみ焦点を当てるのは次の理由による。

まず，引用目的と引用感情は引用文脈分析を行った多くの先行研究で扱われる主要なカテゴリである。また，値を判定する上で対象となる被引用論文が言及される周囲のテキストの意味内容を理

解する必要があることから自動化の難易度が高いといえる。他方，その他のカテゴリについては文献 [ZDM13] において構文的 (syntactic) と呼ばれるカテゴリであり，テキストの意味を理解することなく値を判定することが可能であるため，自動化も相対的に容易である。したがって，引用目的と引用感情は自動化によるメリットが特に大きいと考えられ，本研究ではこの 2 カテゴリのみを取り上げることにした。

文献 [Nis23] では 1,174 件のデータを作成しているが，本研究ではこれらのうち，引用論文の本文を JATS(Journal Article Tag Suite)-XML 形式で取得できる 181 件を対象にして，LLM によるアノテーションを実施した。まず，JATS-XML 形式で収集できる論文のみを対象とした理由としては，1. テキスト抽出に係わる手間・エラーがほぼ無いと期待できること，2. 上記より，仮に LLM によるアノテーションが上手くいった場合に大規模に適用しやすいこと，が挙げられる。逆に，対象を PDF でのみ公開されている論文に絞った場合は，PDF から本文テキストや引用情報を正しく抽出する必要がある。この際，PDF は組み版用のデータフォーマットであるので，ページをまたいでいたり多段組になっていたりするものは，機械的にデータを抽出することが困難なことも多い。結果，単純にテキストを自動抽出しただけでは利用できないケースも多くなり，このデータ抽出精度が不十分なことにより正確な分類が行えない可能性が高まる。したがって，LLM の推定精度の問題と，ベースとなるデータの抽出精度の問題の掛け合わせとなり，評価が困難となる。手作業でテキストを抽出すればこの問題を回避できるが作業コストが高い。こうした背景から，対象を JATS-XML 形式で収集できる論文のみに絞った。

次に，アノテーションそのものについて，本研究が文献 [Nis23] をベースとしていることについては，A. 文献 [Nis23] は過去の引用文脈分析に関する先行研究をレビューしたうえでカテゴリとその値をシンプルに整理していること，B. アノテーション作業に用いたマニュアルが公開されており同様の条件により LLM にアノテーションを行わせられること，C. 最終的に分析に使用されたデータ（正解データ）が公開されているため LLM によるアノテーション結果の推定精度を評価できること，の 3 点による。

なお C. について，先行研究 [Nis23] では，最終的に分析に使用したデータを以下の手順により作成している。

1. マニュアルに基づいて 2 名のアノテーター（研究者と Research Asisstant(RA) として雇用された大学院生）が独立して全データについてアノテーションを行う。
2. 結果が一致していなかったデータについては相互に判定理由を（相手を説得しないように留意しながら）述べたのち各人が任意で値の修正を行う。[1)]
3. 最終的に各カテゴリごとに最も値の修正数が少なかったアノテーターのデータを分析に用いる。

[1)] この作業は「ディスカッション」[Lin18] と呼ばれる

3.2 LLM の種類

LLM を用いたアノテーションは，OpenAI 社の提供する OpenAI API ChatCompletion を通じて行った。LLM 自体のモデルは `gpt3.5-turbo-0310` である。また，返答の「ブレ」に係わる指標である `temperature` は 0.5 とした[2)]。

3.3 プロンプト

LLM では，自然言語による指示（プロンプト）を与えることで任意の作業を実行させることができる。他方で，このプロンプトの表現の違いにより作業結果も異なり得る。例えば，ある論文の要約を作成するにしても「以下に与える論文を要約してください」「論文を 200 字程度で要約してください」「論文の概要を手短にまとめてください」「論文の目的と結果を中心に 200 字程度にまとめてください」など，複数のパターンがあり，それぞれで結果が異なりうる[3)]。そこで本研究では，先行研究 [Nis23] において人間のアノテーターがアノテーション作業を行う際に用いたマニュアルとほぼ同内容のプロンプトを基本としつつ，複数のパターンのプロンプトを設定した。具体的なプロンプトは引用目的と引用感情で異なるため，以下ではそれぞれにおけるプロンプトのパターンについて説明する。なお，実際のプロンプトについては付録に収録した。

■ 引用目的におけるプロンプトのパターン

引用目的における先行研究 [Nis23] のマニュアルは，1. とり得る値の種類，2. 各値の定義，3. アノテーションの手順，4. 値を判定する際の根拠となるキーワードと例文，の 4 つの要素からなる。

これに対応する形で，引用目的におけるプロンプトも，基本的な指示に加えて，1. 値の種類のみ（Simple），2. 値の種類・値の定義（Basic），3. 値の種類・値の定義・アノテーションの手順（Precise），4. 値の種類・値の定義・アノテーションの手順・キーワードと例文（Full），のそれぞれを含んだ 4 パターンを設定した。4 つのプロンプトのうち，マニュアルとほぼ同内容であるのは，すべての要素を含む 4 つ目のパターン（Full）である。ただし，オリジナルのマニュアルには，作業対象のファイルの扱いなど，アノテーションの判断に直接作用しない指示も含まれている。本研究ではこうした指示は除外してプロンプトを設定している。また，オリジナルのマニュアルではアノテーションに際して「当該の引用が行われているセクションのタイトル」も加味するように指示されているが，今回の実験ではタイトル抽出の手間を削減するため，その部分を除外して分析を行っている。

この他，公開されているマニュアルや作業対象となる論文は英語のものであるが，マニュアルのオリジナルは日本語で書かれている。このことから，本研究では上記 4 パターンのプロンプトをそれぞれ日本語と英語で表記することで，合計 8 パターンのプロンプトを設定した。図 1 に例とし

[2)] 公式の説明 (`https://platform.openai.com/docs/api-reference/chat/create`, Last access:2023/May/06) では，0 から 2 の範囲の値を取り，デフォルトは 1。0.2 程度だと常にほぼ同じ返答，0.8 程度だとランダムな結果を返すとされる。それらを考慮して今回は中間的な値を設定した。

[3)] この他，LLM 側のパラメータやデータの性質によっては，全く同じプロンプトであっても結果が異なることがある。

Prompt (Purpose, Simple, English.):

```
 Classify the type of purpose for which the author of the following text
is citing the target.
 The type of citation purpose category is Background, Comparison,
Criticize, Evidence, or Use.
 The label of the target literature is Target:  onwards, up to a new
line.  The thesis is the text after Text:.
 Just report the classification result.
Target: {分析対象となる引用文献のラベル}
Text:
{分析対象となる文章 (Paragraph)}
```

図 1　Prompt の例

て，値の種類のみを含む最もシンプルなパターンのプロンプト（Simple, ENSimple, EN）を示す。

■ 引用感情におけるプロンプトのパターン

引用感情のオリジナルのマニュアルに含まれる要素も，基本的には引用目的の場合と同様である。ただし，引用感情の場合はアノテーションの手順に関する指示が少ない。そこで，引用感情のプロンプトでは，基本的な指示に加えて，1. 値の種類のみ（Simple），2. 値の種類・値の定義（Basic），3. 値の種類・値の定義・キーワードと例文（Full），のそれぞれを含んだ 3 パターンについて，それぞれ日本語と英語で表記した合計 6 パターンを設定した。以上のうちすべての要素を含むパターン（Full）は概してオリジナルのマニュアルと同内容であるが，引用目的の場合と同じく，オリジナルのマニュアルに含まれるアノテーションの判断に直接作用しない指示については除外した。

3.4　評価指標

LLM によるアノテーション作業の性能を評価した先行研究では，1. 作業結果の一貫性/信頼性 (Consistency/Reliability)，2. 作業結果の正答率，の 2 つの観点のいずれかにより評価が行われることが多い。本研究においてもこの 2 つの観点により，LLM による引用文脈分析のデータ作成作業を評価する。後者については，[NM23] を正解データとして用いる。

また，いずれの観点においても複数の指標をとり得る。本研究では，前者についてはコーダー間の単純一致率と Cohen’s Kappa，後者については図 2 に示す様々な指標（正解率 (Accuracy)，特異性 (Specificity)，精度 (Precision)，再現率 (Recall)）を適時用いる。

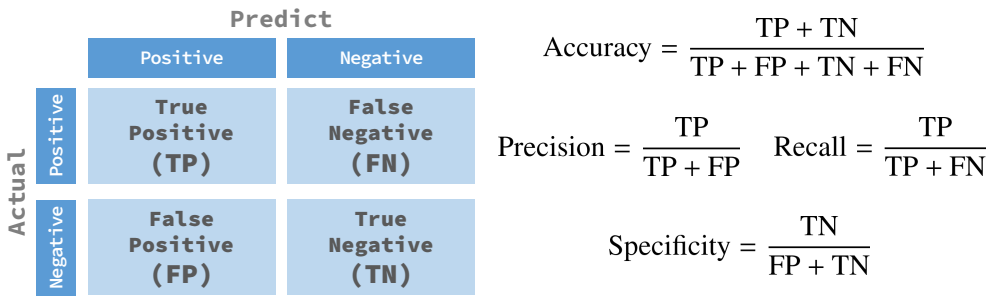

図 2　正答率に関する考え方

4　実験

4.1　データの分布

まず，アノテーションデータの分布を表 2 から表 5 に示す。

表 2　引用分野の分布

cite_pattern	Cnt	Pct
NS-NS	32	17.7%
NS-SSH	71	39.2%
SSH-NS	41	22.7%
SSH-SSH	37	20.4%
Sum	**181**	**100.0%**

表 3　SDGs の分布

SDGs	Cnt	Pct
SDG7	130	71.8%
SDG13	51	28.2%
Sum	**181**	**100.0%**

表 4　引用目的の分布

Purpose	Cnt	Pct
Background	128	70.7%
Comparison	2	1.1%
Criticize	12	6.6%
Evidence	30	16.6%
Use	9	5.0%
Sum	**181**	**100.0%**

ベースとなる文献 [Nis23] では，自然科学 (NS) と人文社会 (SSH) の間の引用傾向を調査するため，SDGs の 7 と 13 という異なる 2 つのテーマを対象に分析している。

表 5　引用感情の分布

Sentiment	Cnt	Pct
Positive	32	17.7%
Neutral	121	66.9%
Negative	28	15.5%
Sum	**181**	**100.0%**

表 2 をみると，今回の分析用に抽出した 181 件のデータ中，NS と SSH の関係については NS-SSH（自然科学が人文社会を引用するパターン）が多いものの，相対的には各カテゴリの比率はそろっておりバランスが良い。

表 3 をみると，テーマについては SDG7 が 7 割と偏りがある。

引用目的及び感情については，個別分析の項で触れる。

4.2　一貫性

作業結果の一貫性という観点から，まず，人間と LLM それぞれの作業結果を比較する。LLM には，先行研究 [Nis23] において使用されたマニュアルとほぼ同内容のプロンプト（Full, EN）を与え，181 件の全データを対象として，引用目的と引用感情それぞれについて 2 回アノテーションを行わせた。表 6 は，先行研究 [Nis23] において 2 名のアノテーターが互いに独立に作業した時点（「ディスカッション」前）の結果と LLM による 2 回の作業結果とについて，それぞれ一致率と Cohen's Kappa を示している。引用目的と引用感情いずれにおいても，LLM の方が一貫性が高いといえる。

表 6　先行研究（人間）と LLM の一貫性比較

	引用目的		引用感情	
	先行研究*	GPT	先行研究*	GPT
単純一致率（%）	71.8%	90.1%	75.7%	91.2%
Cohen's Kappa	0.29	0.77	0.45	0.72

* [Nis23]

次に，方法セクションで述べた全種類のパターンのプロンプトによる LLM の作業結果の一貫性の比較を行う。引用目的については全 8 パターン，引用感情については全 6 パターンのプロンプトを用いて，作業対象となる全 181 件のデータについて LLM に 2 回ずつアノテーションを行わせた。表 7 では，引用目的について，各プロンプトにおける結果が一致しなかった件数，一致率，Cohen's Kappa を示している。

表 7 をみると，引用目的については英語版の値の種類・値の定義・アノテーションの手順を与えたプロンプト（Precise, EN）において最も一貫性が高く，1 回目と 2 回目で回答が異なるものは 8 件 (4.4%)，一貫性を保つものが 95.6% であった。最も一貫性が低いのは，英語のシンプルなプロンプト (Simple, EN) で異なるものが 30 件 (33.1%)，一貫性は 66.9% であった。同パターン以外の

表 7　引用目的のプロンプト別一貫性

Level of Order	LN	Cnt	Agreement	Kappa
Simple	(JP)	37	79.6	0.58
Simple	(EN)	60	66.9	0.40
Basic	(JP)	20	89.0	0.74
Basic	(EN)	17	90.6	0.80
Precise	(JP)	21	88.4	0.71
Precise	(EN)	8	95.6	0.86
Full	(JP)	17	90.6	0.81
Full	(EN)	18	90.1	0.77

(N=181)

プロンプトは，人間のコーダーによる（ディスカッション前の）一致率を上回っている。人間用のマニュアルと同様である，値の種類・値の定義・アノテーションの手順・キーワードと例文のすべての要素を含むプロンプト（Full）ではかえって一貫性が低下しており，必ずしも詳細かつ具体に指示を与えるほど精度が上がるわけではない点は興味深い。

表 8 は，引用感情における各プロンプトの一貫性をまとめている。

表 8　引用感情のプロンプト別一貫性

Level of Order	LN	Cnt	Agreement	Kappa
Simple	(JP)	1	99.4	0.91
Simple	(EN)	3	98.3	0.79
Precise	(JP)	7	96.1	0.65
Precise	(EN)	9	95.0	0.77
Full	(JP)	7	96.1	0.80
Full	(EN)	16	91.2	0.72

(N=181)

表 8 をみると，今回の試行の範囲では日本語で基本的な指示と値の種類のみを与えたプロンプト (Simple) において最も一貫性が高く，1 回目と 2 回目で回答が異なるものは 1 件 (0.6%)，一貫性を保つものが 99.4% であった。最も一貫性が低いのは，英語の詳細マニュアルに例示を付けたプロンプト（Full）で異なるものが 16 件 (8.8%)，一貫性は 91.2% であった。いずれのパターンも，人間による（ディスカッション前の）一致率を上回っている。英語版だけに着目すると，指示を詳細・具体化するほどに一貫性が低下している。また，シンプルなものの一貫性が高い点は引用目的とは大きく異なるもので興味深い。

4.3　正答率

作業結果の正答率という観点から，まず，人間と同様の指示を与えた場合の LLM による作業結果の正答率をみる。引用目的と引用感情それぞれについて，先行研究 [Nis23] において使用されたマニュアルと同内容のプロンプトを与えた場合の LLM による作業結果の正解率 (Accuracy) を調

べたところ，引用目的は 61.3 %，引用感情は 64.6 %であった。なお，一貫性で論じたように LLM には各パターンのプロンプトに 2 回ずつ作業を行わせているが，ここでは初回の試行結果をとりあげた。

次に，引用目的と引用感情それぞれについて，最も一貫性が高かったパターンのプロンプトによる作業結果の正答率をみる。引用目的の場合，英語版の値の種類・値の定義・アノテーションの手順を与えたプロンプト (Precise, EN) の一貫性が最も高かったため，このパターンによる 1 回目の作業結果と正解データの関係を表 9 にまとめた。

表 9　引用目的：推定結果との一致度

		Predict					
	Purpose	**BKG**	**CMP**	**CRT**	**EVS**	**Use**	**SUM**
Actual	Background (BKG)	**107**	2	3	16	0	**128**
	Comparison (CMP)	1	**0**	0	1	0	**2**
	Criticize (CRT)	11	0	**0**	1	0	**12**
	Evidence (EVS)	22	2	0	**6**	0	**30**
	Use	7	0	0	2	**0**	**9**
	SUM	**148**	**4**	**3**	**26**	**0**	**181**

表 9 をみると，例えば正解が「背景 (BKG: Background)」の場合に LLM が正しく BKG と推定できたもの (True Positive) は 107 件あることが分かる。一方で，「根拠 (EVS:Evidence)」と推定したものの正しくは BKG であるものは 16 件，BKG と推定したものの EVS であったものは 22 件あるなど，正解データとの相違も見られる。また，「比較 (CMP:Comparison)」「批評 (CRT:Criticize)」「利用 (Use)」は元々頻度の少ない値であるとはいえ，1 件も正しく推定できていない。

同様に，引用感情の場合に最も一貫性の高かった英語版の値の種類のみを与えたプロンプト（Simple, EN）による 1 回目の作業結果と正解データの関係を表 10 にまとめた。

表 10　引用感情：推定結果との一致度

		Predict			
	Sentiment	**PG**	**NT**	**NG**	**SUM**
Actual	Positive (PG)	**3**	29	0	**32**
	Neutral (NT)	1	**120**	0	**121**
	Negative (NG)	0	24	**4**	**28**
	SUM	**4**	**173**	**4**	**181**

表 10 をみると，例えば正解が「中立的 (NT: Neutral)」で，正しく NT と推定できたものは 120 件あることが分かる。一方で，NT と推定したものの，「肯定的 (PG: Positive)」であったもの 24 件など相違も見られる。

これまでは一つのプロンプトによる 1 回目の作業結果の正答率についてみてきた。他方で，人間がアノテーションを行った場合には，複数人による作業結果を統合・調整することで正解データを作成することが多い。そこで，引用目的と引用感情それぞれについて設定した各数種類のプロンプ

トによる 2 回ずつの作業結果について多数決を取ることで LLM による結果をまとめ，この結果と正解データとの比較を行った。表 11 は，引用目的について，LLM による作業結果と正解データとの関係をまとめたものである。

表 11　引用目的：推定結果との一致度（多数決版）

		Predict				
	Purpose	**BKG**	**CMP**	**CRT**	**EVS**	**Use**	**SUM**
Actual	Background (BKG)	**103**	1	2	21	1	**128**
	Comparison (CMP)	1	**0**	0	0	1	**2**
	Criticize (CRT)	11	0	**0**	1	0	**12**
	Evidence (EVS)	15	1	0	**13**	1	**30**
	Use	6	0	0	3	**0**	**9**
	SUM	**136**	**2**	**2**	**38**	**3**	**181**

表 11 では EVS の正解がやや増えてはいるものの，表 9 の結果と基本的な傾向は変化無く，表 9 の精度が他のプロンプトを用いた場合と比較して著しく異常である可能性は考えにくいことを示唆している。

同様に，表 12 では，引用感情について多数決により一つに定めた LLM の作業結果と正解データとの関係を示す。

表 12　引用感情：推定結果との一致度（多数決版）

		Predict			
	Sentiment	**PG**	**NT**	**NG**	**SUM**
Actual	Positive (PG)	**3**	29	0	**32**
	Neutral (NT)	1	**119**	1	**121**
	Negative (NG)	0	23	**5**	**28**
	SUM	**4**	**171**	**6**	**181**

表 12 と表 10 の結果と傾向はほぼ変化無く，判定結果が比較的ロバストであることが示唆される。

4.4　実験結果の考察

実験の結果，LLM は一貫性という観点からは人間を上回るものの，正答率，特に Accuracy という観点からは高い品質のデータを作成できているとは言い難いことが伺える。表 9，11 および表 10，12 の結果を見ると，もともとクラスごとに偏りが大きいとはいえ，決してうまく推定できているわけではない。例えば，引用目的について表 9 のうち BKG と推定したもの 148 件中，実際に BKG であったものは 107 件で 72.3%。元の分布のうち BKG の割合が 70.7% であったことを考えると，何も考えずに常に BKG と推定した場合に比べて改善は 1.6% である。BKG と推定したが，実際には BKG ではなかったもの (False Positive)，BKG 以外のクラスを推定したが，実際には BKG だったもの (False Negative)，双方一定数存在している点も留意事項である。

引用感情については，表 10 を見ると PG と判定したものは全て，実際にも PG であって，False Negative が少ないという観点ではよく推定できている。ただし，その数は PG 全数 28 件中の 4 件であって多くはない。NG もほぼ同様である。大勢を占める NT については NT と判定した 173 件中 120 件が実際に NT だが，何も考えずに常に NT と推定した場合に比べて改善は 1.5% である。3 割程度は PG/NG の誤判定であることにも留意が必要である。

PG/NT との判定が出た場合は推定結果を信じるとして，その割合は 8/181 件であり，全数から考えると誤差の範囲である。それ以外のケースでは基本的に推定結果が信頼できないため，人手による見直しは避けられないと考えられる。

他方で，一定水準以上の一貫性を有しながら，かつ正答率は低いという本実験の結果は，LLM と人間の間でテキストの「解釈」の仕方に違いがあることを示しているとも考えられる。そこで，LLM が値を正しく推定できなかったか，若しくは複数回の試行間で回答が一致しなかったケースについて，具体的に判定作業の対象となるテキストを確認することで，LLM によるテキストの解釈の仕方を推測・検討した。

検討の結果，LLM が誤答若しくは一貫性の低い判定を行うテキストでは，対象となる被引用論文とその周囲のセンテンスの関係が明示的に示されていないことが示唆された。例えば下記のテキスト [BPZH19] p.207 [4]では，LLM は「批判 (CRT)」と推定しているのに対して，正解は「背景 (BKG)」である。

> (...) This use of technical devices in an attempt to suppress political debates has been widely documented elsewhere (e.g. Latour, 2004 ; Lupton and Mather, 1997). At an extreme this use of GIS and mapping reinforces and aggravates existing divides and inequalities.

このとき，確かに対象となる被引用論文 [LM97] が含まれているセンテンスやその直後に来るセンテンスには，「批判」の判断の根拠となり得る語 (widely, reinforces and aggravates existing divides and inequalities) が含まれている。しかし，これらのセンテンスで CRT の対象となっているのは当該の被引用論文ではなく，“this use of technical devices/GIS and mapping” である。つまり，ここでは引用論文の研究テーマに関する背景情報を示すために当該の被引用論文に言及していると考えられることから，BKG が正解となっている。これは人間であれば比較的容易に判断できるものであるが，被引用論文と周囲のテキストとの関係が明示的に語として表されているわけではないことから，LLM には判断が困難であったと考えられる。ただし，誤答若しくは回答の不一致例の中には，判定理由を推察することが困難なケースも複数含まれており，上記のパターンによってすべてを説明できるわけではない。

以上より，LLM によるアノテーション作業は正解データとの比較・一致の観点から不満足なものであるため，LLM によるアノテーション作業結果をそのまま分析に使用することには問題があるといえる。つまり，引用文脈分析において現状の LLM を人間のアノテーターの代替として利用することは困難であることが明らかとなった。

[4] SDG13 における NS-SSH の引用ペア（B/f0083）である。

5　LLM の利用局面の検討

5.1　人間のデータ作成の補助

これまでは，人間用のマニュアルをベースに LLM に指示を与えた際のどれぐらいの性能を発揮するかという観点から評価を行ってきた。結果として，その観点では LLM を用いた作業結果は人間に及びそうに無く，未知データについては人間の判定が必要なことが示唆された。

ただし，LLM の活用を考える際，完全にアノテーション作業を代替させるという方向性だけで無く，人間による作業のサポートとして利用するという方向性もあり得る。例えば，表 9 の PG の例のように，「PG 判定のものについてはほぼ確実に PG である」と考えて良ければ，つまり精度が十分に高ければ，その部分については人間が作業する手間を省くことができる。

また，「引用目的」についての判定を考えたとき，例えば BKG と判定されたものの中にも，「確実に BKG である」として BKG と推定したもの，「他のクラスの可能性もあるが強いて言えば BKG である」として BKG と判定したもの，などその確からしさには差がある可能性が高い。この場合，仮に，誤判定されたものが LLM の中でも BKG の確実性が低かったなどの特徴があるとすれば，判定が難しそうなものだけを人間が処理し，それ以外は自動処理した結果を信じる，と言うような使い方が考えられる。

一般にテキスト分類作業はクラス（変数がとり得る値）数が多いほど難易度が高く，LLM によるアノテーションについてもクラス数が増えるにつれて精度が低下することが明らかとなっている [DGRÁTS22]。このことから，今回のタスクにおいてもクラスの数を減らして作業を行わせることで，精度を高められる可能性がある。

こうした観点から，LLM を人間のサポートとして利用するという方向性でプロンプトを見直し，活用の可能性を検討した。

5.2　引用目的

引用感情は「背景 (BKG)」のみで 7 割，「根拠 (EVS)」が 1.6 割程度であるので，この 2 つと，それ以外の 3 クラスを一つにまとめたものとで編成し直し，3 値問題として試行した。

まず，3 クラス目を「その他 (Other)」としてプロンプトを実行した結果を表 13 に示す。これを見ると，「その他」に分類されたものはなく，推定結果は BKG か EVS のどちらかである。また，EVS と推定したケースが増え，推定は全般的に悪化している。「その他」の語に引きずられているのか，曖昧なクラスを避けるように判断を下しているように推察される。なお，2 回試行時の一貫性（一致率）は 91.7% である。

前述の通り「その他」という語に引きずられて判断が悪化している可能性が考えられたことから，この「クラス名」が判定に影響を及ぼしているという仮説に基づいて，3 クラス目を「一般 (GEN: General)」「保留 (PDG: Pending)」と置き換えた試行を行った。結果をそれぞれ，表 14,15 に示す。この場合，3 クラス目に属する結果もいくつか観察できるものの，傾向としては表 13 と同様であ

表 13　引用目的：3 クラス目を「その他」とした場合

		Predict			
	Purpose	**BKG**	**EVS**	**OTH**	**SUM**
Actual	**Background (BKG)**	**65**	63	0	**128**
	Evidence (EVS)	3	**27**	0	**30**
	Other (OTH)	14	9	**0**	**23**
	SUM	**82**	**99**	**0**	**181**

る。なお，2 回試行時の一貫性（一致率）は 86.2%，　91.7% である。

表 14　引用目的：3 クラス目を「一般」とした場合

		Predict			
	Purpose	**BKG**	**EVS**	**GEN**	**SUM**
Actual	**Background (BKG)**	**37**	71	20	**128**
	Evidence (EVS)	0	**30**	0	**30**
	General (GEN)	5	16	**2**	**23**
	SUM	**42**	**117**	**22**	**181**

表 15　引用目的：3 クラス目を「保留」とした場合

		Predict			
	Purpose	**BKG**	**EVS**	**PDG**	**SUM**
Actual	**Background (BKG)**	**9**	113	6	**128**
	Evidence (EVS)	0	**29**	1	**30**
	Pending (PDG)	1	21	**1**	**23**
	SUM	**10**	**163**	**8**	**181**

一般にクラス数の削減により精度や再現率などの指標が向上されることが期待されるが 3 クラスにまとめても精度は向上しなかった[5)]。その原因をクラス名の有する意味によるものと考え，語を入れ替えたが改善の傾向が見られなかった。ただし，1 番目のクラス名 (Background) と 2 番目のクラス名 (Evidence) は固定であるので，このクラス名の意味が影響している可能性を排除できない。そこで，3 番目のクラス名を「不明 (UKN: Unknown)」とし，さらにこれまで「Background」など意味のあるクラス名にしていたものを「BKG」など意味が読み取れない形に隠蔽して，プロンプト上での説明のみに基づいて推定するような形での実行を試行した。結果を表 16 に示す。この場合も傾向としては表 13 と同様である。なお，2 回試行時の一貫性（一致率）は 86.2% である。

ここまでの試行で 3 クラスにまで絞り込んでも精度向上は見込みにくいことが明らかになった。そこで表 17 では方向性を変えて，BKG(PB) か否か (NB)，EVS(PE) か否か (NE) の 2 値判定をそ

5) 一般的にはクラス数を削減することによる教師データの増加も寄与する一方，LLM ではモデル自体は変化しないこともあり，クラス数削減による改善の割合は元々大きくはないとは考えられる。

表 16　引用目的：3 クラス目を「不明」とした場合

		Predict		
	Purpose	**BKG**	**EVS**	**UKN**	**SUM**
Actual	**Background (BKG)**	**107**	21	0	**128**
	Evidence (EVS)	22	**8**	0	**30**
	Other (OTH)	19	4	**0**	**23**
	SUM	**148**	**33**	**0**	**181**

れぞれ行い，この組み合わせと，正解との対応を見た。ここでもクラス名は隠蔽をしている。個別に見れば最も単純な 2 値分類であるので精度の向上が期待されるが，表 17 を見る限り，ここでも正解との特定の傾向は観察できない。なお，2 回試行時の一貫性（一致率）は，BKG が 90.6%，EVS が 91.7% である。

表 17　引用目的：BKG か否か，EVS か否かの 2 値問題にした場合

		Predict				
	Purpose	**NBNE**	**PBPE**	**NBPE**	**PBNE**	**SUM**
Actual	**Background (BKG)**	6	68	9	45	**128**
	Evidence (EVS)	0	17	3	10	**30**
	Other (OTH)	4	9	1	9	**23**
	SUM	**10**	**94**	**13**	**64**	**181**

表 18 は表 17 を念頭に，もっとも数の多い BKG かそれ以外（不明 (UKN: Unknown)）かという単純な 2 値問題にまで縮退して試行した。また，これまで同様クラス名の隠蔽も行った。表 18 を見ると，BKG と推定したもの 43 件のうち 38 件 (88.4%) は実際に BKG であり精度は高い。数は多いとはいえないものの，ある程度の割合に達しているため，このケースで BKG と判定されたものについては，自動判定で代替できる可能性がある。なお，2 回試行時の一貫性（一致率）は 91.2% である。

表 18　引用目的：BKG か否かの 2 値問題にした場合

		Predict		
	Purpose	**BKG**	**UN**	**SUM**
Actual	**Background (BKG)**	38	90	**128**
	Evidence (EVS)	2	28	**30**
	Other (OTH)	3	20	**23**
	SUM	**43**	**138**	**181**

5.3 引用感情

引用感情は目的と違い，もともと 3 値分類であることから，クラスを減らす方向は考えにくい。そこで本件ではクラスを追加する方向で試行した。

前述の通り，判定にも「確実に A である」というものから，「強いて言えば A である」というものまで，一定の幅があることが想定される。このとき，確実性が低いものと高いものを選り分けることができれば，精度等の改善への寄与が期待できる。

そこでまず，確度の低いものを切り分けるために「保留 (PD: Pending)」を加えたケースを表 19 に示す。これを見ると PD と判定されるものもいくつか観察されるが 30 件程度と数は多くなく，また 1 件ではあるが，実際には PG であるものを NG と真逆に解釈するケースも観測されるなど悪化している部分も見受けられる。

表 19　引用感情：保留を追加

		Predict				
	Sentiment	**PG**	**NT**	**NG**	**PD**	**SUM**
Actual	**Positive (PG)**	**14**	17	0	1	**32**
	Neutral (NT)	19	**78**	0	24	**121**
	Negative (NG)	2	16	**5**	5	**28**
	SUM	**35**	**111**	**5**	**30**	**181**

PD: Pending

確度が低いものを別のクラスに選別するという取組では改善がみられず，また，クラス数の増加は現状では好ましくない。そこで表 20 では「中立的 (Neutral)」のクラスに対して，実際に中立であるものと，判定が困難であるものとを両方とも収容することにした。そこで，このクラスのみクラス名を隠蔽することにし，その際人間にとって意味が理解しやすいようクラス名を「不明 (UN: Unknown)」と改めた。結果，PG，NG の数が増えたが誤判定がほとんどで意味をなしていない。

表 20　引用感情：中立的と不明を混合

		Predict			
	Sentiment	**PG**	**UN**	**NG**	**SUM**
Actual	**Positive (PG)**	**15**	8	9	**32**
	Neutral (NT)	19	**57**	45	**121**
	Negative (NG)	1	8	**19**	**28**
	SUM	**35**	**73**	**73**	**181**

UN: Unknown

5.4 引用文脈分析における LLM のユースケースの検討

以上より，LLM によって部分的にアノテーション作業を代行させることも基本的には困難であるといえる。他方で，引用文脈分析における LLM 利用の余地が全くないというわけではないことも，これまでの分析結果は示唆している。

文献 [PWF23] では，アノテーション作業一般における LLM のユースケースを，(1) 人間が作成したデータのクオリティの確認，(2) 人間のレビューの優先順位の決定，(3) 教師あり分類器の微調整と検証のためのデータの作成，(4) コーパス全体の分類，の 4 点に類型化している。このうち，(1) 以外のケースについては，本稿においてこれまで示してきた通り，引用文脈分析に対しては適用が困難といえる。(4) については，実験セクションでみたように，正答率が低いことから LLM の作成したデータをそのまま分析に用いることには問題がある。同様の理由により，(3) の用途で用いることも避けるべきである。また，(2) はアノテーション作業の部分的代行に連なるユースケースであるが，本章でこれまで論じてきたように，この用途での利用にも懸念がある。

他方で，(1) については，検討の余地があるように思われる。このユースケースは，人間と LLM の作業結果を比較することで，人間の作業結果のクオリティを検討することを意味する。これを引用文脈分析の文脈に落とし込むと，複数の人間のアノテーターが作成したデータについて，分析に使用するためのデータを一つに絞り込む際に LLM の作業結果を参考情報として活用するという用途が想定される。実験セクションでみたように，LLM による作業結果の正答率は低い一方で一貫性は人間を上回り安定している。また，LLM には判定理由を出力させることもできることから，LLM を人間とは異なる基準・傾向をもつアノテーターとして捉えることもできる。

こうした性質を踏まえると，LLM の作業結果及び判定理由は，「ディスカッション」[Lin18] と呼ばれる，複数の人間のアノテーターが相互に独立して作成したデータを一つに絞り込む際に，相手を説得しないように留意しつつ相互に判定理由を説明し，必要であれば任意で自身の作業結果の修正を行うというプロセスにおいて，第三者的な立場からの参考情報として活用することができよう。これにより，特定のアノテーターの主観が大きく作用する可能性をより低減できる可能性がある。

また，一般にアノテーション作業で高品質なデータを作成することは困難であり，人間のコーダーを雇用するにはコストもかかることから，時には一人のアノテーターが作成したデータに依拠せざるを得ない場合もあることが指摘されている [RSA+23]。作業対象が論文という特殊なテキストである引用文脈分析の場合，アノテーターに対してより高度な技能が要求されることから，一般のアノテーション作業以上に作業者を確保することが困難であると考えられる。このとき人間が一人で作業することを避ける，もしくは複数人が作成したデータを多数決により一つに絞り込むなどのために，LLM を N 人目のアノテーターとして採用することも考えられる。

6 おわりに

本研究の目的は，LLM の引用文脈分析への応用可能性を探ることであった。研究の結果として，現状の ChatGPT では，引用目的および引用感情という 2 つの作業について人間を代替できるほど高い精度でデータ作成を行うことができないことが明らかとなった。また，特定の値については LLM の回答結果を用いることとして残りの部分を人間が担当するといったように，人間のアノテーション作業を LLM に部分的に代行させることも難しいことがわかった。

他方で，LLM の作業結果は一定の一貫性と判断根拠の説明可能性を有していることから，第三のアノテーターとして捉えることはできる。このことから，引用文脈分析における LLM のユースケースとしては，以下の 2 つが考えられる。第一に，複数の人間のアノテーターによる作業結果を一つに絞り込む際の参考情報として LLM による作業結果を利用することができる。第二に，人間のアノテーターの人数を確保することが難しい場合に，N 人目のアノテーターとして LLM を用いることは可能である。

本研究の貢献は，引用文脈分析に LLM を活用することを試みようとしている研究者に対して，その限界と活用し得る局面を示唆している点にある。また，アノテーション作業一般における LLM の性能の検証やユースケースの提案を行った先行研究に対して，引用文脈分析という特殊なタスクにおけるその応用可能性とユースケースを検討した点において，本研究の知見は新規なものである。

他方で，本研究には次のような限界もある。今回の分析では LLM の中でも `gpt3-turbo-0310` を用いていた。これは実験実施時において，もっとも精度が良いと考えられるモデルであった。その後，本稿執筆途中において，`gpt3-turbo-0310` を上回る性能を持つとされている `gpt4` というモデルの一般提供も開始された。こちらのモデルを使用した分析も試行した結果，むしろ正答率については `gpt3-turbo-0310` よりも悪化することが分かっているが（付録 B 参照），今後新たなモデルが登場することで，本研究の結論を覆すような性能を LLM が発揮するようになることも考え得る。このことから，本研究の知見は現時点における LLM の応用可能性をスナップショット的に探ったものであり，今後の技術動向に合わせて継続的に分析を行っていくことが求められる。

参考文献

[BMR+20] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. arXiv (preprint), 2020.

[BPZH19] Maud Borie, Mark Pelling, Gina Ziervogel, and Keith Hyams. Mapping narratives of urban resilience in the global south. Global Environmental Change, 54:203–213, jan 2019.

[Cen23] Center for S&T Foresight and Indicators. Science map2020, March 2023.

[DGRÁTS22] Rodrigo Dorantes-Gilardi, Aurora A. Ramírez-Álvarez, and Diana Terrazas-Santamaría. The role of highly intercited papers on scientific impact: the mexican case. Applied Network Science, 7(1), aug 2022.

[FY15] Toyofumi Fujiwara and Yasunori Yamamoto. Colil: a database and search service for citation contexts in the life sciences domain. Journal of Biomedical Semantics, 6(1), oct 2015.

[HLG+23] Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. AnnoLLM: Making large language models to be better crowdsourced annotators. arXiv (preprint), 2023.

[IHA+21] Sehrish Iqbal, Saeed-Ul Hassan, Naif Radi Aljohani, Salem Alelyani, Raheel Nawaz, and Lutz Bornmann. A decade of in-text citation analysis based on natural language processing and machine learning techniques: an overview of empirical studies. Scientometrics, 126(8):6551–6599, jun 2021.

[Lin18] Chi-Shiou Lin. An analysis of citation functions in the humanities and social sciences research from the perspective of problematic citation analysis assumptions. Scientometrics, 116(2):797–813, may 2018.

[LM97] M. Lupton and C. Mather. 'the anti-politics machine': GIS and the reconstruction of the johannesburg local state. Political Geography, 16(7):565–580, sep 1997.

[Nis23] Kai Nishikawa. How and why are citations between disciplines made? a citation context analysis focusing on natural sciences and social sciences and humanities. Scientometrics, 128(5):2975–2997, feb 2023.

[NM23] Kai Nishikawa and Mie Monjiyama. Data on a citation context analysis focusing on natural sciences and social sciences and humanities. 2023.

[PWF23] Nicholas Pangakis, Samuel Wolken, and Neil Fasching. Automated annotation with generative ai requires validation. arXiv (preprint), 2023.

[Rei23] Michael V. Reiss. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. arXiv (preprint), 2023.

[RSA+23] Christopher Michael Rytting, Taylor Sorensen, Lisa Argyle, Ethan Busby, Nancy Fulda, Joshua Gubler, and David Wingate. Towards coding social science datasets with language models. arXiv (preprint), 2023.

[TB19] Iman Tahamtan and Lutz Bornmann. What do citation counts measure? an updated review of studies on citations in scientific documents published between 2006 and 2018. Scientometrics, 121(3):1635–1684, sep 2019.

[ZDM13] Guo Zhang, Ying Ding, and Staša Milojević. Citation content analysis (CCA): A framework for syntactic and semantic analysis of citation content. Journal of the American Society for Information Science and Technology, 64(7):1490–1503, may 2013.

付録 A　引用感情に関する別解

本編中，LLM の活用について種々の取組（確度の低いものを隔離することを目的とした保留クラスの導入やクラス名の隠蔽）を行ったが，これら確度の低いものを隔離することで Positive, Negative の感情正答率を向上させるという試行は効果を上げなかった。

そこで大きく方針を変え，GPT を用いた感情推定とは異なる推定手法での推定を試行した。具体的には，AWS(Amazon Web Service) の感情推定機 (Amazon Comprehend) を適用して感情推定を行わせた。ここではまず，LLM を用いて与えた文章全体から対象となる引用文献に関連する箇所の文章のみを抽出し，次に，Amazon Comprehend で感情推定を行うという二段階の作業で分析する。Amazon Comprehend の出力は Positive，Negative，Neutral，Mixed の 4 値である。結果を表 21 に示す。

表 21 を見ると，ほぼ全てが Neutral と判定されており，活用の難しさが伺える。

表 21　引用感情：商用感情推定器を活用

		Predict				
	Sentiment	**PG**	**NT**	**NG**	**MX**	**SUM**
Actual	**Positive (PG)**	**0**	31	0	1	**32**
	Neutral (NT)	0	**118**	3	0	**121**
	Negative (NG)	0	26	**2**	0	**28**
	SUM	**0**	**175**	**5**	**1**	**181**

MX: Mixed

ただし，この作業遂行に際して，アノテーション作業を行う対象となる特定の範囲のテキストを取り出す作業については，LLM が十分な精度を発揮することが示唆された。

引用文脈分析では，カテゴリの値を判定するために，引用論文の中から対象となる特定の被引用論文が言及されている所定の範囲のテキストを抜き出すことが必須のプロセスとなる。しかし，論文によって文中での被引用論文への言及の仕方は異なることから，Colil[FY15] や Microsoft Academic のように整備されたデータベースに収録されている論文を除けば，この抜き出し作業を自動化することはそれ自体が一つのハードルとなり得る。特に，manual data processing の automatic data processing に比べたときの利点の一つは，特定のデータベースやフォーマットに依存せずに分析対象とする論文群を選定できることにあると考えられるが [Nis23]，このような場合には抜き出しの自動化はより困難となり，多くの manual data processing を採用する研究では手作業で抜き出し作業を行なっていると推察される。そのため，LLM によりこの作業を代行するだけでも，アノテーション作業による時間的コストを低減できる可能性がある。

付録 B　GPT4 による試行例

実験開始時点 (2023 年 5 月) では，API を通じて一般ユーザが利用可能なモデルのうち，最も高性能と期待できるものは `gpt-3.5-turbo` であったため，本編ではこれを用いた。

ところで，2023 年 7 月からは GPT4 も利用できるようになった。一般的に GPT4 の方が高性能とされるため，追加でいくつかの事例について調査した。

具体的には，英語版プロンプトについて，また，4 章記載の範囲で検証した。

B.1　引用目的

引用目的について表 22 から表 25 にまとめた。

表 22　引用目的：Simple（英語）

		Predict					
	Purpose	**BKG**	**CMP**	**CRT**	**EVS**	**Use**	**SUM**
Actual	Background (BKG)	**35**	8	4	56	25	**128**
	Comparison (CMP)	0	**2**	0	0	0	**2**
	Criticize (CRT)	5	1	**0**	2	4	**12**
	Evidence (EVS)	4	2	0	**19**	5	**30**
	Use	0	0	0	2	**7**	**9**
	SUM	**44**	**13**	**4**	**79**	**41**	**181**

表 23　引用目的：Basic（英語）

		Predict					
	Purpose	**BKG**	**CMP**	**CRT**	**EVS**	**Use**	**SUM**
Actual	Background (BKG)	**41**	18	4	39	26	**128**
	Comparison (CMP)	0	**2**	0	0	0	**2**
	Criticize (CRT)	4	6	**0**	1	1	**12**
	Evidence (EVS)	5	6	0	**15**	4	**30**
	Use	0	0	0	3	**6**	**9**
	SUM	**50**	**32**	**4**	**58**	**37**	**181**

表 24　引用目的：Precise（英語）

		Predict					
	Purpose	**BKG**	**CMP**	**CRT**	**EVS**	**Use**	**SUM**
Actual	Background (BKG)	**28**	10	6	45	39	**128**
	Comparison (CMP)	0	**1**	0	0	1	**2**
	Criticize (CRT)	4	3	**0**	1	4	**12**
	Evidence (EVS)	3	3	0	**14**	10	**30**
	Use	0	0	0	1	**8**	**9**
	SUM	**35**	**17**	**6**	**61**	**62**	**181**

表 25　引用目的：Precise+ 例示（英語）

		Predict					
	Purpose	BKG	CMP	CRT	EVS	Use	SUM
Actual	Background (BKG)	51	4	3	22	48	128
	Comparison (CMP)	0	1	0	0	1	2
	Criticize (CRT)	6	3	0	0	3	12
	Evidence (EVS)	4	0	0	13	13	30
	Use	0	0	0	0	9	9
	SUM	61	8	3	35	74	181

GPT3 を用いた本編では一致度の高さから英語版の詳細マニュアルを採用し，表 9 にまとめた。上記の追試では表 24 が対応する。

まず，GPT3 では特定のクラスに偏るような出力もあったところ，GPT4 ではその不均衡が減るような分類を行っていそうな様子が見える。しかしながら，今回は正解データが不均衡であるため，かえって正解から遠ざかっている様子が見える。実際に，表 9 と表 24 を比べても，ボリュームゾーンの BKG の推定が減った結果，悪化していると言える。なお，不均衡を是正した結果，CMP，Use の正解がいくつか見られている。特に，Use の推定はゼロであったところ，大幅に数が増えている。

以上より，モデルによっても大きく結果が変わること自体は把握できたが，人手での分類に頼らざるを得ないという本編の結果を覆すようなものではなかった。

B.2　引用感情

引用目的について表 26 から表 28 にまとめた。

表 26　引用目的：Simple（英語）

		Predict			
	Sentiment	PG	NT	NG	SUM
Actual	Positive (PG)	5	27	0	32
	Neutral (NT)	6	114	1	121
	Negative (NG)	0	22	6	28
	SUM	11	163	7	181

表 27　引用目的：Basic（英語）

		Predict			
	Sentiment	PG	NT	NG	SUM
Actual	Positive (PG)	14	18	0	32
	Neutral (NT)	18	101	2	121
	Negative (NG)	0	16	12	28
	SUM	32	135	14	181

表 28　引用目的：Precise（英語）

Actual \ Predict	PG	NT	NG	SUM
Positive (PG)	27	3	2	32
Neutral (NT)	40	56	24	120
Negative (NG)	2	4	22	28
SUM	69	63	48	181

引用目的と同様に，引用感情についても追試した。

GPT3 を用いた本編では英語版のシンプルを採用し，表 10 にまとめた。上記の追試では表 26 が対応する。

この引用感情については，一致率 (Accuracy) はわずかに悪化しているものの，GPT3 との傾向は似通っており，大きな差はないと言える。

付録 C　プロンプト：実験用

プロンプト中，{ } で囲んだ部分には，実際のデータが埋め込まれる。

C.1　引用感情

C.1.1　Simple（日本語）

以下のテキストの著者が対象文献を引用する目的の種類を分類してください。
引用目的のカテゴリの種類は Background、Comparison、Criticize、Evidence、または Use です。
対象となる文献のラベルは Target: 以降，改行までです。
論文は Text: 以降の文章です。分類結果だけを答えてください。
Target: {分析対象となる引用文献のラベル}
Text:
{分析対象となる文章 (Paragraph)}

C.1.2　Simple（英語）

Classify the type of purpose for which the author of the following text is citing the target.
The type of citation purpose category is Background, Comparison, Criticize, Evidence, or Use.
The label of the target literature is Target: onwards, up to a new line. The thesis is the text after Text:.
Just report the classification result.
Target: {分析対象となる引用文献のラベル}
Text:
{分析対象となる文章 (Paragraph)}

C.1.3　Basic（日本語）

以下のマニュアルに基づいて、以下のテキストの著者が対象文献を引用する目的のカテゴリの種類を分類してください。
引用目的の種類は Background、Comparison、Criticize、Evidence、または Use です。
マニュアル:
———
テキストのテーマやトピックに関する一般的な背景情報を示すか、もしくは要約するために対象論文を引用している場合、Background と判断します。
テキストと対象論文、もしくは対象論文とその他の引用論文との間で結果や方法を比較する

ために対象論文を引用している場合、Comparison と判断します。
対象論文に対して何らかの評価や検討を加えるために引用している場合、Criticize と判断します。ポジティブな評価もネガティブな評価も含みます。
テキストの著者の意見や意思決定（手法の選択など）、解釈、主張、判断などの根拠や補強材料として対象論文を引用している場合、Evidence と判断します。
対象論文で示された手法やモデル（数式）、概念、理論、仮説、データ、ソフトウェアなどを利用するために引用を行っている場合、Use と判断します。

———

対象となる文献のラベルは Target: 以降，改行までです。
論文は Text: 以降の文章です。分類結果だけを答えてください。
Target: {分析対象となる引用文献のラベル}
Text:
{分析対象となる文章 (Paragraph)}

C.1.4 Basic（英語）

Classify the type of purpose for which the author of the following text is citing the target.
The type of citation purpose category is Background, Comparison, Criticize, Evidence, or Use based on the following manual.
Manual:

———

If the target is cited to present or summarize general background information about the research theme or topic of the text, you classify the text as Background.
If the target is cited to compare results or methods between the text and the target or between other cited papers, you classify the text as Comparison.
If the target is cited to provide some evaluation or review of the text, you classify the text as Criticize. Both positive and negative evaluations are included here.
If the target is cited to support or validate the author's claims, decisions (e.g., choice of methodology), interpretations, judgments, opinions, etc., you classify the text as Evidence.
If the target is cited to use methods, models, data, software, concepts, theories, hypotheses, etc. presented in the target, you classify the text as Use.

———

The label of the target literature is Target: onwards, up to a new line.
The thesis is the text after Text:. Just report the classification result.
Target: {分析対象となる引用文献のラベル}
Text:
{分析対象となる文章 (Paragraph)}

C.1.5 Precise（日本語）

以下のマニュアルに基づいて、以下のテキストの著者が対象文献を引用する目的のカテゴリの種類を分類してください。
引用目的の種類は Background、Comparison、Criticize、Evidence、または Use です。
マニュアル:

————

分類の手順は次の通りです。
1. 対象文献が含まれるセンテンスをまず読みます。当該のセンテンスを読んでカテゴリを明らかに判断できる場合、その時点で分類する。
2. 対象文献が含まれるセンテンスだけでは判断に迷う場合、当該のセンテンスの前後のセンテンスを一つずつ読みます。前後のセンテンスを一つずつ読んでカテゴリを明らかに判断できる場合、その時点で分類する。
3. 前後のセンテンスを一つずつ読んでも判断に迷う場合、パラグラフの最初から順番にセンテンスを読んでいく。カテゴリを明らかに判断できた時点で分類する。
4. 手順 3 までを実施して分類できない場合は Background に分類する。テキストのテーマやトピックに関する一般的な背景情報を示すか、もしくは要約するために対象論文を引用している場合、Background と判断します。
テキストと対象論文、もしくは対象論文とその他の引用論文との間で結果や方法を比較するために対象論文を引用している場合、Comparison と判断します。
対象論文に対して何らかの評価や検討を加えるために引用している場合、Criticize と判断します。ポジティブな評価もネガティブな評価も含みます。
テキストの著者の意見や意思決定（手法の選択など）、解釈、主張、判断などの根拠や補強材料として対象論文を引用している場合、Evidence と判断します。
対象論文で示された手法やモデル（数式）、概念、理論、仮説、データ、ソフトウェアなどを利用するために引用を行っている場合、Use と判断します。

————

対象となる文献のラベルは Target: 以降，改行までです。
論文は Text: 以降の文章です。分類結果だけを答えてください。
Target: {分析対象となる引用文献のラベル}
Text:
{分析対象となる文章 (Paragraph)}

C.1.6 Precise（英語）

Classify the type of purpose for which the author of the following text is citing the target based on the following manual.

The type of citation purpose category is Background, Comparison, Criticize, Evidence, or Use.
Manual:

———

The classification procedure is as follows:
1. Read the sentence containing the target first. If the category can be clearly determined by reading the sentence containing the target, the category is determined at that point.
2. If you are not sure about the decision based on the sentence containing the target alone, read the one sentence before and after it. If the category can be clearly determined by reading the sentences before and after the sentence containing the target, the category is determined at that point.
3. If the category cannot be determined after reading one sentence before or after the sentence containing the target, read the sentences in order from the beginning of the paragraph. If the category can be clearly determined, the category is determined at that point. If the target is cited to present or summarize general background information about the research theme or topic of the text, you classify the text as Background.
If the target is cited to compare results or methods between the text and the target or between other cited papers, you classify the text as Comparison.
If the target is cited to provide some evaluation or review of the text, you classify the text as Criticize. Both positive and negative evaluations are included here.
If the target is cited to support or validate the author's claims, decisions (e.g., choice of methodology), interpretations, judgments, opinions, etc., you classify the text as Evidence.
If the target is cited to use methods, models, data, software, concepts, theories, hypotheses, etc. presented in the target, you classify the text as Use.

———

The label of the target literature is Target: onwards, up to a new line.
The thesis is the text after Text:. Just report the classification result.
Target: {分析対象となる引用文献のラベル}
Text:
{分析対象となる文章 (Paragraph)}

C.1.7 Full（日本語）

以下のマニュアルに基づいて、以下のテキストの著者が対象文献を引用する目的のカテゴリの種類を分類してください。
引用目的のカテゴリは Background、Comparison、Criticize、Evidence、または Use です。
マニュアル:

———

分類の手順は次の通りです。
1. 対象文献が含まれるセンテンスをまず読みます。当該のセンテンスを読んでカテゴリを明らかに判断できる場合、その時点で分類する。
2. 対象文献が含まれるセンテンスだけでは判断に迷う場合、当該のセンテンスの前後のセンテンスを一つずつ読みます。前後のセンテンスを一つずつ読んでカテゴリを明らかに判断できる場合、その時点で分類する。
3. 前後のセンテンスを一つずつ読んでも判断に迷う場合、パラグラフの最初から順番にセンテンスを読んでいく。カテゴリを明らかに判断できた時点で分類する。
4. 手順 3 までを実施して分類できない場合は Background に分類する。テキストのテーマやトピックに関する一般的な背景情報を示すか、もしくは要約するために対象論文を引用している場合、Background と判断します。
次のようなキーワードが含まれている場合、Background と判断します。
Background のキーワードの例: “overview”, “review”, “summarize”
Background と判断する基準は次の通りです。
Background の判断基準: “最近の研究動向に関する情報を要約するために引用している場合”, “研究テーマに関する先行研究を単に紹介・言及している場合”, “一般的・全体的な情報（e.g., 政策動向、関連する研究領域や理論、etc）を述べる際に引用を行っている場合”, “他のどのカテゴリにも該当しない場合”
テキストと対象論文、もしくは対象論文とその他の引用論文との間で結果や方法を比較するために対象論文を引用している場合、Comparison と判断します。
次のようなキーワードが含まれている場合、Comparison と判断します。
Comparison のキーワードの例: “although”, “compare”, “comparison”, “contrast”, “however”, “in contrast”, “on the contrary”, “on the other hand”, “while”
Comparison と判断する基準は次の通りです。
Comparison の判断基準: “テキストの結果と対象論文の結果を比較して、テキストの結果の優位性を主張している場合”, “対象論文とその他の引用論文同士を比較して、一方の論文の優位性や短所を指摘している場合”
次のセンテンスは Comparison と判断されるセンテンスの例です。
Comparison のセンテンスの例：“However, while neurobiology posits that the rewarding properties of social behavior may have evolved to facilitate group cohesion and cooperation [4], our model suggests that polarization (as opposed to cohesion) across groups may be a side-effect of these rewarding properties.”
対象論文に対して何らかの評価や検討を加えるために引用している場合、Criticize と判断します。ポジティブな評価もネガティブな評価も含みます。
Criticize と判断する基準は次の通りです。
Criticize の判断基準: “対象論文の貢献や利点を評価している場合”, “対象論文の弱点や誤りを指摘している場合”

次のセンテンスは Comparison と判断されるセンテンスの例です。
Comparison のセンテンスの例: “The method in [4] reports a high result for the Media-lab dataset but does this using a dataset-specific SE and so it not a universal method.”
テキストの著者の意見や意思決定（手法の選択など）、解釈、主張、判断などの根拠や補強材料として対象論文を引用している場合、Evidence と判断します。
次のようなキーワードが含まれている場合、Evidence と判断します。
Evidence のキーワードの例: “aligns with”, “be consistent with”, “indicate to us”, “similar to”, “support”, “therefore”, “thus”
Evidence と判断する基準は次の通りです。
Evidence の判断基準: “テキストの著者が、自分の意見や仮説、意思決定をサポートするために対象を引用している場合”, “テキストの著者が、自分の研究や自分が支持する先行研究の方法論を正当化するために対象論文を引用している場合”, “テキストの著者の研究の仮定や限界点を正当化するために対象論文を引用している場合”, “今後の研究方向を提案する際に、その提案をサポートするために対象論文を引用している場合”
次のセンテンスは Evidence と判断されるセンテンスの例です。
Evidence のセンテンスの例：“Our findings emphasize that building digital seizing capabilities are contingent on pacing strategic actions, which aligns with dynamic capabilities research in hypercompetitive contexts [4].”
対象論文で示された手法やモデル（数式）、概念、理論、仮説、データ、ソフトウェアなどを利用するために引用を行っている場合、Use と判断します。
次のようなキーワードが含まれている場合、Use と判断します。
Use のキーワードの例: “based on”, “be carried over”, “provided by”, “use” Use と判断する基準は次の通りです。
Use の判断基準: “対象論文のデータセットを利用している場合”, “対象論文で提案/開発された方法を利用している場合”, “ある概念や理論に関する対象論文の定義を引用している場合”
次のセンテンスは Use と判断されるセンテンスの例です。
Use のセンテンスの例：“Our Arabic part-of-speech tagger uses the simplified PATB tag set proposed by [4]”

————

対象となる文献のラベルは Target: 以降，改行までです。
論文は Text: 以降の文章です。分類結果だけを答えてください。
Target: {分析対象となる引用文献のラベル}
Text:
{分析対象となる文章 (Paragraph)}

C.1.8 Full（英語）

Classify the type of purpose for which the author of the following text is citing the target based on the following manual.

The type of citation purpose category is Background, Comparison, Criticize, Evidence, or Use. I will give you the text later.

Manual:

———

The classification procedure is as follows:

1. Read the sentence containing the target first. If the category can be clearly determined by reading the sentence containing the target, the category is determined at that point.

2. If you are not sure about the decision based on the sentence containing the target alone, read the one sentence before and after it. If the category can be clearly determined by reading the sentences before and after the sentence containing the target, the category is determined at that point.

3. If the category cannot be determined after reading one sentence before or after the sentence containing the target, read the sentences in order from the beginning of the paragraph. If the category can be clearly determined, the category is determined at that point.

If the target is cited to present or summarize general background information about the research theme or topic of the text, you classify the text as Background.

If the sentences around the target contain the following keywords, the text is considered Background.

Keywords of Background: “overview”, “review”, “summarize”

The following are the criteria for classifying the text as Background.

Criteria of Background: “When the target is cited to summarize information on recent research trends”, “If the text simply introduces or refer to the target on its research topic”, “When the target is cited to state general or overall information (e.g., policy trends, relevant research areas or theories, etc.)”, “If the text does not fall into any other category”

If the target is cited to compare results or methods between the text and the target or between other cited papers, you classify the text as Comparison.

If the sentences around the target contain the following keywords, the text is considered Comparison.

Keywords of Comparison: “although”, “compare”, “comparison”, “contrast”, “however”, “in contrast”, “on the contrary”, “on the other hand”, “while”

The following are the criteria for classifying the text as Comparison.

Criteria of Comparison: “When cited to compare the results of the citing paper with those of previous studies and claim the superiority of the citing paper’s results”, “When comparing two

previous studies and pointing out the advantages or disadvantages of one study over the other"
If the sentences around the target are similar to the following examples, the text is considered Comparison.
An example of Comparison: "However, while neurobiology posits that the rewarding properties of social behavior may have evolved to facilitate group cohesion and cooperation [4], our model suggests that polarization (as opposed to cohesion) across groups may be a side-effect of these rewarding properties."
If the target is cited to provide some evaluation or review of the text, you classify the text as Criticize. Both positive and negative evaluations are included here.
The following are the criteria for classifying the text as Criticize.
Criteria of Criticize: "When the target is cited to evaluate the contribution or advantage of the text", "When the text is cited to point out a weakness or wrong in the text"
If the sentences around the target are similar to the following examples, the text is considered Criticize.
An example of Criticize: "The method in [4] reports a high result for the Media-lab dataset but does this using a dataset-specific SE and so it not a universal method."
If the target is cited to support or validate the author's claims, decisions (e.g., choice of methodology), interpretations, judgments, opinions, etc., you classify the text as Evidence.
If the sentences around the target contain the following keywords, the text is considered Evidence.
Keywords of Evidence: "aligns with", "be consistent with", "indicate to us", "similar to", "support", "therefore", "thus"
The following are the criteria for classifying the text as Evidence.
Criteria of Evidence: "When the target is cited to support the text' s author' s claim, hypothesis, or decision", "When the target is cited to justify the methodology of the text' s research or previous research the text' s author supports", "When the target is cited to justify the assumptions and limitations of the text", "When the text' s author proposes a future research direction, the author cites the target to support his proposal"
If the sentences around the target are similar to the following examples, the text is considered Evidence.
An example of Evidence: "Our findings emphasize that building digital seizing capabilities are contingent on pacing strategic actions, which aligns with dynamic capabilities research in hypercompetitive contexts [4]."
If the target is cited to use methods, models, data, software, concepts, theories, hypotheses, etc. presented in the target, you classify the text as Use.
If the sentences around the target contain the following keywords, the text is considered Use.
Keywords of Use: "based on", "be carried over", "provided by", "use"
The following are the criteria for classifying the text as Use.

Criteria of Use: “When the target is cited to use a dataset presented in the target”, “When the target is cited to use a method proposed or developed in the target”, “When the author of the text cites a definition on a concept or theory presented in the target”
If the sentences around the target are similar to the following examples, the text is considered Use.
An example of Use: “Our Arabic part-of-speech tagger uses the simplified PATB tag set proposed by [4].”

———

The label of the target literature is Target: onwards, up to a new line.
The thesis is the text after Text:. Just report the classification result.
Target: {分析対象となる引用文献のラベル}
Text:
{分析対象となる文章 (Paragraph)}

C.2 引用感情

C.2.1 Simple（日本語）

以下のテキストを Positive、Negative、または Neutral に分類してください。
対象となる文献のラベルは Target: 以降，改行までです。
論文は Text: 以降の文章です。分類結果だけを答えてください。
Target: {分析対象となる引用文献のラベル}
Text:
{分析対象となる文章 (Paragraph)}

C.2.2 Simple（英語）

Classify the text into Positive, Negative, or Neutral.
The label of the target literature is Target: onwards, up to a new line.
The thesis is the text after Text:. Just report the classification result.
Target: {分析対象となる引用文献のラベル}
Text:
{分析対象となる文章 (Paragraph)}

C.2.3 Precise（日本語）

以下のマニュアルに基づいて、以下のテキストを、Positive、Negative、または Neutral に分類してください。

マニュアル:

—

対象文献が含まれるセンテンスを、Positive、Negative、または Neutral に分類します。対象文献が含まれるセンテンスが接続詞などによって複数の文節に分かれており、分節によって Positive か Negative かが異なる場合には、対象文献が含まれる文節だけを読んで分類します。
対象文献が肯定的な意味のセンテンスで引用されている場合、Positive と判断します。
対象文献が否定的な意味のセンテンスで引用されている場合、Negative と判断します。
対象文献が引用されているセンテンスが Positive でも Negative でもない場合、Neutral と判断します。

—

対象となる文献のラベルは Target: 以降，改行までです。論文は Text: 以降の文章です。分類結果だけを答えてください。
Target: {分析対象となる引用文献のラベル}
Text:
{分析対象となる文章 (Paragraph)}

C.2.4 Precise（英語）

Classify the text into Positive, Negative, or Neutral based on the following manual.
Manual:

—

You classify the sentence in which the target is cited. If the sentence is divided into multiple clauses by conjunctions, etc., and whether the sentence is Positive or Negative differs depending on the clauses, you should read and classify only the clause containing the target.
If the target is cited in a sentence with a positive meaning, you classify it as Positive.
If the target is cited in a sentence with a negative meaning, you classify it as Negative.
If the target is cited in a sentence with neither Positive nor Negative meaning, you classify it as Neutral.

—

The label of the target literature is Target: onwards, up to a new line. The thesis is the text after Text:. Just report the classification result.
Target: {分析対象となる引用文献のラベル}
Text:
{分析対象となる文章 (Paragraph)}

C.2.5 Full（日本語）

以下のマニュアルに基づいて、以下のテキストを、Positive、Negative、または Neutral に分類してください。

マニュアル:

—

対象文献が含まれるセンテンスを、Positive、Negative、または Neutral に分類します。対象文献が含まれるセンテンスが接続詞などによって複数の文節に分かれており、分節によって Positive か Negative かが異なる場合には、対象文献が含まれる文節だけを読んで分類します。

対象文献が肯定的な意味のセンテンスで引用されている場合、Positive と判断します。

次のようなキーワードが含まれているセンテンスを、Positive と判断します。

Positive のキーワードの例：“be able to...”, “best”, “can...”, “could”, “develop”, “enhance”, “important”, “promote”, “robustly”, “support”, “well”

次のセンテンスは Positive と判断されるセンテンスの例です。

Positive のセンテンスの例：“The best known and simplest stochastic representation for discrete geophysical time series is the AR(1) model (Ghil et al. 2002; Bretherton and Battisti 2000).”

Positive のセンテンスの例：“These patterns find empirical support in Popp and Newell’s (2012) study of firm-level R&D spending and patents.”

Positive のセンテンスの例：“Although national and transnational connections may be necessary to secure access to resources and technical expertise, it is argued that local participation in the governance of social-ecological systems provides legitimacy (Biermann and Gupta 2011, Dryzek and Stevenson 2011), accommodates diverse interests and values (Brown 2003, Lebel et al. 2006), and taps local ecological knowledge (Berkes and Folke 2002, Gerhardinger et al. 2009, Raymond et al. 2010).”

対象文献が否定的な意味のセンテンスで引用されている場合、Negative と判断します。

次のようなキーワードが含まれているセンテンスを、Negative と判断します。

Negative のキーワードの例：“but”, “despite”, “even though”, “however”, “ignore”, “less”, “nevertheless”, “problematic”, “suffer”, “undermine”

次のセンテンスは Negative と判断されるセンテンスの例です。

Negative のセンテンスの例：“When women are unable to obtain sufficient water for menstrual ablutions or hygiene (e.g., cleaning menstrual cloths), they may suffer extreme stigma and humiliation (Rashid and Michaud 2000:54).”

Negative のセンテンスの例：“The need for better empirical information about energy-efficiency R&D is well known but difficult to solve due to lack of disaggregated data (although see on the contrary Popp (2002) and Popp and Newell (2012)).”

Negative のセンテンスの例：“Determination of the functions of fungal species, which typically

requires their isolation in pure culture and the study of their effects on defined substrates, has well-documented limitations [19–21].”
対象文献が引用されているセンテンスが Positive でも Negative でもない場合、Neutral と判断します。

—

対象となる文献のラベルは Target: 以降，改行までです。論文は Text: 以降の文章です。分類結果だけを答えてください。
Target: {分析対象となる引用文献のラベル}
Text:
{分析対象となる文章 (Paragraph)}

C.2.6 Full（英語）

Classify the text into Positive, Negative, or Neutral based on the following manual.
Manual:

—

You classify the sentence in which the target is cited. If the sentence is divided into multiple clauses by conjunctions, etc., and whether the sentence is Positive or Negative differs depending on the clauses, you should read and classify only the clause containing the target.
If the target is cited in a sentence with a positive meaning, you classify the text as Positive.
Sentences containing the following keywords are considered Positive.
Examples of Positive keywords: “be able to...”, “best”, “can...”, “could”, “develop”, “enhance”, “important”, “promote”, “robustly”, “support”, “well”
The following sentences are examples of sentences that are considered Positive.
Example of Positive sentences: “The best known and simplest stochastic representation for discrete geophysical time series is the AR(1) model (Ghil et al. 2002; Bretherton and Battisti 2000).”
Example of Positive sentences: “These patterns find empirical support in Popp and Newell’s (2012) study of firm-level R&D spending and patents.”
Example of Positive sentences: “Although national and transnational connections may be necessary to secure access to resources and technical expertise, it is argued that local participation in the governance of social-ecological systems provides legitimacy (Biermann and Gupta 2011, Dryzek and Stevenson 2011), accommodates diverse interests and values (Brown 2003, Lebel et al. 2006), and taps local ecological knowledge (Berkes and Folke 2002, Gerhardinger et al. 2009, Raymond et al. 2010).”
If the target is cited in a sentence with a negative meaning, you classify the text as Negative.
Sentences containing the following keywords are considered Negative.

Examples of Negative keywords: “but”, “despite”, “even though”, “however”, “ignore”, “less”, “nevertheless”, “problematic”, “suffer”, “undermine”

The following sentences are examples of sentences that are considered Negative.

Example of Negative sentences: “When women are unable to obtain sufficient water for menstrual ablutions or hygiene (e.g., cleaning menstrual cloths), they may suffer extreme stigma and humiliation (Rashid and Michaud 2000:54).”

Example of Negative sentences: “The need for better empirical information about energy-efficiency R&D is well known but difficult to solve due to lack of disaggregated data (although see on the contrary Popp (2002) and Popp and Newell (2012)).”

Example of Negative sentences: “Determination of the functions of fungal species, which typically requires their isolation in pure culture and the study of their effects on defined substrates, has well-documented limitations [19–21].”

If the target is cited in a sentence with neither Positive nor Negative meaning, you classify the text as Neutral.

—

The label of the target literature is Target: onwards, up to a new line. The thesis is the text after Text:. Just report the classification result.

Target: {分析対象となる引用文献のラベル}

Text:

{分析対象となる文章 (Paragraph)}

付録 D　プロンプト：LLM 用

D.1　引用目的

D.1.1　第 3 のクラスを Other に（英語）

Please classify the type of purpose for which the author of the following text is citing the target. The type of citation purpose category is Background, Evidence, or Others based on the following manual.

Manual:

—

If the target is cited to present or summarize general background information about the research theme or topic of the text, you classify the text as Background.

If the target is cited to support or validate the author’s claims, decisions (e.g., choice of methodology), interpretations, judgments, opinions, etc., you classify the text as Evidence.

If the type of citation purpose category is neither Background nor Evidence, you classify the text as Others.

—

The label of the target literature is Target: onwards, up to a new line.

The thesis is the text after Text:. Just report the classification result.

Target: {分析対象となる引用文献のラベル}

Text:

{分析対象となる文章 (Paragraph)}

D.1.2 第 3 のクラスを General に（英語）

Please classify the type of purpose for which the author of the following text is citing the target. The type of citation purpose category is General, Background or Evidence based on the following manual.

Manual:

—

Usually annotate as 'General'.

For example, 1. a comparison of results or methods between the text and the subject or other cited papers.

2. some kind of evaluation or review of the text.

3. use of methods, models, data, software, concepts, theories, hypotheses, etc.

All other items that are not background, evidence, etc., that cannot definitely be said to be background or evidence, or that cannot be classified in one specific category, etc., all belong to the 'General' category.

In addition, If the target is cited to present or summarize general background information about the research theme or topic of the text, you classify the text as Background.

If the target is cited to support or validate the author's claims, decisions (e.g., choice of methodology), interpretations, judgments, opinions, etc., you classify the text as Evidence.

—

The label of the target literature is Target: onwards, up to a new line.

The thesis is the text after Text:. Please, return only the classification results in one word.

Target: {分析対象となる引用文献のラベル}

Text:

{分析対象となる文章 (Paragraph)}

D.1.3 第 3 のクラスを Pending に（英語）

Please classify the type of purpose for which the author of the following text is citing the target. The type of citation purpose category is Pending, Background or Evidence based on the following manual.

Manual:

—

Wherever possible, judge the case as 'Pending'.

Anything that cannot definitely be categorised as Background or Evidence, or that cannot be placed in a specific category, belongs to 'Pending'. In addition,

If the target is cited to present or summarize general background information about the research theme or topic of the text, you classify the text as Background.

If the target is cited to support or validate the author's claims, decisions (e.g., choice of methodology), interpretations, judgments, opinions, etc., you classify the text as Evidence.

—

The label of the target literature is Target: onwards, up to a new line. The thesis is the text after Text:. Please, return only the classification results in one word.

Target: {分析対象となる引用文献のラベル}

Text:

{分析対象となる文章 (Paragraph)}

D.1.4 第 3 のクラスを Unknown とし，クラスラベルを隠蔽（英語）

Please provide a classification of the citation in the text.

Classification target is only one citation.

The label of the target citation is [Target:] onwards, up to a new line. The text is after [Text:].

The criteria for classification are as follows.

There are three classes: UKN, BKG and EVS.

The criteria for classification are given below.

Basically, classified as UKN.

This option (i.e. UKN) is extremely strongly recommended.

Try to select this option whenever possible.

Where the purpose of the citation is presumed to be background, such as including present or summarize general background information about the research theme or topic, it is classified as BKG.

This option (i.e. BKG) is second most strongly recommended.

Where the purpose of the citation is presumed to be evidence, such as support or validate the author's claims, decisions (e.g., choice of methodology), interpretations, judgments, opinions, etc., and where there is no room for any other interpretation at all, it is classified as EVS.

This option (i.e. EVS) is deprecated, avoid as much as possible.

Please, return only the classification results in just one word.

Target: {分析対象となる引用文献のラベル}

Text:
{分析対象となる文章 (Paragraph)}

D.1.5 BKG か否か，EVS か否かの 2 値分類とし，クラスラベルを隠蔽（英語）

Please provide a classification of the citation in the text.
Classification target is only one citation.
The label of the target citation is [Target:] onwards, up to a new line. The text is after [Text:].
The criteria for classification are as follows.
There are two types of classes.
Basically, classified as NB.
However, where the purpose of the citation is presumed to be background, such as including present or summarize general background information about the research theme or topic, it is classified as PB.
Please, return only the classification results in just one word.
Target: {分析対象となる引用文献のラベル}
Text:
{分析対象となる文章 (Paragraph)}

D.1.6 BKG か否かの 2 値分類とし，クラスラベルを隠蔽（英語）

Please provide a classification of the citation in the text.
Classification target is only one citation.
The label of the target citation is [Target:] onwards, up to a new line. The text is after [Text:].
The criteria for classification are as follows.
There are 2 classes: UKN and BKG.
The criteria for classification are given below.
Basically, classified as UKN.
This category includes, for example, citations as evidence, criticism, comparison, discussion, etc.
Try to select this option whenever possible.
The other hands, Where the purpose of the citation is presumed to be background, such as including present or summarize general background information about the research theme or topic, it is classified as BKG.
Please, return only the classification results in just one word.
Target: {分析対象となる引用文献のラベル}
Text:
{分析対象となる文章 (Paragraph)}

D.2 引用感情

D.2.1 第 3 のクラスを Not Sure に（英語）

Please classify the text into Positive, Negative, Neutral, or Not Sure based on the following manual.

Manual:

—-

You classify the sentence in which the target is cited.

If the target is cited in a sentence with a positive meaning, you classify it as Positive.

If the target is cited in a sentence with a negative meaning, you classify it as Negative.

If the target is cited in a sentence with neither Positive nor Negative meaning, you classify it as Neutral.

When you are unsure how to classify a sentence, you classify it as Not Sure.

—-

The label of the target literature is Target: onwards, up to a new line.

The thesis is the text after Text:. Just report the classification result.

Target: {分析対象となる引用文献のラベル}

Text:

{分析対象となる文章 (Paragraph)}

D.2.2 第 3 のクラスを Pending に（英語）

Please classify the text into Positive, Negative, Neutral or Pending based on the following manual.

Manual:

—-

You classify the sentence in which the target is cited.

Wherever possible, judge the case as 'Pending'.

If the target is cited in a sentence with a definitely positive meaning, you classify it as 'Positive'.

If the target is cited in a sentence with a definitely negative meaning, you classify it as 'Negative'.

If the target is cited in a sentence with a definitely neutral meaning, you classify it as 'Neutral'.

—-

The label of the target literature is Target: onwards, up to a new line.

The thesis is the text after Text:. Just report the classification result.

Target: {分析対象となる引用文献のラベル}

Text:

{分析対象となる文章 (Paragraph)}

D.2.3 第 3 のクラスを Unknown とし，そのクラスラベルを隠蔽（英語）

Please provide a classification of the citation in the text.
Classification target is only one citation.
The label of the target citation is [Target:] onwards, up to a new line. The text is after [Text:].
The criteria for classification are as follows.
There are 3 classes: Positive, Negative and UKN.
The criteria for classification are given below.
Basically, classify them as UKN.
This category also includes the classification result of neutral.
However, if the target citation is presumed to be cited positively or negatively, output the presumption.
Please, return only the classification results in just one word.
Target: {分析対象となる引用文献のラベル}
Text:
{分析対象となる文章 (Paragraph)}

D.2.4 対象引用文献に関する言及のみを抽出（英語）

Please only extract statements relating to the target citation.
The target is only one citation.
The label of the target citation is [Target:] onwards, up to a new line. The text is after [Text:].
Please, return only extract statements. Do not include any other text.
Target: {分析対象となる引用文献のラベル}
Text:
{分析対象となる文章 (Paragraph)}