

卒業論文

タスク指示文自動生成を用いた評価と作成の反復による
依頼者の意図を反映した指示文作成

田中 涼

2025年3月6日

岐阜大学 工学部 電気電子・情報工学科 情報コース
鈴木研究室

本論文は岐阜大学工学部に
学士（工学）授与の要件として提出した卒業論文である。

田中 涼

指導教員：

鈴木 優 准教授

タスク指示文自動生成を用いた評価と作成の反復による 依頼者の意図を反映した指示文作成*

田中 涼

内容梗概

クラウドソーシングでは、依頼者がタスク指示文を作成し、ワーカがその指示文を用いてタスクを遂行することによってデータセットが作成される。作成されるデータセットの品質や一貫性はワーカへの指示文の明瞭さに影響されるが、依頼者が自身で作成した指示文を説明不足かどうか判断することは難しい。そこで我々は、指示文を用いてタスクを遂行する能力は LLM と人間で等しいと仮定し、LLM にとって説明不足でない指示文を作成し、依頼者にとって望ましい指示文の作成を目指した。具体的には評価用 LLM と作成用 LLM を使い、指示文の作成とその指示文に対するフィードバックを交互に繰り返し行うことによって、明瞭な指示文を作成する。評価用 LLM は指示文を用いてタスクを遂行し、付与したラベルと正解ラベルとの一致率によって指示文を評価し、評価した指示文をフィードバックとして作成用 LLM に与える。作成用 LLM は正解ラベル、および指示文とその評価を元にして、新しい指示文を作成する。実験の結果、ワーカが提案手法による指示文を用いた場合、依頼者による指示文を用いた場合よりも高い Accuracy を出すことが可能であることを確認した。

キーワード

LLM, クラウドソーシング, データセット作成, タスク指示文, フィードバック

*岐阜大学 工学部 電気電子・情報工学科 情報コース 卒業論文, 学籍番号: 1213033096, 2025 年 3 月 6 日.

目次

図目次	iv	
表目次	v	
第 1 章	はじめに	1
第 2 章	基本的事項	4
2.1	ニューラルネットワーク	4
2.2	LLM	4
2.3	クラウドソーシング	5
2.4	能動学習	5
2.5	GAN	6
2.6	評価指標	6
2.7	対応のない t 検定	6
2.8	独立性のカイ 2 乗検定	7
第 3 章	関連研究	8
第 4 章	提案手法	10
4.1	評価用 LLM	12
4.2	作成用 LLM	14
第 5 章	評価実験	17
5.1	実験 1(提案手法による指示文の作成)	17
5.1.1	実験手順	17
5.1.2	使用データ	17
5.1.3	実験条件	18
5.1.4	実験結果と考察	18
5.2	実験 2(提案手法によって作成された指示文の評価)	23
5.2.1	実験手順	23

5.2.2	使用データ	23
5.2.3	実験条件	23
5.2.4	実験結果と考察	24
第 6 章	おわりに	26
	謝辞	28
	参考文献	29
	発表リスト	31

図目次

1.1	提案手法の概要図	3
4.1	提案手法の概要図	11
4.2	評価用 LLM に与えるプロンプト	12
4.3	評価用 LLM に与えるプロンプトの具体例	12
4.4	評価用 LLM の出力の具体例	13
4.5	作成用 LLM に与えるプロンプト	15
4.6	作成用 LLM に与えるプロンプトの具体例	16
5.1	依頼者によって作成された指示文	18
5.2	提案手法によって作成された指示文 (1/2)	21
5.2	提案手法によって作成された指示文 (2/2)	22

表目次

4.1	評価用 LLM によって付与されたラベルと正解ラベルの比較例 . . .	13
5.1	評価用 LLM による指示文の評価	19
5.2	ワークごとの Accuracy とグループごとの平均 Accuracy	24
5.3	各グループが付与したラベルの正誤答数と Accuracy	25

第1章 はじめに

本研究では、クラウドソーシングにおいてワーカーにとって明瞭さの高い指示文を作成することによって、依頼者にとって望ましいデータセットを作成することを目標とする。クラウドソーシング [1] とは、インターネットを通じて仕事を不特定多数の作業者に依頼する仕組みである。仕事を依頼する利用者を依頼者、依頼する仕事の内容をタスク、タスクを遂行する作業者をワーカー、仕事を遂行する手順をまとめた文章を指示文と呼ぶ。指示文の内容は、ワーカーが依頼者の望むデータセットを作成できるかどうかに関係している [2, 3]。特に、指示文が説明不足であった場合、ワーカーは依頼者の想定とは異なる手順や方法でタスクを遂行してしまう可能性がある。例として、次の指示文に従って文章 1、文章 2 にラベルを付与することを考える。

指示文 与えられた文章に主観的、もしくは客観的のラベルを付与してください。両方含まれる場合でもどちらか片方のラベルだけを付与してください。

文章 1 2024/1/8 現在、金の店頭小売価格は税込 14,783 円/g である。

文章 2 スーパーの卵 1 パック 160 円か、安いな

文章 1 は客観的な事実だけを述べているため、ほとんどのワーカーは確信を持って客観的であると判断できる。一方、文章 2 は客観的な事実をもとに主観的な意見を述べているため、ワーカーは文章 2 に対して「安いな」という意見に注目すれば主観的、「スーパーの卵 1 パック 160 円」という事実注目すれば客観的と判断する。ところがどちらが主であるかについて指示されていないため、ワーカーは指示文において文章のどの点に注目すべきかが分からず、各々の判断のもとラベルを付与することとなる。その結果、タスクを依頼した際にラベルを付与したワーカーの勝手な判断によって付与されるラベルが変わることとなる。このように作業者は説明不足の指示文を作業依頼者と異なる解釈をしてしまい、依頼者の望むラベルが文章に付与されない。

依頼者が持つ偏見や依頼者とワーカー間での知識の相違などを原因として、依頼者は依頼者自身が作成した指示文の明瞭さを判断することは難しい。例えば、先ほどの例で挙げた指示文を作成した依頼者が、文章が主観的であるか客観的であるかは

文末に依存するという偏見を持っていたとする。この場合、この指示文はワーカにとっては説明が不足しているが、依頼者は説明が十分であると考えている。依頼者は自身がどのような偏見を持っているかを自覚することが難しいため、依頼者自身が自身の作成した指示文を評価することは難しい。

そこで我々は、指示文を用いてタスクを遂行する能力は大規模言語モデルとワーカで等価であると仮定し、作業者にとって明瞭な指示文を作成する代わりに大規模言語モデルにとって明瞭な指示文を作成することを目指した。大規模言語モデル (Large Language Model ; LLM)[4] とは、ある文章に続く単語を文脈から予測するモデルである言語モデルのうち、大量のテキストデータとニューラルネットワークによって構築された機械学習モデルである。LLM を用いて指示文を作成するという発想は GAN[5](敵対的生成ネットワーク) を基にしている。GAN では、画像の生成モデルと画像が実際に撮影された本物か生成モデルによる偽物かを見分ける識別モデルを繰り返し使用することによって、より実際に撮影された画像に近い画像を得ることができる。この手法において機械学習モデルが繰り返し使用されている点に着目し、指示文を作成する LLM である作成用 LLM と指示文を評価する LLM である評価用 LLM を交互に繰り返し使用することによって、LLM およびワーカにとってより明瞭な指示文を得られると考えた。指示文を作成するにあたり、依頼者自身が文章に付与したラベル、正解ラベルを使用する。正解ラベルを使用する理由は、正解ラベルには依頼者の偏見が反映されており、その偏見を指示文に与えることによって指示文の説明不足が解消されると考えたためである。

提案手法の概要図を図 1.1 に示す。提案手法では、まず評価用 LLM が指示文と分類されていない文章を受け取り、その指示文に従って文章にラベルを付与する。こうして付与されたラベルと正解ラベルの一致率を指示文に評価として与える。作成用 LLM は指示文を作成する際、正解ラベルが付与された文章、指示文とその評価を受け取り新しい指示文を作成する。

評価実験では、提案手法によって作成された指示文が依頼者によって作成された指示文と比べてより明瞭であると言えるのか確かめた。実験では第一著者が依頼者としてタスク、正解ラベル、指示文を用意した。被験者をグループ A と B に分け、グループ A には依頼者が作成した指示文を、グループ B には提案手法によって作成された指示文をそれぞれ与え、それぞれのグループに与えられた指示文に従って

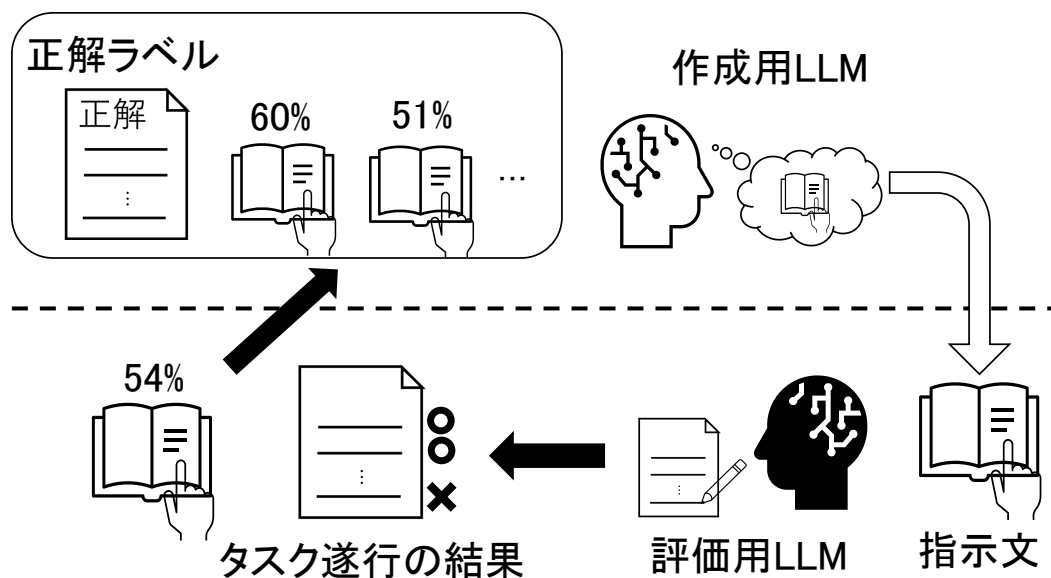


図 1.1: 提案手法の概要図

文章にラベルを付与するよう依頼した。

実験の結果、グループ B の方がグループ A と比べて 3% 高い Accuracy を記録したことから、提案手法を用いて明瞭な指示文を作成できることが分かった。本研究の貢献は以下のとおりである。

- 2 種類の LLM を交互に繰り返し使用することによって、明瞭な指示文が作成できることを確認した。
- 明瞭な指示文を用いることによって、グループのタスクにおける Accuracy が 3% 向上することを確認した。

本論文の構成は以下の通りである。2 章では、基本的事項について述べる。3 章では、関連研究について述べる。4 章では、明瞭な指示文を作成する手法について述べる。5 章では、評価実験について述べる。6 章では、本研究のまとめと今後の展望について述べる。

第 2 章 基本的事項

2.1 ニューラルネットワーク

ニューラルネットワークとは、人間の脳の神経細胞であるニューロンを基にした構造を持つ技術や、その技術を用いて作成された機械学習モデルのことである。機械学習モデルとは、既知の入力とそれに対する出力を学習することによって、未知の入力に対する出力を行うモデルである。機械学習モデルを用いた例として、過去 3 年間の気温とアイスの売り上げから学習を行い、今日の気温からアイスの売り上げを予測するモデルが考えられる。既存の機械学習モデルは、人間がモデルに対して入力のどこに注目すべきなのか設定していたため、予測の精度がモデルを作成した人物によって変動する。一方ニューラルネットワークは、入力のどこに注目すべきなのかを自分で学習するため、予測の精度が作成した人間に左右されない。また、ニューラルネットワークは既存の機械学習モデルと比べて複雑な構造をしているため、より複雑な問題に対応することが可能である。

2.2 LLM

LLM(Large Language Model；大規模言語モデル)とは、大量のテキストデータとニューラルネットワークを使用して作られた言語モデルである。言語モデルとは、文章に対してある単語がその文章の続きにくる確率を学習することによって、文脈に従って文章の続きを生成するモデルである。例えば「今日の天気は」という文章に対しては、晴れや曇りといった単語は文章に続く確率が高いと、犬や楽しいは文章に続く確率が低いと予測し、今日の天気はの続きの単語を生成する。従来 of 言語モデルは長い文章において、離れた二つの単語の関係を把握することができないことが課題となっていた。しかし、ニューラルネットワークを用いた言語モデルは、長い文章においても単語同士の関係を把握でき、より正確な確率に基づいた文章の続きの単語を生成することが可能となった。LLM の代表例として、

OpenAI 社が開発した ChatGPT^{*}、Meta 社が開発した Llama[†]、Google 社が開発した Gemini[‡]などが挙げられる。

2.3 クラウドソーシング

クラウドソーシングとは、業務を特定の個人や団体に業務を委託するのではなく、インターネットを通じて不特定多数の人々に業務を依頼する方法である。業務の依頼はクラウドソーシングプラットフォームで公開され、作業者はその業務の参加条件などを参照し、どの業務を遂行するか選ぶ。クラウドソーシングにて遂行される業務内容は、プログラミング、ブログ内容の執筆、データ入力などの事務作業、アンケートなど多岐に渡る。クラウドソーシングは業務の規模に応じて報酬を設定できるため、アウトソーシングする場合と比べて費用対効果に優れる。また、作業者は業務時間が指定されず業務内容を選ぶことができるため、各々が望むタイミングで望む業務を遂行することができる。作業依頼者から見たクラウドソーシングの欠点として、作業者の中にスパムワーカが含まれた場合、業務が正しく遂行されない可能性がある点が挙げられる。スパムワーカとは、報酬を受け取るために業務を乱雑に遂行して数をこなそうとする作業者や、嫌がらせのために業務をでたらめに遂行する作業者のことを指す。

2.4 能動学習

能動学習とは、モデルが自らの学習に必要な情報を選び出し、その情報を人間や他の情報源から取得することによって効率的に学習を行う仕組みである。例として能動学習を用いて文章にラベルを付与するモデルを作成することを考える。モデルは既存の正解ラベルが付与された文章から、どのようにラベルを付与すべきか学習を行う。モデルはその学習結果に基づき正解ラベルが付与されていない文章にラベルを付与する。モデルは自身がどのラベルを付与すべきか判断の難しい文章を集

^{*}<https://chatgpt.com/>

[†]<https://www.llama.com/>

[‡]<https://gemini.google.com/>

め、それらの文章の正解ラベルを他の情報源から取得する。学習と分類、学習に必要な情報の取得を繰り返すことによって、モデルは効率的に学習を行う。

2.5 GAN

GAN(敵対的生成ネットワーク)とは、画像の生成モデルと画像が本物か生成モデルによる偽物かを見分ける識別モデルを繰り返し使用することによって、より実際に撮影された画像に近い画像を得ることができる機械学習モデルである。生成モデルは学習が進んでいない状態ではほとんどノイズのような画像を出力するが、識別モデルが本物であると判断した画像から学習をすることで、識別モデルによって偽物であると看破されない画像を生成することができる。識別モデルは本物と判断した偽物の画像や、偽物と判断した本物の画像から、生成モデルによって生成される画像を偽物であると看破できるように学習する。生成モデルと識別モデルがお互いを出し抜くように学習し続けることによって、最終的に識別モデルによって識別されないほど精巧な偽物の画像が得られる。

2.6 評価指標

本研究では、提案手法の評価のために Accuracy を使用する。Accuracy は正答率を意味しており、この研究では特に文章に付与されたラベルと正解ラベルとの一致率を表す。Accuracy は以下の式で計算される。

$$Accuracy = \frac{1}{n} \sum_{i=1}^n C_i \quad (2.6.1)$$

ここで n は文章の個数、 C_i は i 番目の文章に正解ラベルと同じラベルが付与されていたならば 1 を、そうでなければ 0 をとる変数である。

2.7 対応のない t 検定

対応のない t 検定は対応のない二つの母集団に対して、二つの母平均に差があると言えるのかどうかを判定する。この検定では「二つの母平均は等しい」という帰

無仮説を立て、二つの母集団からそれぞれ抽出した二つの標本を用いて検定量を計算する。特に二つの標本のサイズが同じ場合、この検定における検定量 t は以下の式で計算される。

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{n}}} \quad (2.7.1)$$

n は標本のサイズ、 \bar{x}_1 と $\hat{\sigma}_1^2$ は一つ目の、 \bar{x}_2 と $\hat{\sigma}_2^2$ は二つ目の標本における標本平均と不偏分散を表す。検定量 t から導出される p 値が有意水準を満たした場合、帰無仮説が棄却される。

2.8 独立性のカイ 2 乗検定

独立性のカイ 2 乗検定は 2 種類の特徴量に基づいて集計された標本に対して、2 種類の特徴量は独立であると言えるのか判定する。独立性のカイ 2 乗検定では「2 種類の特徴量に相関はない」という帰無仮説を立て、以下の式で求められるカイ 2 乗値を用いてその帰無仮説が棄却されるか検定する。

$$\chi^2 = \sum_{i=1}^n \sum_{j=1}^m \frac{(x_{ij} - y_{ij})^2}{y_{ij}} \quad (2.8.1)$$

n は 1 種類目の、 m は 2 種類目の特徴量を持つカテゴリ数をそれぞれ表す。 x_{ij} と y_{ij} は、1 種類目の特徴量のカテゴリが i かつ 2 種類目の特徴量のカテゴリが j であった場合の実測値、期待値を表す。自由度 df は以下の式で計算される。

$$df = (n - 1)(m - 1) \quad (2.8.2)$$

自由度とカイ 2 乗値から導出される p 値が有意水準を満たした場合、帰無仮説は棄却される。

第3章 関連研究

クラウドソーシングにおいて指示文を改善することによって、質の高いデータセットを得ることを試みた研究を紹介する。Manamら [6] は、指示文に含まれる曖昧な表現を検出するため、スペルミスを検出する Bing Spell Check, 専門用語や一般的ではない単語を検出する Textstat, 「多くの」「たくさんの」といった曖昧な表現を検出する Hedge Word Detection など、12 種類の自然言語処理に使用されるプログラム群によって構成される TaskLint を提案した。TaskLint が検出できる曖昧さの例として、一般的でない単語の使用, 単語の多様性, 回答するフォーマットの未指定が挙げられている。TaskLint を利用した指示文を使ったタスクは、利用しなかった指示文を使ったタスクに比べて 1.58 倍正解ラベルの付与率が高く、タスクの完了時間が短くなることが明らかになった。Monam らは TaskLint では検出できない曖昧さの一つに、指示文の文脈が曖昧さに関わっている場合を挙げている。例として、最新の Samsung iPhone の重量を調べてください、という指示文には Samsung が販売している最新のスマートフォンの重量を調べるのか、最新の iPhone の重量を調べるのか、という曖昧さが含まれている。LLM は文脈を読むことができるため、LLM を使用することによって TaskLint では検出できない曖昧さが取り除かれた明瞭な指示文を作成することができると考え、LLM に指示文の作成を行わせた。

Nouri ら [7] は指示文に含まれる曖昧さを不明確な手順、難解な表現、定義されていない用語の使用など 7 種類の要素に分類し、これらの要素がそれぞれ指示文に含まれているのかどうかを判断するモデルの構築を目指した。モデルは文章の長さなど指示文に含まれる計算可能な特徴量から、各要素がそれぞれ含まれているのかを学習する。実験の結果、モデルは 55% から 71% の精度で指示文に曖昧さが含まれているのかを正しく予測することができた。本研究では LLM が指示文の内容など計算可能ではない特徴量を基にして、指示文に曖昧さが含まれているのかどうか判断できると考えたため、LLM に指示文の作成を行わせた。

Nouri ら [8] は、指示文に曖昧さが含まれているかどうかを判定するだけでなく具体的な改善案を提供する ClarifyIt を提案した。ClarifyIt は入力された指示文に対して、言葉遣いとその表現、タスク遂行のステップの明示、タスク承認基準の明

確さなど 8 種類の明確性のうちどの明確性がどの程度欠けているのか検出した後、その明瞭性の種類と程度に応じて改善案を提案する。改善案にはその改善案がどの程度信頼できるのかを示す信頼性スコアが付与されており、ユーザは改善案とスコアを参照することで、曖昧さの排除された指示文を作成することができる。実験では依頼者が ClarifyIt を用いて指示文の改善を行った。ワーカは 8 種類の明瞭性それぞれの観点において、改善前の指示文と改善後の指示文のどちらが明瞭であるか投票した。投票の結果、明瞭性の種類によるものの、60% から 78% のワーカが改善後の指示文の方が明瞭であると答えたことが分かったため、Nouri らは改善後の指示文の方が改善前の指示文よりも明瞭であると結論付けた。指示文が曖昧であるかの判断を、Nouri らの研究においては ClarifyIt が指示文の内容から曖昧さを検出するのかどうかに基づいて行っているが、本研究では評価用 LLM による指示文を用いたタスク遂行の結果をに基づいて行っている。

LLM を繰り返し使用することによって、1 回のみ使用した場合よりも望ましい返答が得られることを目指した研究を紹介する。Radha ら [9] は、LLM による応答の品質を向上するため、自律的な思考の反復 (Autonomous Iteration of Thought ; AIoT) を提案した。AIoT は、内部対話エージェント (Internal Dialogue Agent ; IDA) と LLM エージェント (Large Language Model Agent ; LLMA) を繰り返し利用することで、ユーザに対する応答の改善を目指す。LLMA はユーザのプロンプトを受け取り、そのプロンプトに対する応答を生成する。IDA は LLMA によって生成された応答がユーザにとって十分なものか判断し、不十分と判断した場合に LLMA から十分な応答が生成されるようなプロンプトを生成し、そのプロンプトを LLMA に渡す。LLMA は渡されたプロンプトから再度応答を生成する。これらの手順を、IDA が LLMA による応答が十分であると判断するまで繰り返す。AIoT により生成された返答は、単純に LLM に返答を生成させた場合と比べて 14.11% 高い精度を達成した。AIoT と本研究の相似点として、AIoT における LLMA の出力を IDA が評価している構造と本研究における作成用 LLM の出力を作成用 LLM が評価している構造が挙げられる。一方 AIoT と本研究の相違点として、AIoT はユーザのプロンプトに対する適切な応答を作成することを目指している一方、本研究は依頼者の偏見が反映された指示文を作成することを目指している。

第 4 章 提案手法

提案手法では、評価用 LLM と作成用 LLM を交互に繰り返し利用することによって、LLM にとって明瞭な指示文を作成する。提案手法の概要を図 4.1 に示す。

提案手法における文字の定義を以下に示す。作成用 LLM が指示文を作成する回数を n 、作成用 LLM によって i 回目に作成された指示文を s_i 、評価用 LLM による s_i の評価を e_i とする。指示文の作成のために用いる文章の個数を m 、そのうち j 個めの文章を d_j 、その文章の正解ラベルを c_j とする。それらの文章によって構成された文章群を $D = \{d_1, d_2, \dots, d_m\}$ 、それらの正解ラベルによって構成された正解ラベル群を $C = \{c_1, c_2, \dots, c_m\}$ とする。また、指示文の評価のために用いる文章の個数を m' 、 j' 個目の文章を $d'_{j'}$ 、文章によって構成される文章群を $D' = \{d'_1, d'_2, \dots, d'_{m'}\}$ とする。 $d'_{j'}$ の正解ラベルを $c'_{j'}$ 、それらの正解ラベルによって構成される正解ラベル群を $C' = \{c'_1, c'_2, \dots, c'_{m'}\}$ とする。評価用 LLM が指示文 s_i を用いて $d'_{j'}$ に付与したラベルを $l_{ij'}$ 、それらのラベルによって構成されるラベル群を $L_i = \{l_{i1}, l_{i2}, \dots, l_{im'}\}$ とする。ただし s^* は依頼者によって作成された指示文、 e^* は s^* の評価、 L^* は評価用 LLM が s^* を用いて m' 個の文章に付与したラベルによって構成されるラベル群である。

提案手法の手順を以下に示す。

1. 評価用 LLM が s^* に対する e^* を Accuracy を尺度として求め、 s^*, e^* を作成用 LLM にフィードバックする。
2. 作成用 LLM が D, C, s^*, e^* を参照して s_1 を作成する
3. 評価用 LLM が s_1 に対する e_1 を Accuracy を尺度として求め、 s_1, e_1 を作成用 LLM にフィードバックする。
4. 作成用 LLM は D, C, s^*, e^* に加えて、 s_1, e_1 を参照して s_2 を作成する。
5. 以降、作成用 LLM は i 回目の指示文作成にて $D, C, s^*, e^*, s_1, e_1, \dots, s_{i-1}, e_{i-1}$ を参照して s_i を作成し、評価用 LLM は e_i を計算する。
6. (5) を e_n が得られるまで繰り返す。
7. $\{e_1, e_2, \dots, e_n\}$ のうち e_k が最大であったとき、 s_k を提案手法によって作成された指示文とする。

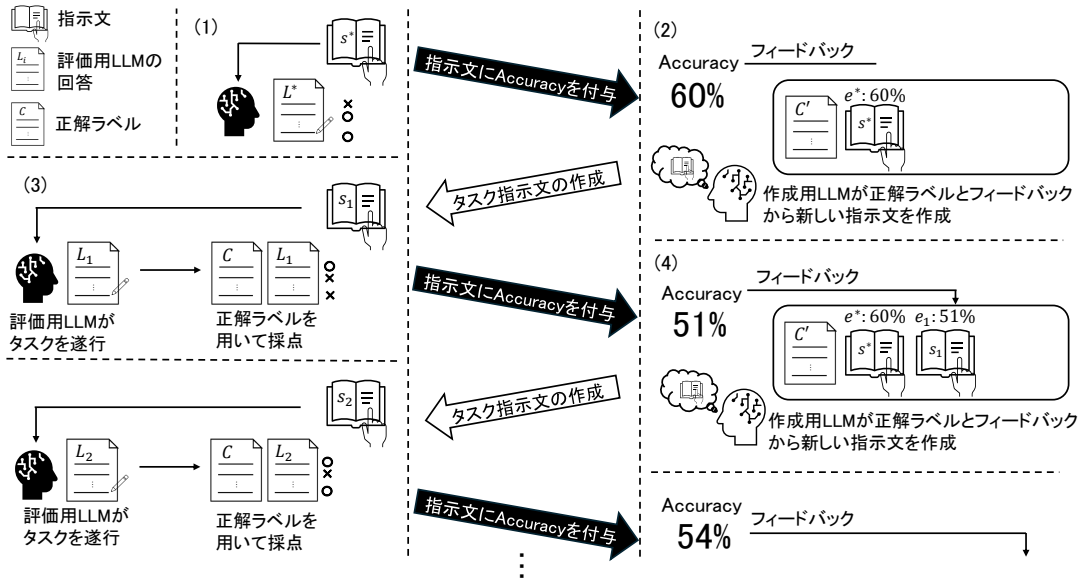


図 4.1: 提案手法の概要図

s_i の Accuracy は以下の式で計算される.

$$Accuracy = \frac{1}{n} \sum_{j=1}^{m'} f(l_{ij}, c'_j) \quad (4.0.1)$$

この式において $f(x_1, x_2)$ は $x_1 = x_2$ ならば 1 を, そうでなければ 0 を返す関数である.

また, (7) にて s_k と同じ評価を持つ指示文が複数存在した場合, その中で最後に作成されたものを $s_{k'}$ として, $s_{k'}$ を提案手法によって作成された指示文とする. 後に作成された指示文を選択する理由は, その指示文は後に作成された分多くのフィードバックを基に作成されており, 同じ正答率を持つ指示文の中でも明瞭であると考えたためである.

評価用 LLM は 4.1 節にて, 作成用 LLM は 4.2 節にて詳細を述べる.

文章に、タスク指示文に従って、ラベルを付与してください。波括弧で囲まれた部分を一つの文章としてください。また、返答は波括弧で囲われた文章、波括弧で囲われていないラベルという順番で行ってください。 [task instruction][unlabeled text]

図 4.2: 評価用 LLM に与えるプロンプト

文章に、タスク指示文に従って、ラベルを付与してください。波括弧で囲まれた部分を一つの文章としてください。また、返答は波括弧で囲われた文章、波括弧で囲われていないラベルという順番で行ってください。

文章が主観的であれば 0 のラベルを、客観的であれば 1 のラベルをそれぞれ付与してください。

{ 絵とかストーリーが凝っててめっちゃ好き } { 書類提出の締め切りは 12/18 となります。よろしくお願ひします。 } { スーパーの卵 1 パック 160 円か、安いな } …

図 4.3: 評価用 LLM に与えるプロンプトの具体例

4.1 評価用 LLM

評価用 LLM は指示文とラベルが付与されていない文章とを入力として、指示文に従ってラベルが付与された文章を出力する役割をもつ LLM である。評価用 LLM には、図 4.2 に示すプロンプトを与える。

[task instruction] は評価用 LLM に評価させる指示文を、[unlabeled text] はラベルが付与されていない文章をそれぞれ表す。例として、文章が主観的か客観的か判断するタスクのための指示文を作成することを考える。このとき評価用 LLM に与えるプロンプトは図 4.3 のようになる。

ここで、文章を波括弧で囲んだ状態で与えているのは、文章が二つ以上の文で構成されていた場合において、どこまでが一つの文章なのか区別するためである。ま

{ 絵とかストーリーが凝っててめっちゃ好き}1 { 書類提出の締め切りは 12/18
 となります。よろしくお願ひします。}0 { スーパーの卵 1 パック 160 円か、
 安いな}1 …

図 4.4: 評価用 LLM の出力の具体例

た、文章を囲うための記号は文章に含まれにくい記号として、波括弧を使用した。

評価用 LLM は s_i に基づいてタスクを遂行し、 L_i を出力する。その出力結果は、図 4.4 に示すような文章とラベルが交互に並ぶ一つの文字列である。文章とラベルを区別するために、文章を波括弧で囲い、ラベルは波括弧で囲わずに出力させる指示を評価用 LLM に与えている。

文章とラベルを区別したのち、 l_{ij} と c'_j を比較することによって e_i を求める。具体例を表 4.1 に示す。この場合、評価用 LLM は初めの二つの文章はどちらも正解ラベルと同じラベルを付与しているが、三つ目の文章は正解ラベルと異なるラベルを付与している。この例において評価用 LLM によってラベルが付与された 50 個の文章のうち、30 個が正解ラベルと同じラベルを付与されていたならば、 s_i の Accuracy である e_i は 60% となる。こうして算出された e_i は s_i とともに作成用 LLM に対してフィードバックされる。

表 4.1: 評価用 LLM によって付与されたラベルと正解ラベルの比較例

d'_j	l_{ij}	c'_j
絵とかストーリーが凝ってて～	1	1
書類提出の締め切りは～	0	0
スーパーの卵 1 パック～	1	0

4.2 作成用 LLM

作成用 LLM は文章とそれに付与された正解ラベル，指示文およびその指示文の評価を入力として，新しい指示文を出力する役割をもつ LLM である．文章とその正解ラベルだけを使用して指示文を作成しようとした場合，指示文に書く内容やその書き方をランダムに決定することになり，作成される指示文の明瞭さもランダムになるため，作成用 LLM が何度指示文を作成しても明瞭な指示文を作成できない可能性がある．作成用 LLM に対して文章とその正解ラベルだけではなく，指示文とその評価を与えることによって，作成用 LLM はそれらの指示文が持つ特徴を参照して新しい指示文を作成する．

作成用 LLM には，図 4.5 に示すプロンプトを与える．図における正答率は Accuracy のことを指す．[true labeled text] は作成用文章とその文章に付与された正解ラベルを，[feedback] は評価用 LLM によって評価された指示文をそれぞれ表す．具体例として，4.1 節にて評価が付与された指示文をフィードバックとして使用する際のプロンプトを図 4.6 に示す．

文章にラベルの番号だけを付与している理由は，フィードバックされた指示文ごとに各ラベルの説明が異なるためである．作成用 LLM は与えられた $D, C, s^*, e^*, s_1, e_1, \dots, s_{i-1}, e_{i-1}$ をもとに，新しい指示文 s_i を作成する．

以下はあるタスク指示文に基づいてラベルを付与された文章です。波括弧の中に、文章とそれに付与されたラベルが書かれています。この結果からどのようなタスクに基づいてラベルが付与されたのか予測し、それをタスク指示文として生成してください。このラベルを付与した人がどのようなラベルを付与する傾向があるかを考察し、それをタスク指示文に反映させてください。指示文は200字以上400字以内で生成してください。タスク指示文だけを生成し、返答に太字を使わないでください。

[true labeled text]

追加の情報として、今まで作成されたタスク指示文とその正答率を以下に示します。

[feedback]

また、これらのタスク指示文とまったく同じタスク指示文を生成しないでください。

図 4.5: 作成用 LLM に与えるプロンプト

以下はあるタスク指示文に基づいてラベルを付与された文章です。波括弧の中に、文章とそれに付与されたラベルが書かれています。この結果からどのようなタスクに基づいてラベルが付与されたのか予測し、それをタスク指示文として生成してください。このラベルを付与した人がどのようなラベルを付与する傾向があるかを考察し、それをタスク指示文に反映させてください。指示文は200字以上400字以内で生成してくださいタスク指示文だけを生成し、返答に太字を使わないでください。

{ 今朝は意外と冷え込んでいてびっくりした, ラベル:0} { 新型スマホ発売来月か~, ラベル:1} …{ 12/30 から 1/3 まで休業します。 , ラベル:1}

追加の情報として、今まで作成されたタスク指示文とその正答率を以下に示します。

{ 文章が主観的であれば 0 のラベルを, 客観的であれば 1 のラベルをそれぞれ付与してください. :60%}

また、これらのタスク指示文とまったく同じタスク指示文を生成しないでください。

図 4.6: 作成用 LLM に与えるプロンプトの具体例

第 5 章 評価実験

実験 1 の目的は提案手法を用いることによって、LLM にとって明瞭な指示文を作成することができるのか確かめることである。実験 2 の目的は LLM にとって明瞭な指示文が、ワーカにとっても明瞭であり、ワーカの Accuracy の向上に貢献するのか確かめることである。

5.1 実験 1(提案手法による指示文の作成)

5.1 節では、作成用 LLM と評価用 LLM を用いて新しい指示文を作成する。また、こうして作成された指示文が s^* と比べて明瞭であると言えるか確認する。

5.1.1 実験手順

依頼者はタスクを用意し、そのタスクをワーカに依頼するための指示文を作成する。依頼者はワーカに作業させる文章を用意し、それらの文章に正解ラベルを付与する。正解ラベルが付与された文章全体のうち、40% の文章を評価用 LLM が指示文を評価するために、残り 60% の文章を作成用 LLM が指示文を作成するために使用する。正解ラベルが付与された文章を指示文の作成用と評価用に分割することで、作成用 LLM が自身に与えられた正解ラベルだけに沿った指示文を作成することを防ぐ。依頼者は提案手法にて評価と作成を繰り返す回数を設定し、提案手法を適用する。

5.1.2 使用データ

この実験では、第一著者を依頼者とした。文章は、SNS 上の投稿のうち画像や動画が含まれていない投稿からランダムに 100 個収集したものを使用した。画像や動画が含まれている投稿を避けた理由は、その投稿の文章を分類するにあたりそれらのコンテンツの内容を確認する必要があるものが多く、コンテンツに付与された文章のみを参照してその文章へのラベル付けを行うことが困難であると判断したため

文章を書いた人の感情を推測し，0: ニュートラル，1: ポジティブ，2: ネガティブのうちいずれかのラベルを付与してください。

図 5.1: 依頼者によって作成された指示文

である。5.1.1 項にて示した通り，40 個の文章を評価用 LLM が，60 個の LLM を作成用 LLM が使用する。

5.1.3 実験条件

この実験では，感情分類タスクをワーカに依頼するための指示文を作成することを目指した。感情分類タスクは，文章を書いた人物の感情がニュートラル，ポジティブ，ネガティブのいずれに当てはまるものであるかを判断するタスクである。文章にラベルを付与する際には，ニュートラルを 0，ポジティブを 1，ネガティブを 2 として，番号のみを文章に付与した。このタスクをワーカに依頼するにあたって依頼者役が作成した指示文を図 5.1 にて示す。

評価用 LLM と作成用 LLM には Gemini 1.5 Flash-8B*[10] を使用した。

提案手法にて作成用 LLM が指示文を作成する回数は 20 回とした。回数を 20 回とした理由は，20 回作成させた指示文のうち一番評価が高い指示文と，21 回以上作成させた指示文のうち一番評価が高い指示文で，評価用 LLM による評価がほぼ変わらなかったためである。

5.1.4 実験結果と考察

評価用 LLM が指示文を評価した結果を表 5.1 に示す。Accuracy が 0/40 となっている指示文は，評価用 LLM から出力された文字列がプロンプトの指示を無視した方法によって区切られていた結果，Accuracy を求めることができなかったものである。この表における s^* および s_{10}, s_{11}, s_{18} より，作成用 LLM は s^* よりも高

*<https://ai.google.dev/gemini-api/docs/models/gemini?hl=ja#gemini-1.5-flash-8b>

い評価を持つ指示文を作成可能なことが分かる。Accuracy を求めることができた指示文のうち、 s_1 から s_{10} までの前半に作成された指示文よりも、 s_{11} から s_{20} までの後半に作成された指示文の方が全体的に Accuracy が高い。Accuracy の違いは、作成用 LLM が指示文を作成するに従って、受け取るフィードバックの量が増加する分評価の高い指示文を作りやすくなったことが理由であると考えられる。

作成用 LLM によって作成された指示文の中で一番評価の高い s_{10} と作成用 LLM ではなく依頼者によって作成された s^* に対して、「指示文の違いは Accuracy の違いに影響していない」という帰無仮説のもと、有意水準を 0.05 として独立性のカイ 2 乗検定を行った。検定の結果 p 値は 0.25 であり、帰無仮説を棄却することはできなかったことから、提案手法によって作成された指示文が評価用 LLM の Accuracy に有意な影響を与えて居ないという結果となった。この検定結果は評価用 LLM が指示文の内容に関係なくタスクを遂行している可能性を示唆しているものの、前半に作成された指示文と後半に作成された指示文の間で評価に違いが見られることから、その可能性には疑問が残る。また、 s_6 や s_{17} など適切に評価されなかった指示文が含まれるため、前半に作成された指示文と後半に作成された指示文の 2 群に対して検定を行うことができない。従って、評価用 LLM に適切な指示文の評価を行わせることによって、評価用 LLM が指示文の内容に応じてタスクを遂行した結果が変化するか確かめる必要があると考えた。

表 5.1 より、一番高い評価を持つ s_{10} を提案手法によって作成された指示文とす

表 5.1: 評価用 LLM による指示文の評価

指示文	評価	指示文	評価	指示文	評価
s^*	13	s_7	4	s_{14}	12
s_1	6	s_8	13	s_{15}	4
s_2	13	s_9	12	s_{16}	13
s_3	2	s_{10}	18	s_{17}	0
s_4	9	s_{11}	15	s_{18}	16
s_5	6	s_{12}	12	s_{19}	0
s_6	0	s_{13}	4	s_{20}	13

る。 s_{10} の内容を図 5.2 に示す。提案手法によって作成された s_{10} は、依頼者が文章の感情を推測するときの偏見が含まれていると考えられる。例えば「2. 文脈」の部分では依頼者が文章全体の感情を参照していることが、「4. 情報の種類」の部分では依頼者が事実の書かれた文章であっても感情を見出す場合があることが示されている。加えてこれらの項目には一つの項目につき一つの例が添付されており、これらの例は LLM やワーカがタスクの内容を把握する手助けをすると考えられる。これらの項目の列挙は指示文ごとに若干内容に違いが見られるものの、指示文 4 以降のほとんどの指示文に含まれており、1 項目 1 個の例は s_{10} 以降の指示文のうち半分程度に含まれていた。

また、 s_{10} は他の指示文と比べて指示文の明瞭さを損なう表現が少ないと考えられる。例えば s_3 には「入力テキストに含まれる特定のキーワードやトピック（例えば、「熱中症」「選挙」「経済危機」など）は、ラベルの判断に役立てることができます。」という説明が含まれている。この説明は熱中症などの言葉をラベルの判断に役立てよう推奨しているものの、具体的にどの言葉が含まれていればどのラベルを付与すべきなのか、という情報が不足しているため、却って LLM やワーカを混乱させ得ると考えられる。

以上より s_{10} には依頼者の意図が含まれており、明瞭さを損なう表現が少ないと考えられる。従って、作成用 LLM によって依頼者の偏見を反映した明瞭な指示文を作成することは可能であると考えられる。

加えて、本実験では作成用 LLM はプロンプトにて指示されていない想定外の挙動をする場合があることが分かった。例えば s_{10} には想定外の挙動が三つ含まれている。一つ目は 1 行目にて、ラベル付けの際に理由を記述するという指示を勝手に追加している点である。二つ目はプロンプトにて太字を使わないよう指示したにも関わらず、マークダウン方式を用いた太字を使用している点である。三つ目はプロンプトにて 200 字以上 400 字以内の指示文を作成するよう指示したにも関わらず、 s_{10} はその制限を大きく超えた 1104 文字によって構成されている点である。これらのプロンプトの指示を無視した挙動は指示文に良い作用と悪い作用を与える可能性がある。理由を記述させる指示の追加はワーカに対してどのラベルを付与すべきか熟考させる作用と、どこに理由を記述すれば良いのか混乱させる作用が考えられる。文字数制限の無視は指示文の明瞭さを向上させる作用と、ワーカが指示文の

以下のテキストを、0（中立）、1（肯定的）、2（否定的）のいずれかの感情ラベルに分類し、その理由を簡潔に記述してください。テキストは、ソーシャルメディア投稿、ニュース記事、ブログ記事、日常会話、その他多様な形式で表現される可能性があります。

感情ラベルの分類においては、以下の点を考慮してください。

1. ****感情表現の有無と強度:**** 喜び、悲しみ、怒り、不安、不満、驚きなど、様々な感情表現を正確に認識し、その強弱を判断します。例えば、「今日はとても良い一日だった」は、「今日は良い一日だった」よりも肯定的な感情が強いと判断します。皮肉な表現や、反語的な表現も考慮します。
2. ****文脈:**** 文章全体の内容、前後の文、状況などを考慮し、言葉の意味合いを判断します。例えば、「雨が降っているけど、傘を持ってきたから大丈夫」という文章は、一見ネガティブな情報を含みますが、全体としてポジティブな解決策が含まれているため、肯定的（1）と判断する可能性があります。
3. ****潜在的な感情:**** 文章に直接的な感情表現がなくても、内容や状況から推測される潜在的な感情も考慮します。例えば、「明日は重要な会議がある」という文章は、直接的な感情表現はありませんが、潜在的な緊張や不安を表している可能性が高いです。

図 5.2: 提案手法によって作成された指示文 (1/2)

内容を読み飛ばしてしまう作用が考えられる。

また、評価用 LLM の出力が崩れていたことによって、20 個中 3 個の指示文を適切に評価することができなかった。これらの 0/40 と評価された指示文は作成用 LLM が参照するフィードバックのノイズとなると考えられる。以下に示す二つの方法によって、このノイズを除去できると考えられる。一つ目は、不適切な評価を受けた指示文をもう一度評価用 LLM に評価させる方法である。この方法は評価用 LLM が何度も不適切な出力をした場合に時間がかかってしまうものの、単純に実装することができる。二つ目は、評価用 LLM による返答を json 形式に固定する方法である。今回の実験では csv ファイルの内容を文字列に変換して入力し、出力さ

4. **情報の種類:** 事実、意見、感想、希望などを区別し、それらの情報に基づく感情的なニュアンスを捉えます。例えば、ニュース記事は事実を伝えつつも、その事実がもたらす感情的な反応（市場の混乱、市民の不安など）を考慮する必要があります。

5. **表現様式:** ユーモア、嘆き、命令、自己主張、疑問、質問など、様々な表現様式を考慮し、文章の感情的なトーンを判断します。例えば、皮肉な表現は、一見肯定的な文章であっても否定的（2）と判断される場合があります。ラベル付けを行う際の判断基準は、上記5項目を総合的に考慮し、文章全体から読み取れる感情のニュアンスを適切に判断することです。複数の感情が混在している場合は、最も強い感情や、文章全体に与える影響の大きい感情に基づいてラベル付けします。分類が困難な場合は、最も妥当と思われるラベルを選択してください。

(例)

「今日は素晴らしい天気だ！」 ラベル：1 理由：肯定的な表現「素晴らしい」を用いているため。

「今日の試合は残念だった。」 ラベル：2 理由：残念という否定的な感情を表しているため。

「今日のニュースは経済に関する内容が多かった。」 ラベル：0 理由：感情や意見が明確でないため、中立的な情報と判断する。

図 5.2: 提案手法によって作成された指示文 (2/2)

れた文字列を csv ファイルに変換した。csv ファイルの代わりに json ファイルを用いることによって、評価用 LLM が不適切な出力をする確率が減り、指示文に適切な評価が付与されやすくなると考えられる。

5.2 実験 2(提案手法によって作成された指示文の評価)

5.2 節では, 5.1 節にて作成された指示文を用いてワーカにタスクを遂行させることによってその指示文がワーカにとって明瞭なものであるか検証する.

5.2.1 実験手順

指示文を評価をするため, 指示文を用いてタスクを遂行するようワーカに依頼する. この実験において各ワーカでタスクを遂行する能力はそれぞれ同じと仮定する. ワーカをグループ A と B に分け, ワーカに作業させる文章とグループごとに異なる指示文を与える. グループ A は s^* , グループ B は s_{10} を用いて文章にラベルを付与する. ワーカごとに Accuracy を計算する. グループごとにラベルの多数決を行い, 結果をそのグループが付与したラベルとする. ある文章において最多となるラベルが複数となった場合, その文章には正解ラベルを付与できなかったものとして扱う. 例えばグループ A のワーカによって d_j にラベル 0 が 3 個, ラベル 1 が 1 個, ラベル 2 が 3 個付与されていた場合, ラベル 0 と 2 の二つのラベルが最多であるためグループ A は正解ラベルを付与できなかったものとして扱う.

ワーカごとの Accuracy の比較, およびグループごとの Accuracy の比較によって, 提案手法によって作成された指示文が Accuracy に影響しているか確認する.

5.2.2 使用データ

ワーカには 5.1.2 節にて収集したのと同じ 100 個の文章を分類させた. グループ A は図 5.1 に示した s^* を, グループ B は図 5.2 に示した s_{10} を用いてタスクを遂行した.

5.2.3 実験条件

被験者 14 人をワーカとして, 7 人ずつグループ A と B に分けた. グループ A には図 5.1, グループ B には図 5.2 にて示す指示文 s^* , s_{10} をそれぞれ与えた. ワーカにはラベルを付与するにあたって以下の行為を行わないように制限した.

- ワーカー間での文章に付与するラベルに関する相談
- 自分が所属していない方のグループに与えられた指示文の確認
- 依頼者に対する、指示文の内容に関わる質問

これらの制限によって、指示文以外の要素が Accuracy に影響しないようにした。

5.2.4 実験結果と考察

各ワーカーの Accuracy とグループごとの平均 Accuracy を計算した結果を表 5.2 に示す。この表においてワーカー id に A が含まれるワーカーはグループ A、B が含まれるワーカーはグループ B 所属である。この表より、グループ B に所属するワーカーの方がグループ A に所属するワーカーと比べて高い Accuracy を記録していることが確認できた。一方「グループ間でワーカーの Accuracy に違いはない」という帰無仮説に対して、有意水準を 0.05 として対応のない t 検定を行った結果、 p 値は 0.29 であったため、帰無仮説は棄却されなかった。従って、グループ間におけるワーカーの Accuracy の差は有意なものであるとは言えなかった。

各グループで多数決の結果選ばれたラベルの正答数と誤答数、Accuracy を表 5.3 に示す。この表より、グループ B のほうがグループ A よりも Accuracy が高いことが確認できた。一方「グループの違いと正答数及び誤答数の違いは独立である」

表 5.2: ワーカーごとの Accuracy とグループごとの平均 Accuracy

ワーカー id	Accuracy	ワーカー id	Accuracy
A-1	0.68	B-1	0.63
A-2	0.62	B-1	0.63
A-3	0.64	B-3	0.64
A-4	0.63	B-4	0.64
A-5	0.67	B-5	0.65
A-6	0.56	B-6	0.70
A-7	0.60	B-7	0.73
グループ A 平均	0.63	グループ B 平均	0.66

という帰無仮説に対して、有意水準を 0.05 として独立性のカイ 2 乗検定を行った結果、 p 値は 0.76 であったため、帰無仮説は棄却されなかった。従って、グループの Accuracy の差は有意なものであるとは言えなかった。

ワーカが付与したラベルの Accuracy とグループが付与したラベルの Accuracy との両方において、グループ B の方がグループ A よりも Accuracy が高かったことから、提案手法によって作成された指示文 s_{10} が Accuracy の向上に影響していると考察した。この影響が検定によって有意なものとは見られなかった理由は二つ考えられる。一つ目は、5.1 節にて作成した指示文 s_{10} に依頼者の偏見がうまく反映できていなかったことが理由であると考えた。この場合、指示文作成の手法を改善することによってグループごとの差が顕著に現れることが期待できる。二つ目は、ワーカがタスクを遂行した文章の個数が少なかったことである。今回使用した文章のうち、指示文の明瞭さによって付与されるラベルが変わり得る文章の割合は少なかったため、指示文の明瞭さによる Accuracy への影響が誤差の範疇に収まってしまったと考えた。この場合、ワーカがより多くの文章に対してタスクを遂行することによって、指示文の明瞭さの Accuracy に与える影響が誤差と言えない程度に大きくなると考えた。

表 5.3: 各グループが付与したラベルの正誤答数と Accuracy

	正答数	誤答数	Accuracy
グループ A	69	31	0.69
グループ B	72	28	0.72

第6章 おわりに

本研究は、クラウドソーシングを通じたデータセット作成の際に明瞭な指示文をワーカーに提供することによって、依頼者にとって望ましいデータセットが作成されることを目指した。明瞭な指示文とは、ワーカーがタスクを遂行するにあたって必要な情報が含まれた指示文のことである。

依頼者は自身によって作成された指示文が明瞭であるか判断することは難しい。その理由は、依頼者がタスクに対して当然こうするべきという偏見を持っているためである。

我々は LLM とワーカーで指示文を用いてタスクを遂行する能力は等価であると仮定し、LLM にとって明瞭な指示文を作成することによって、依頼者が自身にとって望ましいデータセットが得られることを目指した。目標の達成のために GAN を参考として、指示文を作成する作成用 LLM と指示文を評価する評価用 LLM を繰り返し用いる手法を考えた。

提案手法では、評価用 LLM はラベルが付与されていない文章に指示文を用いてラベル付けを行い、その Accuracy を指示文の評価として与える。作成用 LLM は正解ラベルが付与された文章、および指示文とその評価を用いて新しい指示文を作成する。

実験では、グループ A のワーカーは依頼者によって作成された指示文を、グループ B のワーカーは提案手法によって作成された指示文をそれぞれ用いてラベル付けを行い、グループ間の Accuracy の違いを確認した。実験の結果、グループ B の方が Accuracy が高かったことから、提案手法によって明瞭な指示文を作成できることが確認できた。

今後の展望として、手法に対する二つの改善案を考えている。一つ目は指示文の評価の際に複数の評価用 LLM を利用するものである。複数の評価用 LLM の意見をまとめることによって、偶然 LLM が付与するラベルを間違えても Accuracy に影響を与えにくくなり、指示文の明瞭さが Accuracy に与える影響が観測しやすくなると考えた。また、複数の評価用 LLM が同じ指示文を用いて同じ文章にバラバラなラベルを付与した場合、指示文の内容が評価用 LLM によって一義に解釈されていないと見なすことができる。従って、文書に付与されたラベルのばらつき具

合を作成用 LLM に与えることによって、作成用 LLM は一義的に解釈できる明瞭な指示文を作成できると考えた。二つ目は Gemini 以外の LLM を利用するものである。具体的な利用方法として、評価用 LLM と作成用 LLM 両方を Gemini 以外の LLM に置換した場合、指示文の評価と作成をそれぞれ異なる種類の LLM に実行させた場合、指示文の評価を複数種類の LLM に実行させる場合を考えている。Gemini 以外の LLM を用いることによって、指示文を作成することを得意とする LLM の発見や、異なる LLM が一義に解釈できる指示文の作成といった効果が期待できる。

謝辞

本研究を進めるにあたって多くの方々のお世話になりました。指導教員である鈴木先生には、研究に行き詰ったときに何度も対面で対応していただき、次に何をすべきかという方針について多くの知見をいただきました。また、積極的な対外発表を促進していただいたおかげで、より多くの方々から研究に関するアドバイスをもらう機会が得られました。本当にありがとうございました。あと2年間よろしく願いいたします。

事務補佐員の井尾さんには、書類提出の際に何度もお世話になりました。そして、書類の内容に含まれるミスを見つける、さながらウォーリーを探せのような作業を何度もさせてしまい申し訳ありません。可能な限りミスを無くすので、あと2年間よろしく願いいたします。

研究室の先輩方には研究に関しても、それ以外のことに関してもお世話になりました。研究に関して、研究が決まり切っていないときはテーマのアイデア出し、ある程度決まってからはその方針や詳細など、自分の行った研究全体を通して多くのタイミングで助けていただきました。また、研究以外でも食事に誘っていただいたり、雑談をしたりなどしていただいたおかげで、モチベーションを維持することができました。

研究室の同期にも、様々なことでお世話になりました。同期の中で自分だけが把握していない情報があったり、院試の書類を出すときには本当にお世話になりました。正直院試の書類の時点でこれならESとか就職関連の書類とかどうなるんだろうという感じですが、できる限りミスらないよう頑張ります。

また、研究室に所属する学生の方々全員に対して、実験に何度も付き合っていたことを感謝します。時には「文章に付与されたラベルだけから指示文を作成して」という七面倒なことを任せたり、「100件の文章をExcelで読み込んでラベルを理由とともに付与してくれ」という時間がかかるタスクを卒論の時期に依頼しました。実験を手伝っていただきありがとうございました。

以上に挙げた方々以外にも研究を進めるにあたってお世話になった方々は多くいます。全てのお世話になった方々に改めてお礼申し上げます。本当にありがとうございました。そしてあと2年、今後ともよろしく願いいたします。

参考文献

- [1] Jeff Howe, et al. The rise of crowdsourcing. *Wired magazine*, Vol. 14, No. 6, pp. 176–183, 2006.
- [2] Dana Chandler and Adam Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, Vol. 90, pp. 123–133, 2013.
- [3] Thomas Gillier, Cédric Chaffois, Mustapha Belkhouja, Yannig Roth, and Barry L Bayus. The effects of task instructions in crowdsourcing innovative ideas. *Technological Forecasting and Social Change*, Vol. 134, pp. 35–44, 2018.
- [4] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.
- [5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, Vol. 27, , 2014.
- [6] VK Chaithanya Manam, Joseph Divyan Thomas, and Alexander J Quinn. Tasklint: Automated detection of ambiguities in task instructions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10, pp. 160–172, 2022.
- [7] Zahra Nouri, Ujwal Gadiraju, Gregor Engels, and Henning Wachsmuth. What is unclear? computational assessment of task clarity in crowdsourcing. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pp. 165–175, 2021.
- [8] Zahra Nouri, Nikhil Prakash, Ujwal Gadiraju, and Henning Wachsmuth. Supporting requesters in writing clear crowdsourcing task descriptions through computational flaw assessment. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 737–749, 2023.

- [9] Santosh Kumar Radha, Yasamin Nouri Jelyani, Ara Ghukasyan, and Oktay Goktas. Iteration of thought: Leveraging inner dialogue for autonomous large language model reasoning, 2024.
- [10] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

発表リスト

[1] 田中涼, 鈴木優 『客観的な視点を利用したインストラクション作成手法』 東海
関西データベースワークショップ 2024, 2024.