Vol.2025-HCI-212 No.42
2025/3/7

人-AI協調アノテーションの有用性検証とAI予測の提示タイミングが人のラベル決定に及ぼす影響

木下 裕一朗[1,a)]　中村 聡史[1]

概要：AIは様々なタスクに活用されており，データに対するアノテーションへの利用可能性も示されている．本稿では，人とAIの協調アノテーションにより，スポーツのネタバレ画像データセットの質が向上するか検証する．また，AIが予測したラベルとその理由を，人がアノテーションする前に提示する場合と，人がアノテーションした後に提示する場合で，ラベル決定に異なる影響を及ぼすか明らかにする．アノテーション実験の結果，AIによる予測が正しい場合は，AIなしでアノテーションするときよりも正確なラベルが付与され，AIによる予測が誤っている場合は，AIなしでアノテーションするときよりも誤ったラベルが付与されることがわかった．また，AIの予測を提示するタイミングによって，ラベルの正確性や一致度，アノテーション作業の負荷の大きさが変わる可能性が示された．

1. はじめに

AIの発展によって，文章の翻訳や要約，画像認識などを高精度に行うことが可能となり，現在AIは様々なタスクで活用されている．高性能なAIを開発し，その性能を正しく評価するためには，使用するデータセットの質が極めて重要となるが，データセットの構築にはコストがかかる．そこで，データアノテーションにおけるAIの利用可能性について研究が行われており，アノテーション支援にAIが有効である [1][2][3] ことや，AIをアノテータとして使用できる可能性 [4][5] が示されている．また，AIをアノテーションに用いることで，コストを削減できる [6][7][8] ことが明らかになっている．

我々はこれまで，スポーツの試合結果が予想できてしまう画像（ネタバレ画像と呼ぶ）の存在に着目し，画像によるスポーツのネタバレ防止に取り組んできた．我々はまず，ネタバレ画像の判定可能性について検証を行うため，スポーツのネタバレ画像データセットを構築し，ネタバレ画像検出手法の提案とその精度評価を行った [9]．精度評価の結果，提案手法が約80%の精度でネタバレ画像を検出できることを示した．しかし，構築したデータセットにおいて一部のラベルが不正確であったため，精度を正しく評価できていなかった可能性がある．そこで我々は，ネタバレ画像データセットの質の改善が必要であると考えた．

既存のネタバレ画像データセットの改善を行う場合は，ラベルが不正確であったデータを抽出し，そのデータに対して再アノテーションを実施すれば良い．しかし，データセットの拡張のために新しいデータに対してアノテーションを行う場合，人間の手動アノテーションによって一部のラベルの正確性に再び問題が発生する可能性がある．そのため，ネタバレ画像アノテーションにおいて上記の問題を解決できるアノテーション手法が求められる．

Wangら [10] は，テキストアノテーションタスクにおける人-AI協調アノテーションの有用性を検証しており，AIによる予測ラベルとその理由を提示することで，予測ラベルが正しい場合は人のアノテーションの正確性が向上することを明らかにした．画像アノテーションにおける人-AI協調アノテーションの有用性については十分に研究されておらず，AIによる予測の適切な提示タイミングについて検証を行った研究は多くない．

そこで本稿では，アノテーションタスクにおけるAIの利用可能性に着目し，人とAIが協調してアノテーションすることがネタバレ画像データセットの質向上に有用であるか検証を行う．具体的には，ネタバレ画像データセットの一部のデータに対して，AIを使用せずに人間がアノテーションしたときの結果と，AIによる予測結果を提示して人間がアノテーションしたときの結果を比較することで，人-AI協調アノテーションの有用性を検証する．また，Wangら [10] の研究では，AIによる予測結果を人間がアノテーションする前に提示していたが，本稿では，人間がアノテーションする前に予測結果を提示するだけでなく，

[1] 明治大学
Meiji University
[a)] kinoshita@nkmr-lab.org

人間がアノテーションした後に予測結果を提示する場合も検証することで，AI の予測結果の提示タイミングによって人間のラベル決定に及ぼす影響は異なるかを調査する．

2. 関連研究

2.1 AI のアノテーションへの利用

機械学習モデルや大規模言語モデル（LLM）は，アノテーションの支援にも活用されている．機械学習モデルの利用には，タスクに特化したモデルのトレーニングが必要となるのに対し，LLM はプロンプトを調整するだけで様々なタスクに利用できるため，LLM のアノテーションへの利用可能性について研究が盛んに行われている．

LLM をアノテータとして利用できることを示す研究 [4][5] が存在するが，その性能はタスクによって異なり [11][12]，LLM が人間のアノテータにとって代わることはできないことを示す研究も多い [13][14][15]．また，LLM を単独でアノテーションに用いると，誤った結果を導いてしまうリスクが指摘されている [16]．そこで，LLM によって得られた結果を人間がチェックする方法 [17] や，人と LLM が協調してアノテーションする方法 [10] が提案されており，それらはアノテーションに要するコストを削減しつつ，データ品質を高められる可能性が示されている．

本稿は画像アノテーションタスクを行うため，テキストだけでなく画像も扱うことができるマルチモーダル LLM（MLLM）を利用する．MLLM は画像に写る物体間の関係理解が可能 [18] で，画像分類の精度向上に有用であることが示されている [19] ため，ネタバレ画像アノテーションに十分活用できると考えられる．本稿は，人-AI の協調アノテーションがネタバレ画像データセットの品質向上に有用であるか検証を行うものである．

2.2 AI が人間の意思決定に及ぼす影響

AI が生成した文章は人間の論理的思考に影響を与え，意思決定を変化させることが明らかになっている [20]．機械学習モデルによる助言が人間の意思決定に与える影響について調査した研究 [21] では，人間が判断した後に助言を提示することで，人間の判断力を改善できることが示されている．ただし，提示された助言を過剰に信頼してしまう場合や，反対に助言を活用しない場合も観察されており，人間が AI による助言を適切に活用することは難しい可能性が示唆されている．また，自然言語処理能力が高い LLM による説明は，人間による説明よりも明確であると認識される傾向が示されており [22]，LLM の判断が誤っている場合は，人間を誤解させるリスクが指摘されている [10]．

本稿は，AI の利用がネタバレ画像データセットの質向上に有用であるかの検証に加え，アノテーション時における AI の予測結果の提示タイミングの違いによって，人間のラベル決定やその正確性に及ぼす影響が異なるか明らかにするものである．

3. 実験

本稿は，以下の 3 手法を用いてスポーツのネタバレ画像データセット [9] の一部のデータに対する再アノテーションを行う実験を実施し，その結果を比較することで，人-AI 協調アノテーションのデータセットの質改善に対する有用性と，AI による予測の提示タイミングの違いが人のラベル決定に異なる影響を及ぼすか検証する．

- AI 予測なし手法：人間が自身だけで考えてアノテーションする
- AI 予測先出し手法：AI による予測ラベルとその理由が表示された状態で，人間がアノテーションする
- AI 予測後出し手法：人間がアノテーションした後に，AI による予測ラベルとその理由が表示される

3.1 使用データと AI による予測ラベルの取得

スポーツのネタバレ画像データセットは，野球，サッカー，バスケットボールの画像約 4,500 枚 (各スポーツ約 1,500 枚ずつ) から構成されており，各画像には 3 名のアノテータによるラベルが付与されている．ラベルは，画像からその試合結果が「明らかにわかる」「なんとなく予想がつく」「わからない」の 3 つである．過去の研究 [9] では，「明らかにわかる」「なんとなく予想がつく」「わからない」の 3 つのラベルにそれぞれ 0，1，2 のスコアを割り当て，3 名のアノテータによるラベルの統合を行った結果，合計スコアが 2 以上の場合にネタバレ画像，2 未満の場合を非ネタバレ画像と決定した．

ネタバレ画像データセットには，試合結果が予想できる画像であるにも関わらず，ラベル統合の結果，非ネタバレと決定された画像（**図 1**）がいくつかみられた[*1]．また，アノテータによってラベルが異なる画像も存在した．そこで本稿では，そのような画像を対象として人-AI 協調アノテーションを行い，ラベルの正確性とラベル一致度が向上するか検証を行う．なお，本稿では，データセットのうちサッカーの画像のみを実験に用いる．

本稿では，筆頭著者によってラベルが正確でない（試合結果を予想できるが非ネタバレと決定されていた）と判断された画像 52 枚と，ラベルがアノテータ間で分かれていた画像 190 枚（ネタバレ画像 17 枚，非ネタバレ画像 173 枚）に加え，ラベルが正確である 182 枚のネタバレ画像と 26 枚の非ネタバレ画像を合わせた，計 450 枚の画像を実験に使用する．

我々は，OpenAI の GPT-4o [23] を使用し，**表 1** に示すプロンプトを用いてこの 450 枚の画像に対する AI の予測ラベルとその理由を取得した．予測ラベルの一貫性を高め

[*1] 試合結果の予想ができない画像が，ネタバレ画像と決定されていた場合は存在しなかった．

図 1: ラベルが誤っていたと考えられる画像の例．選手たちが喜んでいる様子から，試合結果の予想が可能であるにも関わらず，非ネタバレ画像と決定されていた．

表 1: 画像に対するラベルの予測とその理由を取得するために GPT-4o に入力したプロンプト

Instruction: I will provide you with a football image. Please annotate the image. There are three labels to select from: ”明らかにわかる（Clearly identifiable）”, ”予想できる（Predictable）”, and ”わからない（Not identifiable）”. Select the one label that best applies to the image. Additionally, provide a concise explanation for your annotation in Japanese. # Criteria: If the image explicitly contains words related to the outcome of the match or displays the final score, it should be annotated as ”明らかにわかる（Clearly identifiable）”. If the image features players smiling, posing, or appearing sad, and the result of the match can be predicted from this, it should be annotated as ”予想できる（Predictable）”. If the image does not contain any players, or if players are present but none of the above characteristics are observed, it should be annotated as ”わからない（Not identifiable）”. # Output Format: Please provide the label and reason for each image in the following format. Note that both the label and reason should be output in Japanese: label: reason

るため，GPT の temperature パラメータ（0 から 2 の値を指定可能）の値は 0.2 に設定した．

3.2 実験設計

本実験は，Wang ら [10] の実験設計を参考にし，各実験参加者が 3 手法すべてを用いてアノテーションを行う参加者内計画で実施する．

実験参加者が各アノテーション手法で同数の画像にラベル付けを行うようにするため，我々は実験に使用する 450 枚の画像を，事前に 150 枚ずつ 3 つのグループに分割した．各グループに含まれる，過去のアノテーションでラベルが正確でないと判断された画像・過去のアノテーションでラベルがアノテータによって異なっていた画像・ラベルが正確なネタバレ画像/非ネタバレ画像の数については，それぞれグループ間でできるだけ差が出ないようにした．実験参加者は，我々が分割した 150 枚の画像グループごとに，異なる手法を用いてアノテーションを行う．なお，本実験ではアノテーション手法が 3 つあるため，手法の経験順序は 6 通り存在する．順序効果を考慮し，我々は 6 通りの経験順序それぞれについて，その順序でアノテーションを行う実験参加者の数がすべて同数になるようにした．

本実験では，AI による予測ラベルとその理由の提示によって実験参加者のラベル判断基準がアノテーション中に変化するか調べるため，各手法で 150 枚ずつアノテーションする際に，最初に 15 枚の画像に対してアノテーションを行ってもらい，150 枚の画像へのアノテーション後，再び最初に提示した画像と同じ 15 枚の画像に対してアノテーションを実施してもらう．なお，実験参加者には，最初と最後に同じ画像を提示することは知らせない．また，この 15 枚の画像は，150 枚の画像の中から，過去にアノテータによってラベルが異なっていたネタバレ画像と非ネタバレ画像を 5 枚ずつ，そしてラベルが正確でないと考えられる画像 5 枚を抽出してシャッフルしたものであり，15 枚の画像は事前に 3 つに分けた画像グループごとに決定した．このため，実験参加者は 150 枚の画像のうち，15 枚の画像については 3 回アノテーションを行い，その他の 135 枚の画像については 1 回のみアノテーションを行う．つまり，実験参加者は，各手法を用いて 180 回ずつアノテーションを行うことになる．

実験参加者は我々が実装した Web システム（**図 2**）を用いてアノテーションを行い，システムでは各画像に対するラベルとラベルの変更回数，手法ごとの合計アノテーション時間，実験後に実施するアンケートへの回答を記録する．

3.3 実験手順

実験参加者が自身の PC を用いて，任意のブラウザから実験用の Web システムにアクセスすると，最初に実験内容の説明ページが表示される．説明ページでは，各画像に対してその画像から試合結果がわかるか，「明らかにわかる」「予想できる*2」「わからない」の 3 つの中から適切なラベルを選択してアノテーションすることを指示した．また，画像 180 枚にアノテーションするごとにアノテーション手法が変わることと，手法が変わる際にはどのような手法かを説明するページが表示されることを記載した．さらに，各ラベルに該当する画像の例を一枚ずつ示し，実験を通してブラウザの戻るボタンやリロードボタンは使用しないように指示した．

実験内容説明ページで実験参加者がアカウントを作成し，ボタンを押すと，最初に使用するアノテーション手法

*2 過去の研究 [9] では，「なんとなく予想がつく」というラベルであったが，本稿では表現を「予想できる」に変更した．

(a) AI 予測なし手法　(b) AI 予測先出し手法　(c) AI 予測後出し手法（ラベルを選択すると，AI による予測結果が表示される）

図 2: 実験に使用したアノテーションシステムの画面

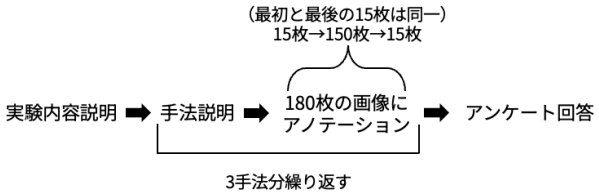

図 3: 実験手順

についての説明ページが表示される．手法の説明ページでは，できるだけ作業を中断せずに続けて 180 枚の画像に対してアノテーションすることを指示した．また，AI 予測先出し/後出し手法の説明ページには，AI による予測が正しいとは限らないことを記載した．実験参加者が手法についての説明ページを読みボタンを押すと，アノテーション画面へと遷移する．

アノテーション画面では，画像は 5 枚ずつ表示される．実験参加者が 5 枚の画像すべてに対してラベルを選択し，送信ボタンを押すと，次の 5 枚が表示される．これを繰り返し，180 枚の画像に対してアノテーションが終わると，次に使用するアノテーション手法についての説明ページに遷移する．このようにして，実験参加者は 3 手法を用いて計 540 枚の画像にアノテーションする．

実験参加者が 540 枚の画像すべてに対してアノテーションすると，アンケート画面に遷移する．アンケート画面では，各アノテーション方法における自身のアノテーションの正確性とアノテーションのやりやすさについて質問した．また，AI を利用した 2 つのアノテーション手法においては，AI が予測した結果の信頼度と AI の予測精度の高さについて質問した．実験参加者は，すべてのアンケート項目に対し，7 段階リッカート尺度を用いて回答する．実験参加者がアンケートに回答した後，システムは実験の終了を参加者に通知する．以上の実験の手順を**図 3** に示す．

4. 結果

4.1 実験参加者とデータの除外

実験参加者は，サッカーの視聴観戦経験あるいは競技経験がある 30 名で，平均年齢は 21.23 歳（標準偏差 2.14）であった．アノテーション実験により，我々は 450 枚の画像に対する 16,200 件 (各実験参加者 540 件ずつ) のラベルを得た．実験に使用した 450 枚の画像のうち，3 回アノテーションする対象であった 45 枚の画像（3.2 節参照）については，3 つの手法ごとにそれぞれ異なる 10 名によって 30 件ずつラベルが付与され，その他の 405 枚の画像については，手法ごとにそれぞれ異なる 10 名によって 10 件ずつラベルが付与された．

実験参加者ごとに，AI 予測なし手法を用いてアノテーションしたときのラベルと，過去のアノテーション [9] で得られたラベルの一致度を求めた結果を**図 4** に示す．この一致度の算出では，過去のアノテーションでラベルが不正確であったと考えられる 52 枚の画像は除外した．また，過去のアノテーションではアノテータが 3 名いたため，3 名の中で最も選択数が多かったラベルを正解ラベルと仮定して一致度を算出した．本実験では，5 名の実験参加者（P2, P5, P17, P22, P25）において過去のアノテーションで得られたラベルとの一致度が 0.5 未満と低い値であった．この 5 名のラベル判断基準は，アノテーションシステム上で最初に説明した基準と大きく異なっていたと考えられるため，該当する 5 名によるデータは分析対象から除外した．本章では，25 名の実験参加者によるデータについて分析を行う．

4.2 ラベル一致度とラベルの正確性

手法ごとにラベルの一致度を求めた結果を**表 2** に示す．本稿は，ラベル一致度の指標に Krippendorff's alpha を用

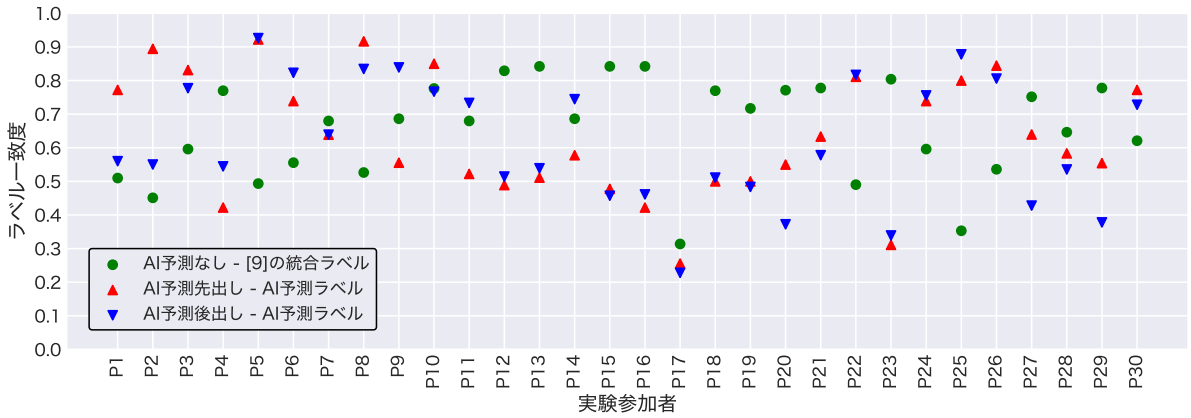

図 4: 各実験参加者における，AI 予測なし手法時の [9] の統合ラベルとの一致度，AI 予測先出し手法時の AI が予測したラベルとの一致度，AI 予測後出し手法時の AI が予測したラベルとの一致度

いた．この指標は −1 から +1 の間の値をとり，値が大きいほど一致度が高いことを表す．本実験では，同一の手法により 540 枚の画像に対してそれぞれ 10 件[*3]ずつラベルが付与されているが，180 枚ごとに各実験参加者が用いるアノテーション手法を変えていたため，各手法において同一の 10 名によってアノテーションされているのは 180 枚の画像のみである．そのため，我々は 180 枚ごとに Krippendorff's alpha を算出し，手法ごとにそのマクロ平均と標準偏差を求めた[*4]．AI 予測先出し手法は，他の手法と比べてラベル一致度が高かったものの，手法間の差は小さかった．一般的には Krippendorff's alpha の値が 0.67 以上であるときに中程度の一致とみなされるため，どの手法においてもラベル一致度は低い結果であった．

次に，手法ごとに各画像に対する 10 件[*3]のラベルを統合し，ネタバレ/非ネタバレの決定を行った． [9] では，試合結果が「明らかにわかる」「なんとなく予想がつく」「わからない」の 3 つのラベルにそれぞれ 0，1，2 のスコアを割り当て，3 名のアノテータによる合計スコアが 2 以上の場合にネタバレ画像，2 未満の場合を非ネタバレ画像とした．本稿では，各ラベルに対応したスコアは同じものを採用するが，アノテータが各画像に対して 10 名存在したため，10 名のアノテータによる合計スコアが 6 以上のものをネタバレ画像，6 未満のものを非ネタバレ画像とする．なお，5 名の実験参加者によるデータを除外したことによってラベル付与数が 9 件となった画像についても，同様の方法でネタバレ/非ネタバレの決定を行った．実験で使用したネタバレ画像と非ネタバレ画像について，本実験でのラベル統合結果が過去と同一になった割合と，過去にラベルが不正確であったことにより非ネタバレとされた画像が，

[*3] 5 名の実験参加者によるデータを分析対象から除外したことにより，一部の画像ではラベルの付与数が 9 件となった．

[*4] Krippendorff's alpha の算出には，Castro により公開されているライブラリ（`https://github.com/pln-fing-udelar/fast-krippendorff`）を使用した．

表 2: 手法ごとに算出した Krippendorff's alpha の値

	AI 予測なし	AI 予測先出し	AI 予測後出し
平均	0.30	0.34	0.29
標準偏差	0.07	0.03	0.13

ネタバレ画像と正しく決定された割合をそれぞれ手法ごとに求めた．結果を**表 3** に示す．

AI 予測なし手法において，ネタバレ画像は高い割合で過去と同一の結果となり，過去のラベル不正確画像の約 5 割が正しくネタバレ画像と決定された．一方で，AI 予測を提示した 2 手法については，非ネタバレ画像において高い割合で過去と同一の結果となったものの，ネタバレ画像をラベル統合によってネタバレと正しく決定できた割合と，ラベル不正確画像をネタバレ画像と正しく決定できた割合は低かった．

Wang ら [10] は，AI による予測ラベルが正しい場合は人間のアノテーション精度が向上するが，予測ラベルが誤っている場合は人間のアノテーション精度が低下することを明らかにした．そこで本稿も，AI によるラベルが正確であった場合とそうでなかった場合に分けてそれぞれラベル統合を行い，その結果が過去 [9] のラベル統合結果と一致した割合を求めた．実験で用いたラベルが不正確であった 52 枚の画像に対して，AI が正しいラベルを予測できた数は少なかったため，AI の予測ラベルが誤っていた場合についてのみ，52 枚のラベル不正確画像が正しくネタバレと決定された割合を求めた．

AI の予測ラベルが正確であるとき，AI 予測先出し/後出し手法の両方が，AI 予測なし手法よりも高い割合で正しくネタバレ/非ネタバレ画像と決定できた（**表 4**）．一方で，AI の予測ラベルが誤っている場合，AI 予測先出し/後出し手法の両方において，ネタバレ/非ネタバレと正しく決定できた割合は，AI の予測ラベルが正確であったときよりも低かった（**表 5**）．

表 3: ラベル統合の結果，正しくネタバレ/非ネタバレ画像と決定された割合と，[9] のラベル不正確画像が，ネタバレ画像と正しく決定された割合

	AI 予測なし	AI 予測先出し	AI 予測後出し
ネタバレ画像	0.83	0.78	0.67
非ネタバレ画像	0.83	0.85	0.89
ラベル不正確画像	0.49	0.30	0.29

表 4: AI の予測ラベルが正確な場合において，正しくネタバレ/非ネタバレ画像と決定された割合

	AI 予測先出し	AI 予測後出し
ネタバレ画像	0.93	0.87
非ネタバレ画像	0.91	0.93

表 5: AI の予測ラベルが誤っている場合において，正しくネタバレ/非ネタバレ画像と決定された割合と，[9] のラベル不正確画像がネタバレ画像と正しく決定された割合

	AI 予測先出し	AI 予測後出し
ネタバレ画像	0.67	0.54
非ネタバレ画像	0.77	0.85
ラベル不正確画像	0.29	0.28

4.3 AI の予測提示がアノテーション行動に及ぼす影響

AI の予測提示によって，アノテーションに要する時間が変化するか調べるため，手法ごとに実験参加者がアノテーションに要した時間の分析を行った．実験参加者が常にアノテーション作業に取り組んでいたわけではない可能性を考慮し，この分析では Rousseeuw ら [24] の基準に基づいて，各実験参加者のアノテーション時間の四分位値から四分位範囲の 1.5 倍を上回っていた，あるいは下回っていた 5 名の参加者によるデータは除外した．各手法における平均アノテーション時間は，AI 予測なし手法が 614.29 秒，AI 予測先出し手法が 690.79 秒，AI 予測後出し手法が 680.94 秒であった（**図 5**）．Shapiro-Wilk 検定によってデータの正規性が棄却されたため（$p < 0.05$），フリードマン検定を実施した結果，手法間のアノテーション時間に有意差は認められなかった（$p = 0.26$）．

次に，AI による予測結果をアノテーション後に提示することによって，実験参加者がどの程度ラベルを変更したか分析を行った．手法ごとに平均ラベル変更回数を求めたが，AI 予測なし手法では 0.08 回，AI 予測先出し手法では 0.07 回，AI 予測後出し手法では 0.07 回という結果であり，手法間で差はなく全体的にラベル変更が行われていなかったことがわかった．

AI による予測結果の提示により，人間のラベル判断基準がアノテーション中に変化するか調べるために，3 回アノテーションする対象であった画像について，同一画像に対する 3 回のアノテーションで 3 回とも同じラベルを選択していた割合を手法ごとに求めた．その結果，AI 予測なし手法では 0.70，AI 予測先出し手法では 0.78，AI 予測後出し手法では 0.75 という割合であったことから，AI による予測を提示することによってアノテーション中にラベルの判断基準が変化するのではなく，反対にラベル判断基準の一貫性が向上する可能性が示された．

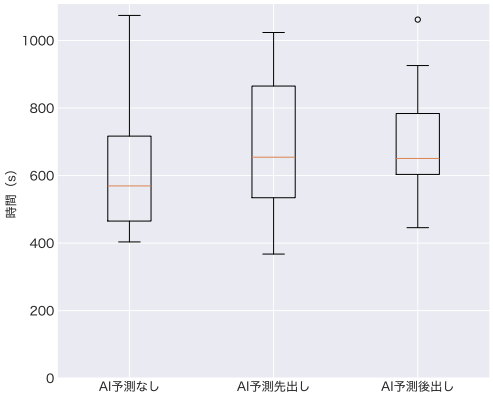

図 5: 手法ごとのアノテーション時間

アノテーション実験の終了後に実施したアンケートに対する回答の平均値を**表 6** に示す．実験参加者は全ての質問に対して 7 段階のリッカート尺度（−3 から +3）を用いて回答し，値が高いほどその質問項目に対する同意を表す．自身のアノテーションの正確性に関する評価は，手法間で差がみられなかった．アノテーションのやりやすさについては，AI 予測なし手法が最も高く，AI 予測先出し手法が最も低い結果となった．AI が予測した結果の信頼度や予測精度の高さについては，AI 予測先出し手法の方が AI 予測後出し手法よりも高い評価であったが，いずれも 0 以下の値であった．

5. 考察

5.1 人-AI 協調アノテーションの有用性

ラベル一致度は AI 予測先出し手法が 3 手法の中で最も高かったものの，手法間で差は小さく，どの手法も低い値であった．本実験では過去のアノテーションでアノテータによってラベルが異なっていた画像を多く使用していたことに加え，実験参加者によって過去のアノテーションの統合ラベルや AI の予測ラベルとの一致度が大きく異なっていた（図 4）ことが一致度の低下につながったと考えられる．

AI によるラベルが正確であるとき，AI の予測を提示した 2 手法によるラベルの正確性は，AI 予測なし手法よりも高かった．一方で，AI によるラベルが誤っていた場合は，AI の予測を提示した 2 手法によるラベルの正確性は低下し，AI 予測なし手法よりも低い値であった．この結果は，先行研究 [10] と同じ結果であり，人-AI 協調アノテーションがネタバレ画像データセットの質向上に有用である可能性が示された一方で，AI による誤ったラベルの提示が人

表 6: 実験後に実施したアンケートに対する回答の平均値

	AI 予測なし	AI 予測先出し	AI 予測後出し
自身のアノテーションは正確だった	1.36	1.24	1.28
アノテーションはやりやすかった	1.40	0.60	1.16
AI が予測した結果を信頼した	N/A	−0.32	−0.80
AI の予測精度は高いと感じた	N/A	0.00	−0.33

間のラベル判断基準を歪めてしまう可能性が示唆された．

本稿では，実験で用いた画像に対する AI の予測ラベルとその理由を取得するために，GPT-4o に複数のプロンプトを入力し，出力結果を比較して最終的に使用するプロンプトを決定したが，実験で用いた GPT-4o によるラベルと過去の 3 名のアノテータの統合ラベルとの一致度は 0.47 であったことから，AI の予測精度が十分でなかったと考えられる．実験後に実施したアンケートでも，AI の予測結果に対する信頼度や予測精度は低く評価されていた．これに対しては，AI のパフォーマンスに関する説明を人間に提示することで，人間のパフォーマンスが向上することが示されている [25] ため，AI のアノテーション精度を事前に提示することでラベル正確性が改善されると考える．また，ネタバレ画像アノテーションのような，主観的な視点による評価タスクでは，プロンプトにペルソナを入力することが有用である可能性 [26] が示されているため，プロンプトを工夫することによってラベルの予測精度が向上する可能性も考えられる．

同一画像に対する 3 回のアノテーションにおいて，3 回すべてでラベルが一致していた割合は，AI 予測を提示する 2 手法の方が，AI 予測なし手法よりも高い値であった．この結果から，AI の予測精度に関わらず，予測結果を提示することでアノテータのラベル判断基準の一貫性が向上する可能性がある．そのため，AI の予測精度を改善することができれば，AI 予測の提示によってアノテータのラベル正確性と一貫性の両方が向上すると予想される．

アノテーションに要する時間については，AI 予測なし手法の方が AI 予測を提示する 2 手法よりも短い時間であった．これは先行研究 [10] と同様の結果であった．AI 予測なし手法は自身で考えたラベルを選択するだけで良いが，AI 予測を提示する手法では，表示された予測ラベルとその理由を読み，それを踏まえて最終的にどのラベルを選択するか考える必要があるため，アノテーション時間が長くなったと考えられる．AI 予測提示手法におけるこのようなラベル決定の過程がアノテータの負荷を増加させたため，アノテーションのやりやすさに関して AI 予測提示手法の評価が低くなった（表 6）と考える．本実験は図 2 に示すようなインタフェースでアノテーションを行ったが，AI による予測ラベルをデフォルトの選択肢にして予測理由のみを表示したり，予測理由の重要な部分を抽出 [27] して強調表示したりすることで，アノテータの負荷を軽減できる可能性がある．

5.2 AI 予測の提示タイミングの違いによる影響

AI 予測先出し手法と AI 予測後出し手法においては，ラベル一致度やラベル正確性で AI 予測先出し手法が上回る結果となった．同一画像に対するラベルの一貫性や，アノテーションに要する時間については，2 つの手法で差はみられなかった．また，実験後のアンケートにおけるアノテーションのやりやすさについての質問では，AI 予測後出し手法の方が高評価であったが，AI が予測した結果の信頼度や予測精度は AI 予測先出し手法の方が高く評価されていた．これらの結果より，AI による予測の提示タイミングによって，提示内容は同じであっても信頼度や精度の認識に差が生じる可能性があり，そのような違いがラベル一致度やラベル正確性に影響を及ぼしたと考える．本稿の結果からは，アノテーションの精度向上の点では AI による予測を先に提示することが望ましいと考えられる．しかし，アノテーションのやりやすさには問題があるため，5.1 節で示したような方法によって改善が必要である．

5.3 展望

今後は，入力プロンプトの工夫や同一画像に対する複数回のアノテーション結果の平均を採用するなどの方法によって，ネタバレ画像アノテーションにおける AI の予測精度を改善し，AI 予測先出し手法の有用性を再検証する予定である．また本稿では，実験参加者によってラベルの判断基準が大きく異なっていたため，今後アノテーションを行う際は，アノテータ間の判断基準を事前に統一できるように練習タスクなどを設ける予定である．さらには，アノテータの負荷を軽減できるようなアノテーションインタフェースの実装にも取り組み，AI 予測先出し手法の有用性の再検証を行った後，ネタバレ画像データセットの再構築と拡張に取り組みたい．

6. おわりに

本稿は，スポーツのネタバレ画像データセットの品質向上を目的として，人-AI 協調アノテーションの有用性を検証し，AI による予測結果の提示タイミングの違いが人間のラベル決定に異なる影響を及ぼすか調査を行った．

我々は AI 予測なし手法，AI 予測先出し手法，AI 予測後出し手法を用いて，ネタバレ画像データセットの一部の

データに対して再アノテーションを行う実験を実施した．実験の結果，AI 予測を提示した 2 手法によって，AI が提示した予測ラベルが正しい場合はラベルの正確性が向上する可能性が示された．また，同一アノテータにおけるラベルの一貫性の向上については，AI 予測を提示することが有用である可能性が示された．AI による予測ラベルの提示タイミングについては，人間がアノテーションした後に提示するよりも，アノテーションする前に提示した方が，ラベルの一致度や正確性は高くなる可能性が示唆された．

本稿では，ネタバレ画像アノテーションにおける AI のラベル予測精度が十分でなかったため，今後は予測精度の改善に取り組み，アノテータ間の判断基準を統一する練習タスクをアノテーション前に実施して，AI 予測先出し手法の有用性を再検証する予定である．また，再検証の結果を踏まえて，ネタバレ画像データセットの再構築と拡張に取り組む予定である．

謝辞 本研究の一部は JSPS 科研費 JP22K12135 の助成を受けたものです．

参考文献

[1] Radeta, M., Freitas, R., Rodrigues, C., Zuniga, A., Nguyen, N. T., Flores, H. and Nurmi, P.: Man and the Machine: Effects of AI-assisted Human Labeling on Interactive Annotation of Real-time Video Streams, *ACM Trans. Interact. Intell. Syst.*, Vol. 14, No. 2, pp. 1–22 (2024).

[2] Mikulová, M., Straka, M., Štěpánek, J., Štěpánková, B. and Hajič, J.: Quality and Efficiency of Manual Annotation: Pre-annotation Bias (2023).

[3] Zhang, Z., Ning, Z., Xu, C., Tian, Y. and Li, T. J.-J.: PEANUT: A Human-AI Collaborative Tool for Annotating Audio-Visual Data, *Proc. of UIST '23*, pp. 1–18 (2023).

[4] Gilardi, F., Alizadeh, M. and Kubli, M.: ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks, *Proceedings of the National Academy of Sciences*, Vol. 120, No. 30, p. e2305016120 (2023).

[5] Zendel, O., Culpepper, J. S., Scholer, F. and Thomas, P.: Enhancing Human Annotation: Leveraging Large Language Models and Efficient Batch Processing, *Proc. of CHIIR '24*, pp. 340–345 (2024).

[6] Kaikaus, J., Li, H. and Brunner, R. J.: Humans vs. ChatGPT: Evaluating Annotation Methods for Financial Corpora, *Proc. of IEEE BigData '23*, pp. 2831–2838 (2023).

[7] He, Z., Huang, C.-Y., Ding, C.-K. C., Rohatgi, S. and Huang, T.-H. K.: If in a Crowdsourced Data Annotation Pipeline, a GPT-4, *Proc. of CHI '24*, pp. 1–25 (2024).

[8] Hamilton, K., Longo, L. and Bozic, B.: GPT Assisted Annotation of Rhetorical and Linguistic Features for Interpretable Propaganda Technique Detection in News Text., *Proc. of WWW '24 Companion*, pp. 1431–1440 (2024).

[9] Kinoshita, Y., Takaku, T. and Nakamura, S.: Detecting Sports Spoiler Images on YouTube, *Proc. of CollabTech '24*, pp. 114–128 (2024).

[10] Wang, X., Kim, H., Rahman, S., Mitra, K. and Miao, Z.: Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels, *Proc. of CHI '24*, pp. 1–21 (2024).

[11] Aldeen, M., Luo, J., Lian, A., Zheng, V., Hong, A., Yetukuri, P. and Cheng, L.: ChatGPT vs. Human Annotators: A Comprehensive Analysis of ChatGPT for Text Annotation, *Proc. of ICMLA '23*, pp. 602–609 (2023).

[12] Zhu, Y., Zhang, P., Haq, E.-U., Hui, P. and Tyson, G.: Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks (2023).

[13] Nguyen, T. H. and Rudra, K.: Human vs ChatGPT: Effect of Data Annotation in Interpretable Crisis-Related Microblog Classification, *Proc. of WWW '24*, pp. 4534–4543 (2024).

[14] Li, J.: A Comparative Study on Annotation Quality of Crowdsourcing and LLM via Label Aggregation (2024).

[15] Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z. and Yang, D.: Can Large Language Models Transform Computational Social Science?, *Computational Linguistics*, Vol. 50, No. 1, pp. 237–291 (2024).

[16] Gligorić, K., Zrnic, T., Lee, C., Candès, E. J. and Jurafsky, D.: Can Unconfident LLM Annotations Be Used for Confident Conclusions? (2024).

[17] Chatrath, V., Lotif, M. and Raza, S.: Fact or Fiction? Can LLMs be Reliable Annotators for Political Truths? (2024).

[18] Nie, J., Zhang, G., An, W., Tan, Y.-P., Kot, A. C. and Lu, S.: MMRel: A Relation Understanding Benchmark in the MLLM Era (2024).

[19] Abdelhamed, A., Afifi, M. and Go, A.: What Do You See? Enhancing Zero-Shot Image Classification with Multimodal Large Language Models (2024).

[20] Ferguson, S. A., Aoyagui, P. A. and Kuzminykh, A.: Something Borrowed: Exploring the Influence of AI-Generated Explanation Text on the Composition of Human Explanations, *Extended Abstracts of CHI '23*, pp. 1–7 (2023).

[21] Green, B. and Chen, Y.: The Principles and Limits of Algorithm-in-the-Loop Decision Making, *Proc. ACM Hum.-Comput. Interact.*, Vol. 3, No. CSCW, pp. 1–24 (2019).

[22] Huang, F., Kwak, H. and An, J.: Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech, *Proc. of WWW '23 Companion*, pp. 294–297 (2023).

[23] OpenAI: Hello GPT-4o (2024). available from https://openai.com/index/hello-gpt-4o (accessed 2025-01-20).

[24] Rousseeuw, P. J. and Hubert, M.: Robust Statistics for Outlier Detection, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 1, No. 1, pp. 73–79 (2011).

[25] Cabrera, A. A., Perer, A. and Hong, J. I.: Improving Human-AI Collaboration With Descriptions of AI Behavior, *Proc. ACM Hum.-Comput. Interact.*, Vol. 7, No. CSCW1, pp. 1–21 (2023).

[26] Fröhling, L., Demartini, G. and Assenmacher, D.: Personas with Attitudes: Controlling LLMs for Diverse Data Annotation (2024).

[27] Majumder, B. P., Camburu, O., Lukasiewicz, T. and Mcauley, J.: Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations, *Proc. of ICML '22*, Vol. 162, pp. 14786–14801 (2022).