人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第36回)
SIG-AM-36-01

タスク指示と発話系列の表現に注目した LLM の「チャット対話分離」プロンプト最適化

Prompt Optimization for LLM-Based Dialogue Disentanglement Focusing on Task Instructions and Utterance-Sequence Representations

高田尚輝 [1]　　森辰則 [1]
Naoki Takada[1]　Tatsunori Mori[1]

[1] 横浜国立大学大学院環境情報学府
[1] Graduate School of Environment and Information Sciences, Yokohama National University

Abstract: Dialogue disentanglement separates intertwined utterances in multi-user online chats into coherent dialogue, enabling reliable downstream processing. While large language models (LLMs) are promising for this task, prompt-based approaches can be unstable across models and open-weight LLMs still lag behind non-LLM methods. This paper proposes an automatic prompt optimization for LLM-based dialogue disentanglement. We decompose a prompt into three components—task instructions, utterance-sequence representation, and output-format constraints—and optimize them using GEPA, an optimization method for compound AI systems. Experiments on benchmark datasets show that the optimized prompts improve dialogue disentanglement accuracy over the original prompts and can surpass carefully hand-crafted prompts.

1　はじめに

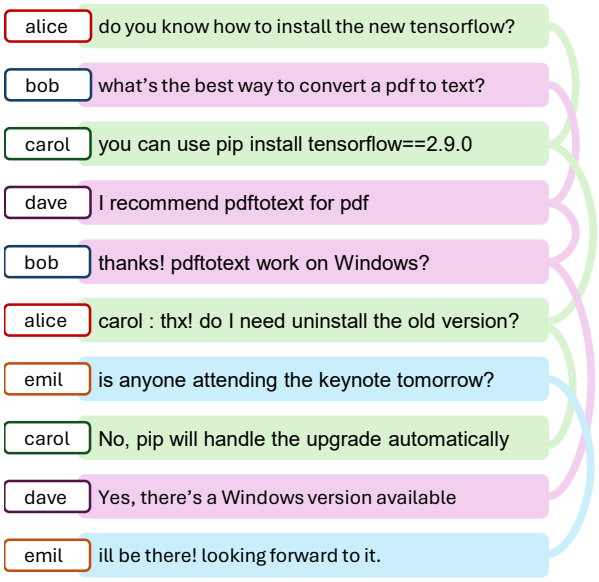

図 1　分離後の対話を色ごとに示す

多数の参加者によって構成される Slack や Discord 等のオンラインチャットにおいて，複数の対話が同時に進行し交錯する現象が生じる．この交錯はユーザがチャット中の文脈を追跡する難易度を上げる．加えて，機械的な解析においても無関係な発話がノイズとして混入し，処理の精度低下を招く．この課題に対し，対話分離 (Dialogue Disentanglement) が提案された [2]．対話分離は図 1 に示すように，対話が交錯している発話系列を応答関係で結びついた一貫性のある対話クラスタへ分割するタスクである．本技術は，対話状態追跡 [3] や応答生成 [4] といった下流タスクの重要な基盤となる．従来，対話分離は教師あり機械学習による発話間ペアワイズ分類として扱われてきた [2]．チャットの各発話について，新しい対話の始まりであるか，もしくはそれ以前のどの発話と応答関係があるのか判定することで対話構造を明らかにする．

大規模言語モデル (LLM) は高度な文脈理解力と推論力を有しており，対話分離への応用が期待される．その初の試みは既存の非 LLM 手法より著しく劣る結果であった [8]．しかし，プロンプトを工夫することでその精度を向上させることができると報告されている．対話単位の判定，後続文脈の付与を導入することで既存手法を凌駕

する精度を達成し，対話分離への LLM 応用の可能性が示された [20]．一方で，対話単位の判定と後続文脈の付与は LLM によって有効性が異なり，精度を低下させる場合もある．そのため，LLM による対話分離の精度を向上させる汎用的手法は未だ存在しない．さらに，LLM が 30B 程度のモデルになると，既存の非 LLM ベースの手法よりも大きく性能が劣ってしまい，対話分離への LLM 応用には改善の余地が残されている．

対話分離は人間にとっても難易度が高い．人手のアノテーションでは多くの揺れが生じ，複数人によるアノテーション結果のすり合わせが必要とされている [5]．そのため，人手で詳細なタスク指示を考えることは困難である．加えて，LLM はプロンプト中での情報の提示形式や出力形式の指示によって性能が大きく変動する [17, 18]．対話分離においてもその傾向は顕著であり，発話系列の表現や出力形式の指示によって精度が大きく変わってしまう (後述する第 5 章，特に表 1 を参照)．そのため，形式の差も考慮したプロンプト設計が重要となる．以上のような背景から，最適な対話分離プロンプトの設計と，その検証を人手で行うことは難しい．そこで本研究では，対話分離における自動プロンプト最適化手法を提案する．プロンプトを次の 3 つの要素に分け，それぞれを最適化する．

タスク指示: 対話分離の定義を説明する．どのようなタスクであるのか，何を根拠に判定を行うべきか説明する．

発話系列表現: 発話番号，発話時間，発話者，テキストメッセージの組を与える．形式変換 (例:JSON で表現) やラベル付けなどを行い，LLM にとって理解しやすい形式で発話系列を提示する．

出力形式指示: 対話分離の判定結果を，LLM が判定に成功しやすい形式で出力する方法を指示する．

本研究の最適化対象は複数要素から成るプロンプトである．そこで，複数プロンプトの自動最適化において最先端の成果を上げている GEPA[15] を利用した手法を提案する．本研究の貢献は設計の難易度が高い対話分離プロンプトを自動最適化する手法を提案したことにある．最適化前のプロンプトを大幅に上回る対話分離精度を達成し，Takada らの研究 [20] で人手の試行錯誤によって作られたプロンプトの結果を上回る精度を達成した．

2 関連研究

2.1 対話分離

対話分離は，長らく 2 段階の処理プロセスとして定式化されてきた [2, 6, 7]．第 1 段階では，指定範囲内の全発話ペアに対して応答関係の度合いを示すスコアを付与する．第 2 段階では特定された関係を統合し，全体的な対話構造を構築する．初期の機械学習モデルは人手の特徴量設計に依存しており，語彙の重複，時間差，ユーザ言及などが利用された [1, 2]．その後，大規模なアノテーション付きコーパス [5] の登場により，エンドツーエンドのニューラルモデル開発が促進された．さらに，BERT 等の事前学習済み言語モデルの導入により，文脈情報の活用が進み精度が向上した [6]．しかし，これらの手法は会話を単なる発話系列として扱う傾向があり，対話構造の考慮が不十分であった．この課題に対し，話者の特性やユーザへの言及といった対話特有の特徴を組み込むアプローチが登場した [7]．DiHRL[8] では，階層的学習損失や easy-first デコードアルゴリズムの統合，大域的な会話特性を捉えることでさらに精度を向上させている．

LLM は自然言語処理において多くの成果を上げているが，対話分離への適用は十分に探求されていない．初の試み [8] は，既存の非 LLM ベースの手法に大きく劣った．この実験で使われたタスク指示は説明性に富むものではなく，「応答関係のある発話を特定し，その発話インデックスを出力せよ．近い発話同士は応答関係を持ちやすい．」のみであり，発話系列表現と出力形式指示においても，簡単な用例を試すに留まっていた．これを受け，Takada らの研究 [20] では，発話系列表現と出力形式指示の改善によって，既存手法を大きく上回る LLM ベースの手法を提案した．発話系列を JSON 形式で表現し，過去の判定から明らかになった対話構造を示す手法 (DLA)，判定対象の発話以降に続く発話を参考情報として挿入する手法 (SC) が提案されている．しかし，一部のクローズモデルでは高精度である一方，30B 程度の LLM では精度が著しく低下してしまう．パラメータの少ないオープンモデルでも高精度な分離ができることは，個人情報を含むチャットデータを扱う上で有用である．本研究はその実現に向け，LLM の対話分離精度を向上させる問題に取り組む．

精度向上を目指すにあたり，より説明性に富んだタスク指示の作成が考えられる．判定の基準が記述されており，どのような発話であっても正しく，一意に判定することができることが望ましい．タスク指示の原案として，人手のアノテーション時に使われた指示書を参考にすることが考えられる．しかし，これまでに多くの対話分離アノテーションが行われた [5, 19] 一方で，詳細に定義された指示書は設計されていない．唯一参照可能であるものも極めて簡単な内容に限る [5]．この指示書は複数のアノテータによる意見統合が前提に作られており，一意にアノテーション可能な高品質なものではない．アノテーション実験では判断の一致度を示すカッパ係数が 0.75 未満であり，判断に揺れが生じていた．対話分離アノテーションでは，複数の応答先が考えられる場合や多様な話

題展開，アノテーション時の個人差があり，高品質な指示書を作成することは難しい．同様に，LLM に与える対話分離のタスク指示作成は困難であり，発話系列表現や出力形式指示まで含めた最適プロンプトの設計は困難を極める．そこで本研究では，自動プロンプト最適化を利用し，最適プロンプト設計を目指す．

2.2 自動プロンプト最適化

LLM の性能を引き出すためには適切なプロンプト設計が求められる．しかし，手動設計は多大な試行錯誤を要し，最適プロンプトの設計は難しい．この問題を解決するべく，自動プロンプト最適化の研究が進められている．初期の研究として，Prompt Tuning[10] や Prefix-Tuning[11] が提案された．これらは自然言語の指示文を探索する代わりに，入力に付与する学習可能な連続ベクトルを最適化する手法である．LLM 本体の重みを固定し，追加した少数のベクトルのみを勾配法で更新する．ファインチューニングよりもコストを抑えられる一方，得られるプロンプトがベクトルであり，人間が内容を読んで解釈することはできない．さらに，ベクトルの次元や注入位置が LLM に依存するため，他の LLM への移植が難しいという課題があった．

これに対し，APE[12] や OPRO[13] といった自然言語プロンプトを最適化する手法は，LLM 自身を最適化器としてプロンプトの探索を行う．これらは最適プロンプトの可読性と汎用性を向上させた．EvoPrompt[14] のように，進化計算的手法を導入することで膨大な探索空間を考慮した最適化手法も提案されている．しかし，その多くは単一タスクの最適化に留まり，複数のエージェントや推論ステップが相互作用する複合システムを最適化する能力には限界があった．近年では，複合システムの最適化に向けて，MIPROv2[16] のようなベイズ最適化やブートストラップ探索を応用した手法も提案されている．本研究では，上記をさらに発展させた，複合システム全体の挙動に対する内省とパレート最適化を導入した GEPA[15] を応用する．GEPA を選んだ理由は 3 つある．第 1 に，複合システムを最適化対象としている点である．第 2 に，自然言語プロンプトを最適化対象としており，利用する LLM が変更された場合でも一定の性能を維持できる点である．第 3 に，先述した 2 点を満たす最先端の精度を誇る手法である点である．しかしながら，GEPA のシステムをそのまま利用することはできない．GEPA の最適化対象はプロンプト中のタスク指示のみであり，発話系列表現や出力形式指示まで扱うことができないからである．よって本研究では GEPA を応用し，発話系列表現と出力形式指示も含めた最適化手法を提案する．

3 LLM による対話分離

3.1 対話分離の定義

チャットデータ中の全発話系列を会話 C と呼び，C から分離された対話の集合を D とする．ここで，各 $d_j \in D$ は，特定の話題や目的を共有し，応答関係で結ばれた発話からなる対話クラスタである．これに基づき，タスクは次のように定式化される．入力は時系列順に並んだ n 個の発話からなる会話 $C = (u_1, u_2, \ldots, u_n)$ である．各発話 $u_i \in C$ は $u_i = (t_i, s_i, m_i)$ で表される．ここで t_i は発話時刻，s_i は発話者，m_i はテキストメッセージである．目的は，C を互いに素な対話クラスタ集合 $D = \{d_1, d_2, \ldots, d_p\}$ に分割することである．この分割は次の 2 つを満たす，C の厳密な分割でなければならない．第 1 は網羅性であり，$C = \bigcup_{d_j \in D} d_j$ が成立することである．第 2 は排他性であり，$\forall j \neq k, d_j \cap d_k = \emptyset$ が成立することである．

3.2 LLM による判定・分離手法

まず，判定対象となる発話 u_{target}(以降対象発話と呼ぶ) を 1 件選び，それより前の発話と共に LLM に与え，新しい対話の始まりであるか，もしくはどの発話と応答関係があるのか，判定させる．ただし，応答先はただ 1 つの発話であることが前提である．C の全発話を判定し，次に，得られた応答関係をもとに先頭の発話から対話クラスタを作る．u_{target} が新しい対話であった場合，u_{target} ただ 1 つを要素とする対話クラスタを作る．u_{target} に応答関係があった場合は，応答先の発話が所属する対話クラスタへ u_{target} を加える．全発話で上記の処理を行い，終了すると，対話クラスタ集合 D を得ることができる．この方法は従来の非 LLM ベースの手法 [2] で用いられており，本研究では判定器を LLM に置き換えて使用する．

上記の方法以外に，u_{target} 以前の発話系列を過去の判定から明らかになった対話構造で表現し，応答先の候補を対話単位で設定する手法 (DLA) や，補足情報として後続文脈を付与する手法 (SC) が提案されている [20]．これらは LLM ごとに精度向上の寄与が異なり，精度が低下する場合もあった．本研究では対話分離プロンプトの自動最適化手法を提案することを目標とし，前段落で述べたように，DLA も SC も導入しない手法を採用する．

3.3 判定のスコア化

GEPA による最適化を行うためには LLM の判定を 1 件ごとにスコア化する必要がある．そのため，判定の成

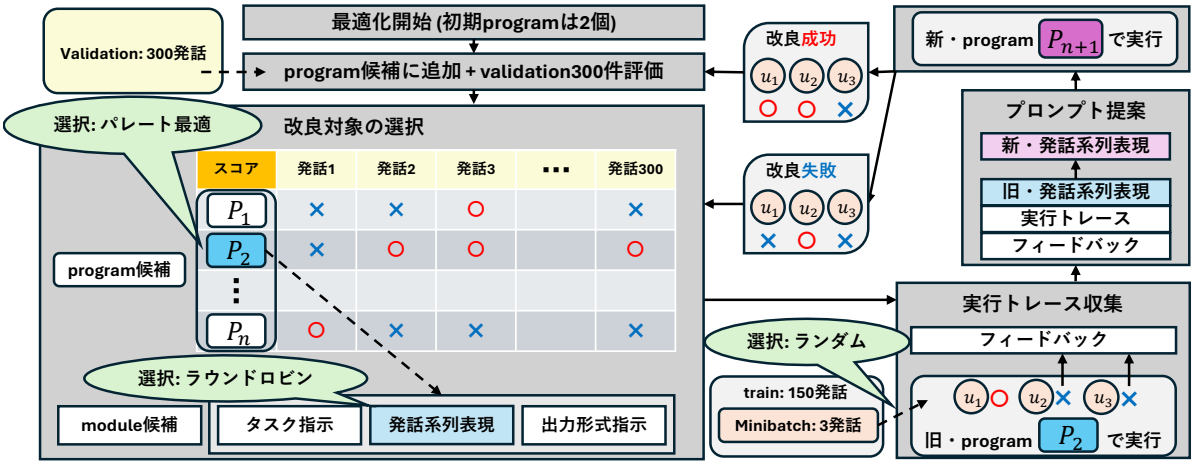

図 2　GEPA による対話分離プロンプト最適化 (改善対象として program 2 の発話系列表現 module が選ばれた例．プロンプト提案する module は発話系列表現に限らず，候補からラウンドロビンで選択される．)

否を 2 値で評価し，成功を 1，失敗を 0 としてスコア化した．判定の成否は次のように定義される．まず，正解データ上で u_{target} が属する対話クラスタを d_{gold} と表す．LLM による判定結果は u_{target} が (i)「新しい対話の始まりである」または，(ii)「ある先行発話と応答関係を持つ」のいずれかである．各判定の成功条件は，(i)「u_{target} が d_{gold} の先頭の発話である」(ii)「d_{gold} 内に，特定された先行発話が含まれる」である．ただし，(ii) の場合は，特定された先行発話が u_{target} よりも前の発話を指定する必要がある．上記の条件を満たさない結果は失敗とみなす．

4　自動プロンプト最適化

4.1　GEPA

1 章で述べたように，対話分離プロンプトを，タスク指示，発話系列表現，出力形式指示の 3 要素に分けて最適化する．この各要素を module と呼び，3 つの module の組を使用して対話分離を行うシステムを program と呼ぶこととする．GEPA[15] を応用して対話分離プロンプトを最適化する手法を図 2 に示す．最適化の過程では，まず，初期 program 群を候補プールに追加し，それらを 300 件の validation データで評価してスコアを算出する．次に，パレート最適な program が改善候補として選ばれる．その過程は 3 段階に行われ，まず各 program において，300 件の validation データにおける各タスクの成否を確認する．次に，ある program の成功したタスクが他の program によって全て成功され，総スコアでも劣っていた場合，その program を候補から除く．最後に，残った候補から総スコアに比例した確率で program が選択される．program が選ばれると，次は，どの module を改善するかをラウンドロビンによって決める．改善対象が決まったら，train データからランダムに抽出した minibatch 3 件を使用し，旧・program で対話分離の判定を行い，実行トレースを収集する．実行トレースには LLM の入出力，推論過程が含まれる．判定に失敗した場合は LLM に失敗理由を生成させ，フィードバックを収集する．フィードバックの生成方法は GEPA で定められておらず，利用者に委ねられている．本研究では，失敗した判定ごとに入出力と真の解を与え，失敗理由を生成させた．minibatch 上でのトレース収集が完了したら，改善を試みる．対象 module の旧・プロンプトと実行トレース，フィードバックが含まれた改善指示を LLM に与え，新・プロンプトを得る．旧，新・プロンプトは対話分離プロンプト全文ではなく，選択された module の要素のみである．新・プロンプトを反映した新・program を minibatch 上で評価し，旧・program の精度を超えた場合に改善成功である．新・program を program 候補に加え，300 件の validation データで評価する．改善失敗の場合は新・program を棄却し，再び改善対象を選びなおす．このサイクルを決められた回数だけ繰り返し，スコアの高い program が最適 program となる．使用したプロンプトやデータセットの詳細は github[*1] にて公開する．

GEPA[15] の改善指示は「実行トレースを注意深く見て新しい指示を作りなさい」である．本研究はこの改善指示を修正し，さらに，module に外部機能を追加することで対話分離プロンプトの最適化を行った．次節ではその詳細を述べる．

[*1] https://github.com/haniwara/sigam2026

4.2 タスク指示

元の改善指示に加え,「出力形式の指示を作ってはいけない」という条件を加えた．この条件が無い場合，得られるプロンプトに出力形式の指示が含まれ，module ごとに分離して改善にすることができないためである．外部機能の追加は無い．最適化過程では出力されたタスク指示の改善案を加工することなくプロンプトに挿入して評価する．

4.3 発話系列表現

新しい発話系列表現の案を得ることを目的とし，元の改善指示を「対話分離を行う LLM にとって理解しやすい発話系列の表現方法を提案せよ」に差し替えた．得られた表現案を任意の発話系列で使うため，発話の形式変換コードを作成する外部機能を設けた．コードは LLM で作る．表現の案を LLM に与え，事前に用意した形式から表現案の形式に変換するコードの作成を依頼する．新・program の評価では，発話系列を形式変換コードで変換し，プロンプトに挿入する．どのような発話系列も提案された表現での挿入を目指す．

4.4 出力形式指示

改善指示を「対話分離を行う LLM にとって判定が正確になる出力形式指示を提案せよ」に差し替えた．提案された出力形式に対応するため，出力解釈コードを作成する外部機能を設けた．コードは LLM で作る．新しい出力形式指示を事前に用意した形式へ変換するコードを作成するよう依頼する．新・program の評価では，新形式で出力される判定を得られたコードで変換し，評価する．

5 評価実験

5.1 実験目的

提案手法を 2 つの観点から評価する．第 1 に，LLM の対話分離精度を向上させることができるか評価する．人手設計のプロンプトと最適プロンプトの精度を比較する．第 2 に，小さなモデルがプロンプト最適化によって高性能なモデルにどこまで迫ることができるか評価する．

5.2 データセット

本研究では，対話分離の標準的なベンチマークである Ubuntu IRC データセット [5] を使用した．本データは，Ubuntu OS に関する技術的な課題解決を目的とした実際のチャットログであり，複数の参加者による交錯した対話が含まれている．データセットは train，dev，test の 3 つに分割されている．このうち dev および test は，複数のアノテータによる検証を経た高品質な正解データであり．dev は 2500 件，test は 5000 件の発話を含む．プロンプト最適化には dev を使い，改善前後での性能差の検証，および，先行研究との比較では test を使用した．

5.3 比較手法

プロンプトが異なる四つの手法について評価する．1 つ目は最適化の初期 program の 1 つで Seed 1 と呼ぶ．タスク指示は DiHRL[8] で試験的に作られた簡単なものである，発話系列表現は，発話の各要素を空白で区切るのみである．出力形式指示は，判定結果を発話番号のみで扱い，新しい対話の始まりである場合は自身の番号，応答先がある場合は応答先の番号の出力を指示する．2 つ目はもう 1 つの初期 program で Seed 2 と呼ぶ．Seed 1 から出力形式指示のみ変更し，新しい対話の始まりであるかを true/false で表すキーと，応答先の発話番号を示すキーの，2 キーでの出力を指示する．3 つ目は Takada らの研究 [20] で Baseline と呼ばれるプロンプトの再現であり，同じく Baseline と呼ぶ．タスク指示は Seed 1 と同じで，発話系列表現は各要素のラベル付き JSON 形式，出力形式指示は Seed2 と同じである．4 つ目は最適化された program を使用する方法であり，Optimum と呼ぶ．Optimum は Seed 1 をタスク指示，発話系列表現，出力形式指示の順に 1 回ずつ改善して得られた。

Baseline は Seed 1, 2 と比べて判定の精度が高い (表 1 参照) ため，2 キーでの出力と JSON 形式の発話表現は精度が向上する要因と考えられる．しかし，Seed 1 のみで始めた最適化では JSON 形式の発話表現が早期に提案される一方，2 キー出力が提案されることは稀であった．初期の改善を円滑に進めるため，Seed 2 を用意した．

5.4 評価指標

対話分離の研究で広く使われている評価指標を採用し，3 つの観点から評価する．第 1 に，クラスタリングの全体的な整合性を評価する，Variation of Information (VI)[9]，Adjusted Rand Index (ARI)[22]，Normalized Mutual Information(NMI)[21]，One-to-One accuracy(1-1)[2]，および Shen-F1 (S-F)[23]．第 2 に，発話 3 件ごとの局所的なクラスタリング精度を測る Local_3[2]．第 3 に，対話クラスタの完全一致した数を評価する，適合率 (P)，再現率 (R)，および F1 スコアを用いた [5]．

表 1　使用 LLM とプロンプトの違いによる成功率の変化

LLM	手法	成功率 (%)
Qwen3-30B	Seed1	80.28
Qwen3-30B	Seed2	81.60
Qwen3-30B	Baseline	91.92
GPT5.2	Baseline	96.12

5.5　ハイパーパラメータ

データセット構築 (5.6 節) を除き，使用する LLM は Qwen3-30B-A3B-Thinking-2507 に限る．リクエスト時のパラメータは (i) 応答関係の判定と (ii) プロンプト改善によって異なる．(i) では LLM での対話分離における再現性確保のため，temperature: 0 とした．(ii) では GEPA と同様に改善提案に多様性を持たせるため，temperature: 0.6，top-p: 0.95，top-k: 20 とした．

応答関係の判定には，候補となる先行発話の範囲を 50 件とした．GEPA のバージョンは v0.023 を使用し，学習データの件数は GEPA[15] に従い，train150 件，validation300 件，minibatch の選択 3 件とした．merge は汎用的な有効性が無いため使用しなかった．

5.6　プロンプト最適化用データセット構築

プロンプト最適化の学習データとして 450 件のタスクを集めたデータセットを構築した．IRC データの dev セット 2500 件を使い，LLM が判定可能な発話のうち難易度が高いもので構成されている．構築方法は次の通りである．まず，使用する LLM とプロンプトを変えながら，表 1 のように判定結果を収集する．次に，いずれの手法でも判定に失敗した 63 件の発話を LLM が判定不可能と候補から除外する．IRC データにはアノテーションミスと思われるものや，対応可能な発話範囲 50 件を超えた応答関係を特定しなければならない場合があるためである．最後に，判定成功率の高い手法で失敗した発話は難易度が高いとみなし，難易度の高い発話でデータセットを構築した．なお，train と validation で難易度は均一化した．

6　実験結果

6.1　ベンチマーク評価

IRC データの test セット 5000 発話に対する各手法の精度を表 2 に示す．上段に非 LLM ベースの最先端手法，中段に高性能なクローズモデルによる分離，下段に本研究の結果を示す．Optimum は Seed と比較して大幅な精度向上を達成した．Baseline と比較しても，全ての評価指標において精度が向上した．この結果は，本研究で提案した自動プロンプト最適化手法の有効性を示す．しかしながら，表の上段にある非 LLM ベース手法の DiHRL[8] や，クローズモデルによる分離 [20] には劣る結果となった．

6.2　最適化の限界

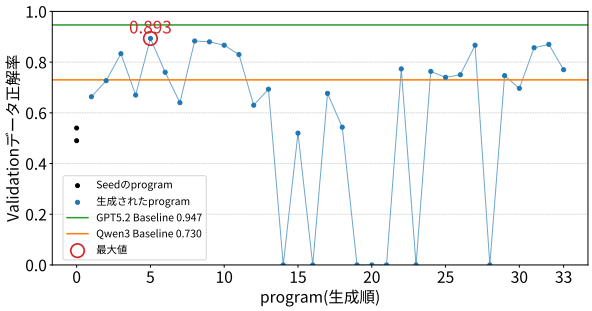

図 3　生成された program の最適化過程

最適化の過程で生成された program とそのスコアについて図 3 に示す．5 個目に生成された program のスコアが最高値となり，それ以降に生成される program では改善が見られなかった．Qwen3-30B の Baseline からの改善は達成したが，GPT5.2 の Baseline スコア以上に改善されることは無かった．最適化回数を増やすことで改善される可能性がある一方，どの program の改善も正解率 0.88 付近に収束し，改善が進行していない．さらに，改善された program は多くの train データで正解してしまい，失敗から内省する事例も少なくなっていた．よって本研究では，最適化の限界に達したとみなし，5 個目の program を最適プロンプトとして結論付けた．

6.3　誤判定の分析

Optimum で IRC test セットを判定した結果のうち，誤判定した事例を定性的に分析した．誤判定は以下の 5 つに分類される．1 つ目は “hi” や “hmm” などの挨拶やリアクションである．同じ人が連続して発話したものだが，間に関係の無い発話が挟まれていると，これらを別の対話であると判定してしまう．2 つ目はサーバーログやシステムメッセージである．応答関係を持たない発話であるが，テキストメッセージに発話者の名前が含まれているため，応答関係があると判定してしまう．3 つ目は “Wine” や “hoary” など，ubuntu OS の専門用語を含む発話である．内容の関連性から応答関係を特定しなけれ

表 2　対話分離精度の比較: 上段に非 LLM ベースの最先端手法，中段にクローズモデルを使った最適化無しプロンプト，下段にオープンモデルを使った最適化前プロンプト (Seed1, Seed2), 人手作成されたプロンプト (Baseline), 最適化プロンプト (Optimum) による結果を示す．各段の最大値を下線，全結果における最大値を太字で表す．

Method		VI	ARI	1-1	NMI	Local_3	S-F	P	R	F1
Elsner [1]		82.10	—	51.40	—	—	—	12.10	21.50	15.50
DiaBERT [6]		93.20	72.80	79.70	—	—	—	42.10	47.90	44.80
Struct [7]		94.60	76.80	84.20	—	—	—	51.80	51.80	51.70
DiHRL [8]		94.23	81.10	84.20	91.85	95.64	87.50	47.97	49.86	48.90
GPT4.1	DLA+SC[20]	95.39	82.22	86.34	96.65	95.14	90.55	46.38	48.73	47.53
Gemini2.5pro	DLA+SC[20]	**97.16**	**92.23**	**90.78**	**97.88**	**97.62**	**92.02**	**58.81**	**63.94**	**61.27**
Qwen3-30B	Seed1	82.97	37.95	56.68	83.83	84.89	63.37	12.03	18.03	14.43
Qwen3-30B	Seed2	83.31	38.98	56.14	84.36	84.68	62.97	11.26	16.34	13.33
Qwen3-30B	Baseline	93.00	70.02	78.46	93.62	94.32	82.81	38.30	40.56	39.40
Qwen3-30B	Optimum	94.12	75.87	82.26	95.51	95.39	84.80	42.22	42.82	42.52

ばならない状況であるが，適切な応答先を見つけることができなかった．4 つ目は抽象的な意見や感想で，応答先が複数考えられる場合である．応答関係は対話ごとに特定の発話者によって構成される．発話者が誰と話しているのか特定し，応答先を判定しなければならない状況で誤判定が起きた．参加者の名前を言及している発話であっても，判定に失敗してしまう場合も存在した．また，対象発話より後ろの発話を参照することができれば，文脈を推測して判定が容易になるような場合も存在した．

7　考察

GEPA[15] では長期の改善プロセスを経て最適解にたどり着く一方，本研究では早期に出た最適解以降に改善が起きなかった．これには次の 2 つが原因として考えられる．第 1 に，多様な事例を考慮したトレースを作成できない点である．話題や参加者の交流，話し口調はチャット内に多様に存在する．その全てに対応できることが理想である一方，改善時には 3 つのトレースしか参考にすることができない．参照するトレースの量が不十分である可能性がある．しかし，トレース 3 件を含んだ改善指示プロンプトは既に 2 万トークンを超えている．単にトレースを増やすだけでは改善指示プロンプトが肥大してしまい，LLM がトレースを十分に考慮することができないと考えられる．トレースを増やす場合は，要約したトレースを挿入するなどしてトークンを減らす必要がある．第 2 に，最適化用データセットの難易度が低い点である．判定難易度の高い発話を特定してデータセットを構築したものの，図 3 ではその半数以上が Seed によって正解されてしまう．高品質な対話分離データセットを大量に用意し，難易度の高い発話を集めることで更なる改善を促す可能性があると考えられる．

8　おわりに

LLM を使った対話分離において自動プロンプト最適化を導入し，その有効性を示した．プロンプトの要素をタスク指示，発話系列表現，出力形式指示に分解し，各要素を最適化した．30B のオープンモデルで最適プロンプトを使った結果は既存の非 LLM ベース手法や最適化の無いクローズモデルの結果には及ばなかったが，その精度を向上させることができた．本研究が，LLM による対話分離プロンプトの最適化に用いられることを期待する．

9　限界と課題

入力可能なトークン数には上限があり，範囲外にある長距離の応答関係は捕捉できない．使用したデータセットはプログラミングの話題に特化しており，他ドメインでの会話内容について有効性は未検証である．また，本チャットデータは公開されており，LLM の事前学習データに含まれている可能性も否定できない．

GEPA のハイパーパラメータ調整，DLA や SC[20] を導入したプロンプトの最適化によってさらに精度向上する可能性がある．

謝辞

本研究の一部は，ＮＥＤＯ（国立研究開発法人新エネルギー・産業技術総合開発機構）の委託事業「経済安全保障重要技術育成プログラム／先進的サイバー防御機能・分析能力強化」（JPNP24003）によるものである．また，本研究の一部は JSPS 科研費 JP24K15084, JP23H00491 の助成を受けたものである．

参考文献

[1] Elsner, M., Charniak, E.: You talking to me? A corpus and algorithm for conversation disentanglement, *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 834–842 (2008)

[2] Elsner, M., Charniak, E.: Disentangling chat, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 117–126 (2010)

[3] Ouyang, Y., Chen, M., Dai, X., Zhao, Y., Huang, S., Chen, J.: Dialogue state tracking with explicit slot connection modeling, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 34–40 (2020)

[4] Cai, X., Fu, Y., Zhao, H., Jiang, W., Pu, S.: Memory Graph with Message Rehearsal for Multi-Turn Dialogue Generation, *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM)*, pp. 108–117 (2022)

[5] Kummerfeld, J. K., Gouravajhala, S. R., Peper, J. J., Athreya, V., Gunasekara, C., Ganhotra, J., Patel, S. S., Polymenakos, L. C., Lasecki, W.: A large-scale corpus for conversation disentanglement, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3846–3856 (2019)

[6] Li, T., Gu, J.-C., Zhu, X., Liu, Q., Ling, Z.-H., Su, Z., Wei, S.: DialBERT: A hierarchical pre-trained model for conversation disentanglement, *arXiv preprint arXiv:2004.03760* (2020)

[7] Ma, X., Zhang, Z., Zhao, H.: Structural characterization for dialogue disentanglement, *arXiv preprint arXiv:2110.08018* (2021)

[8] Li, B., Fei, H., Li, F., Wu, S., Liao, L., Wei, Y., Chua, T.-S., Ji, D.: Revisiting conversation discourse for dialogue disentanglement, *ACM Transactions on Information Systems*, Vol. 43, No. 1, pp. 1–34 (2025)

[9] Meilā, M.: Comparing clusterings by the variation of information, *Learning Theory and Kernel Machines (COLT/Kernel 2003)*, pp. 173–187 (2003)

[10] Lester, B., Al-Rfou, R., Constant, N.: The Power of Scale for Parameter-Efficient Prompt Tuning, *arXiv preprint arXiv:2104.08691* (2021)

[11] Li, X. L., Liang, P.: Prefix-Tuning: Optimizing Continuous Prompts for Generation, *arXiv preprint arXiv:2101.00190* (2021)

[12] Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., Ba, J.: Large Language Models are Human-Level Prompt Engineers, *Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023)* (2023)

[13] Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., Chen, X.: Large Language Models as Optimizers, *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)* (2024)

[14] Guo, Q., Wang, R., Guo, J., Li, B., Song, K., Tan, X., Liu, G., Bian, J., Yang, Y.: Connecting Large Language Models with Evolutionary Algorithms Yields Powerful Prompt Optimizers, *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)* (2024)

[15] Agrawal, L. A., Tan, S., Soylu, D., Ziems, N., Khare, R., Opsahl-Ong, K., Singhvi, A., Shandilya, H., Ryan, M. J., Jiang, M., Potts, C., Sen, K., Dimakis, A., Stoica, I., Klein, D., Zaharia, M., Khattab, O.: GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning, *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR 2026)* (2026)

[16] Opsahl-Ong, K., Ryan, M. J., Purtell, J., Broman, D., Potts, C., Zaharia, M., Khattab, O.: Optimizing Instructions and Demonstrations for Multi-Stage Language Model Programs, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9340–9366 (2024)

[17] Sclar, M., Choi, Y., Tsvetkov, Y., Suhr, A.: Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting, *The Twelfth International Conference on Learning Representations*, (2024)

[18] Tam, Z. R., Wu, C.-K., Tsai, Y.-L., Lin, C.-Y., Lee, H.-Y., Chen, Y.-N.: Let Me Speak Freely? A Study On The Impact Of Format Restrictions On Large Language Model Performance, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1218–1236 (2024)

[19] Chatterjee, P., Damevski, K., Kraft, N. A., Pollock, L.: Software-related slack chats with disentangled conversations, *Proceedings of the 17th International Conference on Mining Software Repositories*, pp. 588–592 (2020)

[20] Takada, N., Mori, T.: Rethinking Dialogue Disentanglement for LLMs via Dialogue-Level Assignment and Subsequent Context, *Proceedings of the 10th Linguistic and Cognitive Approaches To Dialog Agents Workshop (LaCATODA) co-located with the 40th AAAI Conference on Artificial Intelligence (AAAI)* (2026)

[21] McDaid, A. F., Greene, D., Hurley, N.: Normalized mutual information to evaluate overlapping community finding algorithms, *arXiv preprint arXiv:1110.2515* (2011)

[22] Hubert, L., Arabie, P.: Comparing partitions, *Journal of Classification*, Vol. 2, No. 1, pp. 193–218 (1985)

[23] Shen, D., Yang, Q., Sun, J.-T., Chen, Z.: Thread detection in dynamic text message streams, *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 35–42 (2006)

仮想ユーザ像を用いた情報推薦インタフェースの LLM を用いた評価に関する予備的検討

Preliminary Study on LLM-based Evaluation of Recommender Interface Employing Virtual Users

山﨑 洋紀 [1*] 柴田 祐樹 [1] 高間 康史 [1]

Hironori Yamazaki[1], Hiroki Shibata[1], Yasufumi Takama[1]

[1] 東京都立大学大学院 システムデザイン研究科
[1]Graduate School of Systems Design, Tokyo Metropolitan University

Abstract: This paper investigates the use of Large Language Model (LLM) instead of human participants to evaluate the recommender interface employing virtual users. Recommender systems employing virtual users have been proposed as an explainable recommendation without collecting information from users. As a means for preliminary considering the interface design, this paper reports a result of an experiment, in which LLM decides what movies to watch using the prototype interface.

1 はじめに

本稿では，仮想ユーザ像を用いた情報推薦インタフェースの評価に LLM を用いるアプローチを提案し，予備的検討を行った結果を報告する．

情報推薦システムは，大量に蓄積された情報から必要な情報を効率よく得るために重要な存在となっている．ユーザから嗜好に関する情報を収集し，それらの情報に基づいて推薦アイテムを選択する．しかし，このような個人化された推薦システムに対し，プライバシーが侵害されていると感じるユーザもいるため，システムへの信頼低下につながる可能性が指摘されている[1].

推薦の受容性を高める方法として，説明可能情報推薦が研究されている．推薦時に，そのアイテムが推薦された理由を説明することで，推薦の透明性や説得力，満足度が向上すると言われている[2].

これらの背景から，推薦対象ユーザから情報を収集しない説明可能情報推薦を目的として，仮想ユーザ像を用いた情報推薦が提案されている[3, 4]．この手法では，推薦の際に仮想のユーザのプロファイルと，対象アイテムに対する評価値を提示する．ユーザはそれらの情報を参考に推薦アイテムを受容するかどうかを判断するため，ユーザの嗜好に関する情報を収集する必要がないという利点がある．

仮想ユーザ像を用いた情報推薦インタフェースにおいて、仮想ユーザの人数が少ない場合，提示された仮想ユーザ像が対象アイテムを受容するかどうかの判断の手がかりとして不足である可能性がある．そのため，多様なユーザにとって有用な情報推薦システムとするために，多くの仮想ユーザ像を用いるべきであると考える．しかし，多数の仮想ユーザ像を全て同時に提示すると情報過多につながり，仮想ユーザ像を参考にすることが困難になる．従って，同時に提示すべき仮想ユーザ数や，提示すべき情報の内容，ユーザが可能な操作など，ユーザインタフェースの設計において検討すべき事項は多岐に渡る．

インタフェースの評価は通常，実験協力者によるユーザ実験に基づき行うことが一般的であるが，実験協力者を集めるコストや，実験協力者の負担などのため，多数の要素を検討することは困難である．

そこで本稿では，仮想ユーザ像を用いた情報推薦インタフェースの初期段階での検討に，大規模言語モデル（Large Language Model: LLM）を実験協力者の代わりに用いることを検討する．LLM を用いて推薦システムを利用するユーザの行動をシミュレートする手法は，実際のユーザ実験の低コストな代替手段になる可能性として注目されている[5, 6]．既存研究ではユーザのアイテム評価行動を LLM に代替させることがほとんどであるのに対し，本稿ではインタフェースの操作を LLM に行わせ，その結果をインタフェースの評価・改善に利用する点で異なる．

*連絡先: 東京都立大学大学院 システムデザイン研究科
〒191-0065 東京都日野市旭が丘 6-6
E-mail: yamazaki-hironori@ed.tmu.ac.jp

図 1: 仮想ユーザを評価値別に提示するインタフェース

図 2: プロファイル文を表示した画面

LLM によるインタフェース評価の予備的検討として，インタフェースで提示する情報，および可能な操作をプロンプトで LLM に与え，提示された映画を視聴するか否かを判断するタスクをさせる実験を行った．LLM が判断の手がかりとした情報を分析し，インタフェース改善に有益な情報が得られるか，および評価方法の改善点について考察する．

2 関連研究

2.1 仮想ユーザ像を用いた情報推薦

稲田らは，推薦対象ユーザの情報を利用しない説明可能情報推薦を目的として，仮想ユーザ像を用いた情報推薦手法を提案している[3, 4]．ユーザ像はポジティブおよびネガティブな嗜好の説明を 3 文ずつの計 6 文から構成されるプロファイル文として生成し，人手による生成手法と LLM を用いた半自動生成手法を提案している．生成には匿名ユーザの行動履歴データを用いている．

図 3: 仮想ユーザの情報を確認できる画面

2.2 LLM を用いた推薦システムの評価

推薦システムを利用するユーザの行動を，LLM を用いてシミュレーションする研究が行われている．Agent4Rec[5]や RecAgent[6]は，ユーザのプロファイル，過去の行動，推薦システム内で行うことができる行動をプロンプトに入力し，ユーザと推薦システムのやり取りをシミュレートしている．実際の人間の行動をどの程度再現できるかについて検討し，嗜好を高い精度で反映させることや，フィルターバブル現象を再現できることを確認している．

3 提案手法

3.1 仮想ユーザ像を用いた情報推薦インタフェース

本節では，評価対象として想定する，仮想ユーザ像を用いたインタフェースについて説明する．このインタフェースでは，映画を対象アイテムとしており，提示された映画に対する評価値別に数名の仮想ユーザを提示する．対象アイテムに対し，高評価あるいは低評価をつける仮想ユーザの嗜好の傾向を把握し，自身の嗜好と比較することで，提示されたアイテムを選択するか否かを判断することを想定して

現在のページ

映画タイトル：Toy Story
概要ページのリンク：
IMDB
http://www.imdb.com/title/tt0114709/
allcinema
https://www.allcinema.net/cinema/28621
旬報キネマ
なし

仮想ユーザの情報
以下は提示された仮想ユーザの情報です。
仮想ユーザは評価値別に提示されています。
プロファイル文が未確認の仮想ユーザは、[未確認]と表記されます。

評価 5 をつけた仮想ユーザ
ユーザ名：D
プロファイル：個性的で挑発的な作品を評価：「Pulp Fiction (1994)」や「Fight Club (1999)」のように、斬新なストーリーテリングや強烈なテーマに惹かれる。
（中略）
このユーザが映画につけた平均評価値：3.46
高評価した映画（この仮想ユーザが特に好むと判断した映画）：
Night Tide
（中略）
低評価した映画（この仮想ユーザが好まないと判断した映画）：
Porky's II: The Next Day
（中略）

ユーザ名：U
プロファイル：[未確認]

評価 4 をつけた仮想ユーザ
ユーザ名：J
プロファイル：[未確認]

（中略）

評価 1 をつけた仮想ユーザ
なし

図 4: インタフェースが提示している情報の記述例

いる．

図 1 に，インタフェースのスクリーンショットを示す．映画のタイトル，概要ページのリンクと，1~5 の 5 段階の各評価値について最大 3 名の仮想ユーザの名前が表示される．概要ページとして，allcinema[1]，キネマ旬報 WEB[2]，IMDb[3]を利用する．仮想ユーザの名前をクリックすると，その嗜好を表すプロファイル文が表示される（図 2）．この時，仮想ユーザの名前を変更するボタンと仮想ユーザの情報を確認できる画面へのリンクが同時に表示される．仮想ユーザの名前を変更するボタンをクリックすると，仮想ユーザの名前を自由に変更することができる．この機能は，仮想ユーザの識別性を向上させることを目的としている．

仮想ユーザの情報を確認できる画面を図 3 に示す．この画面では，仮想ユーザの現在の名前，元の名前，平均評価値，プロファイル文，その仮想ユーザが高評価した映画と低評価した映画が表示される．仮想ユーザが高評価した映画と低評価した映画はランダムにそれぞれ 5 本選ばれる．

3.2 LLM を利用したインタフェース評価実験

LLM にインタフェースを操作させるため，インタフェースで提示する情報などをテキストとして LLM に入力し，実行する操作を出力させる．評価実験の手順は以下の通りである．

1. LLM にプロンプトを入力する
2. LLM が行動を選択して出力する
3. 選択された行動に従いプロンプトを更新し，1 に戻る

LLM に入力するプロンプトには，システムに関する説明，LLM が模擬するユーザの嗜好に関する情報，インタフェースが提示している情報，これまでに行った行動，思考プロセス，次に行うことができる行動を記述する．

インタフェースが提示している情報の例を図 4 に示す．図 1 に示す画面をベースとしており，映画のタイトル，概要ページのリンク，仮想ユーザの情報から構成される．仮想ユーザの情報は，評価値毎に仮想ユーザの名前とプロファイル文を記述するが，最初はプロファイル文は「未確認」とし，内容はプ

[1] https://www.allcinema.net/
[2] https://kinejun.jp/
[3] https://www.imdb.com/

思考プロセス

あなたが映画を視聴するかどうかを決定する前に、以下の思考プロセスを順に実行してください。
1. あなたの映画の嗜好と、提示された映画との適合性はどうですか？
2. 仮想ユーザのプロファイルと、彼らがこの映画にその評価をつけた理由は、あなたの意思決定にどのような影響を与えますか？
3. 未確認の仮想ユーザのプロファイルが視聴決定に影響を与えうると考えるなら、[CHECK_PROFILE]コマンドでその情報を確認することができます。
4. 確認した概要ページがある場合、該当する概要ページのリンク先の情報を意思決定に反映させることができます。あなたが確認した概要ページは、行動履歴に記載されています。まだ確認していない概要ページから映画の概要を確認したい場合は、[CHECK_MOVIE_INFO]コマンドで概要ページを確認することができます。
5. 総合的に判断し、最終的な視聴決定とその理由を導き出してください。
6. もし今後、あなたが特定の仮想ユーザを分かりやすく識別したいと考えるなら、仮想ユーザの名前を変更することができます。

図 5: 思考プロセスの記述例

ロンプトに含めない.「未確認の仮想ユーザのプロファイル文を確認する」という操作を実行した後に，その仮想ユーザのプロファイル文を含めて記述する.

思考プロセスの記述例を図 5 に示す．映画を視聴するか否かを決定する前に実行する思考プロセスを記述する．自身の嗜好と映画との適合性，仮想ユーザの情報を考慮し，必要であれば仮想ユーザのプロファイル文や概要ページを確認するように指示する.

次に行うことができる行動として，以下の 6 つの選択肢を提示し，LLM にどの行動を選択するかを出力させる.

- 映画の視聴を決定する.
- 映画を視聴しないことを決定する.
- 未確認の仮想ユーザのプロファイル文を確認する.
- 映画の概要ページを確認する.
- 特定の仮想ユーザが高評価・低評価した映画や平均評価値を確認する.
- 仮想ユーザの名前を変更する.

表 1: LLM が模擬するユーザの嗜好情報の例

項目	内容
人物像	スリラー・サスペンス愛好家
映画の嗜好	予想外の展開や緊張感のあるスリラー映画，心理的な駆け引きを描いた作品が好き.
好む映画	『セブン』 『シャッター アイランド』 『プリズナーズ』
好まない映画	恋愛映画やコメディ映画，ストーリー展開が単純な作品

未確認の仮想ユーザのプロファイル文を確認する行動は，仮想ユーザの名前をクリックしプロファイル文を表示する操作に対応する．前述の様に，この行動が選択された場合は，以降のプロンプトからプロファイル文は「未確認」ではなく内容を記述する.

映画の概要ページを確認する行動は，概要ページのリンクをクリックする操作に対応する．この行動が選択されると，行動履歴に概要ページを閲覧したことが追記される．図 5 に記載がある通り，行動履歴に概要ページを確認したことが記述されていれば，LLM はリンク先の情報を意思決定に反映させることができる.

特定の仮想ユーザが高評価・低評価した映画や平均評価値を確認する行動は，仮想ユーザの情報を確認できる画面への遷移に対応する．この画面で確認できる，仮想ユーザが映画につけた平均評価値，高評価，低評価した映画それぞれ 5 本が，図 4 のユーザ D のようにプロンプトに含まれるようになる.

仮想ユーザの名前を変更する行動は，該当機能を利用する操作に対応する. この行動を選択する際は，新しい名前を同時に出力させ，以降はその仮想ユーザは新しい名前でプロンプトに記述される.

映画の視聴可否の決定が選択された場合，インタフェース操作を終了し，インタフェース利用の感想，仮想ユーザの参考度，対象映画を視聴後の評価を出力するよう指示するプロンプトを入力する.

4 評価実験

4.1 実験概要

3 節で述べた手順に従い，LLM に提案インタフェースのプロトタイプを評価させる実験を行った．映

表 2: 視聴可否判断の根拠

模擬ユーザ	視聴する				視聴しない				計
	T1	T2	A1	A2	T1	T2	A1	A2	
概要	1	1	2	2	1	0	2	1	10
評価分布	0	1	1	0	1	1	0	0	4
高評価	1	1	3	3	4	2	2	0	16
低評価	0	1	0	0	1	0	0	1	3
判断材料不足					0	3	1	0	4
作品数	1	2	3	3	4	3	2	2	20

画に関する情報および仮想ユーザ像の生成には MovieLens の 1m データセット[4]を用いた. 仮想ユーザ像は稲田ら[3]の LLM を用いた半自動生成手法によって 22 名生成した.

LLM が模擬するユーザの嗜好情報の例を表 1 に示す. ChatGPT[5]を用いて生成し, 人物像, 映画の嗜好, 好む映画，好まない映画から構成される．「スリラー・サスペンス愛好家」を 2 名（T1，T2）と，「アニメ・ファンタジー好き」を 2 名（A1，A2）用意した．同じジャンルを好むユーザでも，嗜好の詳細や好む映画などが異なるように生成している．

模擬するユーザ 1 名につき 5 本の映画について実験を行った．映画はランダムに 5 本選んだが，好むジャンルが同じ 2 名については同じ 5 本の映画について実験した．

インタフェース操作を行う LLM は OpenAI[6]の GPT-5.2 モデルを使用した．

4.2 結果

模擬ユーザが選択可能な 6 種類の行動のうち，映画の概要ページの確認と仮想ユーザのプロファイル文の確認は，20 回の試行全てにおいて行われた．一方，仮想ユーザが高評価・低評価した映画や平均評価値の確認は 1 回しか選択されず，仮想ユーザの名前の変更は 1 回も行われなかった．

[4] https://grouplens.org/datasets/movielens/
[5] https://chatgpt.com/（2025 年 1 月 17 日，2026 年 1 月 16 日）
[6] https://openai.com/

4.2.1 視聴可否の判断理由

LLM が視聴するかどうかを判断した理由を分析したところ，判断の根拠は映画の概要，仮想ユーザの評価分布，高評価・低評価それぞれの仮想ユーザと自身の嗜好の比較に大別された．また，判断材料が不足しているため視聴しないと判断したとの回答もあった．これらの根拠が主な判断理由として挙げられた回数を，視聴する・しないと判断した場合毎に集計した結果を表 2 に示す．複数の理由を挙げた場合はそれぞれカウントしている．前述のとおり，仮想ユーザが評価した映画を確認する行動は観測されなかったが，プロファイル文を根拠とした場合でも，仮想ユーザが好む映画などの具体例に言及することはなかった．

表 2 より，映画の概要と，高評価をつけている仮想ユーザを根拠とするケースが多い結果となった．一方で，低評価した仮想ユーザが根拠となることは少なかった．また，映画によっては，提示された仮想ユーザの評価分布が，高評価あるいは低評価のどちらかのみの場合があり，視聴判断に影響を与えていた．さらに，判断材料不足は模擬ユーザ T2 によくみられた．

この結果から，高評価している仮想ユーザのプロファイル文は初めから表示し，仮想ユーザの特徴把握の手間を軽減するといった改善案が考えられる．仮想ユーザの評価が高評価・低評価のどちらかに偏る問題については，仮想ユーザ数を増やせば起こりにくくなると想定される．しかし，非常に多数の仮想ユーザを用いているにも関わらず評価に偏りがある場合には，どのような嗜好のユーザにとっても似た評価になる映画と考えられるため,「万人受けする映画」などの説明を記載することで視聴判断の助けになる可能性がある．また，自身の嗜好と比較できる材料が不足しているという理由が挙げられていたことから，インタフェースが提示する情報もしくは仮想ユーザ像が不足している可能性もある．自分の嗜好に関する手がかりを得られないという状況を起こさないためには，それぞれ異なる嗜好を持つ，多様な仮想ユーザ像を提示することが効果的と考える.

4.2.2 低評価仮想ユーザの有無

実験に用いた 10 本の映画のうち 5 本について, 提示された全ての仮想ユーザのプロファイル文が確認された．提示された仮想ユーザの一部のみプロファイル文を確認して視聴可否を判断したケースでは，低評価している仮想ユーザがいなかった．全ての仮想ユーザのプロファイル文が確認された 5 本の映画のうち，低評価している仮想ユーザがいなかった映画は 1 本であり，その場合は評価 3，4 の仮想ユーザしか存在しなかった．

表 3. 仮想ユーザのプロファイル文の極性と解釈

		高評価		低評価	
		P	N	P	N
好む要素	あり	6	0	0	0
	なし	8	0	1	3
好まない要素	あり	0	0	0	0
	なし	1	3	0	0
計		15	3	1	3

この結果から，評価が割れている場合は両方の意見を慎重に確認しているといえるが，4.2.1 節で述べた通り，低評価をつけている仮想ユーザは判断の根拠となることが少ない．この理由について考察するため，模擬ユーザが判断理由において，自身の好む・好まない要素が含まれるかどうかについて，仮想ユーザのプロファイル文を解釈した回数を集計した結果を表 3 に示す．表において，行は好む・好まない要素がプロファイル文に含まれるか否かに対応し，列は対象の映画を高評価・低評価した仮想ユーザに対応する．また，プロファイル文は仮想ユーザのポジティブ（P），ネガティブ（N）な嗜好の両方を記載しているため，それぞれ分けて集計している．

表 3 より，高評価をつけている仮想ユーザのポジティブな嗜好から，好む要素の有無を解釈する回数が最も多く，結果として視聴判断の手がかりとされることが多かったと考える．一方で，低評価をつけている仮想ユーザはネガティブな嗜好の方が参考にされており，好む要素を含まないという解釈に使われている．好まない要素が含まれているかどうかは概要ページで確認していたため，低評価ユーザを手がかりとすることが少なかったと考える．

模擬ユーザが言及したプロファイル文をみると，映画のジャンルや雰囲気を示す言葉を含む文(「スタイリッシュで知的な作品を評価」「奇抜なホラーや独創的な世界観に魅了される」等）や，特定の要素に対するネガティブなプロファイル文(「感傷的すぎる映画には厳しい」等）が言及されやすかった．一方で,「時代を超えたクラシック作品を評価」や「平凡な娯楽映画に興味が薄い」など具体性に欠ける賛辞・批判のプロファイル文や,「期待を裏切るシリーズ作に失望」など限定的すぎるプロファイル文は言及されなかった．

この結果から，映画のジャンルや雰囲気に直結しやすいポジティブな嗜好を持つ仮想ユーザや，好まない要素が明確な仮想ユーザが手がかりになりやすいといえる．従って，図 1 で仮想ユーザを評価値別に提示する際に，このような特徴を満たすプロファイル文を持つ仮想ユーザを強調表示するといった方法が考えられる．

5 おわりに

本稿では，仮想ユーザ像を用いた情報推薦インタフェースの評価に LLM を用いるアプローチを提案した．予備的検討として，提示した映画の視聴可否を LLM に判断させる実験を行った結果，高評価をつけている仮想ユーザの重要性や手がかりとなりやすい仮想ユーザの特徴などの知見が得られた．今後は，各仮想ユーザの特性・役割の違いについて詳細に分析する他，本実験で得られた知見に基づいて，提案するインタフェースの評価実験を行う予定である．

謝辞

本研究の一部は JSPS 科研費 22K19836，23K24953 の助成を受けたものです．

参考文献

[1] F. Ricci, L. Rokach, B. Shapira: Introduction to Recommender Systems Handbook, Spriner, pp. 1-35, 2010

[2] Y. Zhang, X. Chen: Explainable Recommendation: A Survey and New Perspectives, Foundations and Trends in Information Retrieval, Vol. 14, No. 1, pp. 1-101, 2020

[3] 稲田真樹人，柴田祐樹，高間康史: 仮想ユーザ像を用いた説明可能情報推薦にむけた予備的検討，第 39 回ファジィシステムシンポジウム講演論文集，No. 39, pp. 553-558, 2023

[4] 高間康史，柴田祐樹: 仮想ユーザ像を用いた情報推薦システムにおける大規模言語モデルを用いたプロファイル生成の検討，第39回人工知能学生全国大会, No. 1E4-OS-3a-04, 2025

[5] A. Zhang, Y. Chen, L. Sheng, X. Wang, T. S. Chua: On Generative Agents in Recommendation, 47th International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 1807-1817, 2024

[6] L. Wang, J. Zhang, H. Yang, Z. Y. Chen, J. Tang, et al.: User Behavior Simulation with Large Language Model-based Agents, ACM Transactions on Information Systems, Vol. 43, No. 2, Article 55, pp. 1-37, 2025

Latent-Explorerを用いた 論理推論問題に対するLLM推論過程の分析

Analysis of LLM Reasoning Processes for Logical Tasks using Latent-Explorer

木村五郎[1] *　尾崎知伸[1]
Goro Kimura[1], Tomonobu Ozaki[1]

[1] 日本大学 文理学部 情報科学科
[1] Department of Information Science, Nihon University

Abstract: While Large Language Models (LLMs) excel in logical tasks, evaluations based solely on final outputs remain insufficient for understanding their internal reasoning processes. To address this black-box nature, this study investigates the mechanisms of knowledge formation and updates during logical inference. We employ Latent-Explorer, an activation-patching-based method, to extract layer-wise propositions from Llama-2-7b on the LogicQA and RuleTaker datasets. By constructing dynamic knowledge graphs, we analyze internal reasoning profiles through multi-faceted metrics, including conclusion emergence, stability, and convergence. Our findings quantitatively clarify how LLMs transform propositional knowledge across layers to reach final conclusions.

1 はじめに

Transformer アーキテクチャに基づく大規模言語モデル（Large Language Model, LLM）は，自然言語処理タスクにおいて飛躍的な性能向上を遂げ，機械翻訳や質問応答，文章生成など多様な分野に応用されている．また近年では，LLM の論理推論タスクへの適用も注目されており，正答率などの外部的な基準を用いた能力評価が行われている [1, 2]．しかし，論理的帰結の一致のみに着目した評価は，論理推論能力の評価としては必ずしも精密ではなく，推論過程を含めたより詳細な評価手法の確立が求められている．これに対し現状は，LLM 内部の推論過程，すなわち LLM がどの様なタイミングでどの様な知識を想起し，またそれらをどの様に組み合わせることで帰結を形成しているのかを，外部から把握することは容易ではなく，詳細評価の一つの障壁となっている．

これらのことを背景に，本研究では，ブラックボックスとされる LLM の内部推論メカニズムを段階的に評価するための初期的な試みとして，論理推論タスクにおける推論過程を可視化・分析し，その本質や原理の解明を試みる．具体的には，論理推論タスクに関するデータセット LogicQA[2] および RuleTaker[3] のそれぞれに対し，アクティベーション・パッチング（Activation Patching）[4] に基づく動的グラフ生成手法である Latent-Explorer[5] を適用することで，推論過程における各層の出現命題を抽出する．さらに得られる命題集合を分析することで，結論の形成時期や安定性，探索から収束への推移，相関の変化量など，多角的な指標により内部推論過程の分析を行う．これらを通じ，LLM が層ごとにどのような命題レベルの知識を構築・変容させ，最終的な結論へと収束させるのか，その内部推論過程を定量的に明らかにする．

本論文の構成は以下のとおりである．2 章で関連研究について述べる．3 章で，本研究における基礎技術である Latent-Explorer について概説する．次に 4 章で，論理タスクに対する LLM の推論過程を分析する枠組みを提案し，5 章でその評価実験の結果を示す．最後に 6 章でまとめを行い，今後の課題を述べる．

2 関連研究

LLM の内部推論過程はブラックボックスであり，モデルがどのように推論を行い，結論に至るかを明確に理解する手法は限られている．近年この課題を解決するために，LLM 内部の推論過程の可視化や知識の参照過程の追跡に関する研究が活発に行われている．

Heimersheim ら [4] は，アクティベーション・パッチング（AP 法）と呼ばれる，特定のプロンプトに対す

*連絡先：日本大学 文理学部 情報科学科
〒 156-8550 東京都 世田谷区 桜上水 3-25-40
E-mail: chgo22004@g.nihon-u.ac.jp

るモデル隠れ層の活性状態を別のプロンプトに転写することで推論の流れや参照知識を解析する技術を用い，モデル内部の知識参照の可視化手法を提案した．これにより，推論中にどの知識が呼び出されたかを部分的に明らかにすることが可能となった．

Bronzini ら [5] は，LLMs 内部の潜在表現を動的に追跡する手法として，AP 法を基礎とした動的知識グラフ構築手法 Latent-Explorer（LE 法）を提案した．詳細は後述するが，この手法は，モデルが推論過程において参照する知識の変化をグラフ構造として捉え，時系列的な推移を視覚化する．またその具体的な応用として，真偽判定タスクにおける複雑な命題推論の流れを可視化し，モデル内部における知識の変化を明らかにした．しかし現状，真偽判定以外のタスクに対する応用は確認されておらず，異なる系統の問題に対する推論過程の解明に関して検討の余地が残されている．

Jiang ら [6] は，数理推論に対する LLM の精度を高めるため，前向き推論と後ろ向き推論を統合した FORBAR 法を提案した．この手法は，推論の前方展開と逆方向の検証を組み合わせ，論理的な整合性を確認するものである．現時点では，モデル内部でどの知識がどのタイミングで参照されたかを段階的に可視化するまでには至っていないが，特に，数学的な証明に対して高い精度を発揮することが確認されている．

これらの研究ではいずれも，LLM の内部状態を示す情報を獲得することは可能であるが，多段推論における知識の時間的変化の詳細な追跡など，ステップごとの変化に対する形式的・定量的な解析までは行われていない．しかし実際には，複雑な数学的または論理的タスクを対象とした場合，モデルは複数の命題や事実を段階的に結び付けながら推論を進める必要があり，この過程を正確に観測・記録することは，推論メカニズムの解明のために不可欠であると言える．

3 Latent-Explorer

Latent-Explorer（LE 法）[5][1]とは，LLM モデルの各隠れ層に対するアクティベーション・パッチング（AP 法）の適用結果を集約することで，その推論過程を動的グラフとして抽出・可視化する手法である．

AP 法は，以下の手順に従い，L 層からなる LLM モデル $\mathcal{M}$ がタスク T を解く際における第 l 隠れ層での推論状況を自然言語表現 $o^{(l)}$ として抽出する（図 1 参照）．

1. LLM モデル $\mathcal{M}$ に，タスク T を伴うソースプロンプト $P_{\mathcal{S}}$ を表すトークン列 $X^{\mathcal{S}}(T) = x_1^{\mathcal{S}}, \cdots, x_m^{\mathcal{S}}$ を与え，タスク（推論）を実行する．

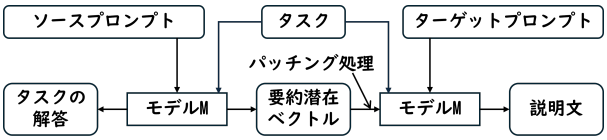

図 1: アクティベーション・パッチングの処理フロー

2. $P_{\mathcal{S}}$ によって事前に既定されるトークン群を対象に，推論の過程で得られる第 l 層における第 t トークン $x_t^{\mathcal{S}}$ の潜在ベクトル表現 $h_t^{(l)}$ を

$$\overline{h^{(l)}} = \sum_t w_t h_t^{(l)} \,/\, \sum_t w_t$$

のように集約し，第 l 層の要約潜在ベクトル $\overline{h^{(l)}}$ を算出する．ここで w_t は $x_t^{\mathcal{S}}$ に対する重みであり，名詞や動詞などのエンティティや関係を表すトークンには大きな重みを，それ以外のトークンには小さい重みを与えることで，重要性を反映した要約表現を獲得する．

3. 再び $\mathcal{M}$ に，T を伴うターゲットプロンプト $P_{\mathcal{T}}$ を表すトークン列 $X^{\mathcal{T}}(T) = x_1^{\mathcal{T}}, \cdots, x_n^{\mathcal{T}}$ を与え，タスクを実行し，結果である自然言語表現 $o^{(l)}$ を得る．なおこのとき，第 l 隠れ層における特定位置の潜在ベクトル表現を，要約潜在ベクトル $\overline{h^{(l)}}$ に置き換える処理（パッチング処理）を行った上で，推論処理を継続・完遂する．このパッチング処理により，結果である $o^{(l)}$ に，$\overline{h^{(l)}}$ で表現される推論の内部状態が反映される．

LE 法は，AP 法の後処理として，テンプレートに基づく命題表現抽出処理を $o^{(l)}$ に適用し，命題集合 $P^{(l)} = \{p_1^{(l)}, \cdots, p_{N_l}^{(l)}\}$ の抽出と，$P^{(l)}$ に基づく知識グラフ $G^{(l)}$ の構築を行う．さらにこの処理を全隠れ層に適用することで，層位置 $l \in \{1, \cdots, L\}$ をタイムスタンプとする動的知識グラフ $G = G^{(1)}, \cdots, G^{(L)}$ を導出する．

以上概観した通り，LE 法は，AP 法の適用も含めたこれら一連の処理を通じ，LLM モデル $\mathcal{M}$ におけるタスク T の推論過程，すなわち時間軸（層方向）に沿って知識がどのように追加・変形されていくかを動的知識グラフ G として抽出・可視化する．

4 LE 法の論理推論タスクへの適用

本研究では，LE 法を用いた論理推論タスクに対する LLM 推論過程の分析に関し，以下の 2 点について検討を行う．

1. タスク構造に即したソースプロンプトの設計
2. 動的グラフを対象とした多角的な評価指標の策定

[1] https://github.com/Ipazia-AI/latent-explorer

4.1 ソースプロンプトの設計

LLM へ与えるソースプロンプト P_S は，モデル内部で形成される潜在表現の構造および抽出される推論過程に直接的な影響を及ぼす．本研究の目的は，外的性能の最大化ではなく，内部推論過程の比較・可視化である．従って，ソースプロンプトには再現性と比較可能性を担保するための設計上の妥当性が求められる．これに基づき，以下の指針に従って設計を行うこととする．

- タスク構造の忠実な反映：元の問題が持つ構造を保ったまま提示する．
- 不要ヒントの排除：正解ラベルや解き方を直接示すような追加文は含めず，同タスクを人間が解く際に与えられる以上の情報を与えない．
- 形式的一貫性：同一タスク内では，すべての問題に対して同一の指示文を用いる．これにより，プロンプト構造の変化による影響を抑え，問題ごとの差は内容そのものに起因するようにする．
- 役割指示：モデルには “論理推論の専門家” として回答する役割のみを与え，中間的な思考ステップの形式や出力スタイルを細かく制約しない．

4.2 評価指標の策定

動的グラフ・命題集合として抽出される推論プロセスを定量的に評価するため，結論導出の速さや安定性，考慮知識の収束率，および層間での知識の更新量に着目し，以下の多角的な評価指標を導入する．

最終帰結初出層： LLM の思考速度を評価することを目的に,「タスクに関するすべての帰結が出現するまでに必要とされた層の数」を評価基準とする．形式的には，タスク T の正答に対する命題表現集合を P_T，L 層モデル $\mathcal{M}$ の T に対する命題集合系列を $P = P^{(1)}, \cdots, P^{(L)}$ とし，最終帰結初出層を次のように定義する．

$$\max_{p \in P_T} \min_{l \in \{1,\cdots,L\}} \{l \mid p \in P^{(l)}\}$$

帰結安定層： 最終帰結初出層が「いつ答えに辿り着いたか」を示すのに対し，帰結の安定性，すなわち LLM が「いつ答えを確信し，維持し始めたのか」を評価することを目的に,「全帰結が最終層まで一貫して出現し始めた最小の層」を評価基準とする．形式的には，次のように定義する．

$$\min_{s \in \{1,\cdots,L\}} \forall_{l \in \{s,\cdots,L\}} P_T \subseteq P^{(l)}$$

推論圧縮率： 推論の洗練度，すなわち推論の途上で生成された膨大な中間仮説や候補（探索の広がり）から，最終的な結論に向けていかに不要な情報を削ぎ落とし，論理を収束させたかという情報の整理能力を定量化することを目的に,「（正規化した）最大層内命題数と最終層内命題数の差」を評価基準とする．形式的には，以下のように定義する．

$$\min_{l \in \{1,\cdots,L\}} \left(1 - \frac{|P^{(L)}|}{|P^{(l)}|}\right)$$

層間知識更新率： 各層における推論の活動量を測定することを目的に,「隣接する層間で行われた知識グラフの正規化更新量」を評価基準とする．形式的には，隣接する層間における導出命題集合の Jaccard 距離と定義する．

$$1 - \left(\frac{\left|P^{(l)} \cap P^{(l-1)}\right|}{\left|P^{(l)} \cup P^{(l-1)}\right|}\right)$$

5 評価実験

5.1 実験設定とソースプロンプト

対象モデルとして Meta 社によるチャット向け事前学習モデル `meta-llama/Llama-2-7b-chat-hf` を採用し，評価実験を行う．またデータセットとして，読解型多肢選択問題を扱う LogicQA[2][2]および伴意判定問題を扱う RuleTaker[3][3] を利用する．以下，各データセットの詳細と，LE 法で利用するソースプロンプトについて説明する．

LogicQA： このデータセットは，読解・論理推論問題から構成されるデータセットであり，短い文章（context）と質問文（question），4 つの解答選択肢（options）から構成される．文章中の事実や因果関係を基に正しい選択肢を推定することが求められ，多段の推論や条件の組み合わせが必要となる問題が含まれる．本タスクに対するソースプロンプトを図 2 に示す．

RuleTaker： このデータセットは，問題文（context）と質問文（query）から構成されるデータセットであり，質問文が伴意される場合は ENTAILMENT を，そうでない場合は NOT ENTAILMENT を出力することが求められる．本タスクに対するソースプロンプトを図 3 に示す．

[2] https://github.com/lgw863/LogiQA-dataset
[3] https://github.com/allenai/ruletaker

```
You are an expert in logical reasoning. Read the passage carefully and answer the question
by choosing one option from (A), (B), (C), (D). Answer only with the letter of the correct option.

Example 1: Passage: Tom had three apples. He gave one apple to Mary and one apple to John.
Question: How many apples does Tom have now?
Options: (A) 1 (B) 2 (C) 3 (D) 4
Answer: A

Example 2: Passage: Alice is older than Betty. Betty is older than Chris.
Question: Who is the youngest person?
Options: (A) Alice (B) Betty (C) Chris (D) None of the above
Answer: C

Now solve the following problem.
Passage: [context of the target problem]
Question: [question of the target problem]
Options: (A) [option A] (B) [option B] (C) [option C] (D) [option D]
Answer:
```

図 2: LogicQA に対するプロンプト

```
You are an expert in logical reasoning. Read the context carefully and determine whether the question
is entailed by the context. Answer only with one of the following labels: ENTAILMENT or NOT_ENTAILMENT.

Example 1:
Context: Tom is tall. Tall things are strong.
Question: Tom is strong.
Answer: ENTAILMENT

Example 2:
Context: Alice is older than Betty. Betty is older than Chris.
Question: Alice is the youngest person.
Answer: NOT_ENTAILMENT

Now solve the following problem.
Context: [context of the target problem]
Question: [question of the target problem]
Answer:
```

図 3: RuleTaker に対するプロンプト

5.2 定量評価

各データセットから 30 問ずつ，計 60 問を選定して定量評価を行った．問題の選定にあたっては，入力長が極端に長い問題は除外し，モデルが安定して推論可能な範囲の問題を選択した．具体的には，LogicQA については，選択肢数や問題形式が標準的な問題を選び，特殊なフォーマットを含むものは除外した．一方 RuleTaker については，多数の推論ステップが必要とされる難易度の高い問題は対象外とした．

実験結果を表 1 にまとめる．なお，表中の各値は全 30 問の平均値である．また層間知識更新率に関しては，全層を前段 (02–08 層)・中段 (09–16 層)・後段 (17–24 層) の 3 つのフェーズに分割して集計した．加えて，後段層に関しては，詳細分析のために，タスクの正誤別での結果も算出した．以下，各評価基準の観点から考察を行う．

正答率： 正答率に関しては，LogicQA が RuleTaker を下回る結果となった．これは，ルールに基づく推論が核となる RuleTaker に対し，LogicQA は複数の選択肢から妥当な結論を導出するために広範な探索を要するという，タスク難易度の差を反映しているものと考えられる．

帰結の導出： 帰結の導出プロセスにおいて，RuleTaker は LogicQA よりも最終帰結初出層・帰結安定層ともに小さな値を示し，両指標の乖離（安定までの遅延）も小さい．これは，RuleTaker が問題文から中間命題を導出すれば結論が定まりやすいという特性を持ち，必要な命題集合が比較的浅い層で揃うためと推察される．対照的に，LogicQA は選択肢を検証するための条件探索を伴うため，結論の到達および安定までにより多くの層を要する傾向が伺える．

表 1: 評価結果

指標	LogicQA	RuleTaker
正答率	0.73	0.90
最終帰結初出層	16.27	9.10
帰結安定層	20.07	11.70
推論圧縮率	0.676	0.479
最大層内命題数	38.47	23.90
層間知識更新率（02–08）	0.447	0.383
層間知識更新率（09–16）	0.206	0.115
層間知識更新率（17–24）	0.236	0.103
層間知識更新率（17–24 正）	0.214	0.091
層間知識更新率（17–24 誤）	0.298	0.215

推論圧縮率： LogicQA は最大層内命題数が大きく，推論圧縮率も高い値を示した．これは，推論の過程で膨大な中間命題を一時的に生成（拡散）し，その後，結論に不要な情報を大幅に削除（収束）するという動的な情報整理が行われていることを示唆している．一方，RuleTaker は最大層内命題数が小さくまた圧縮率も高くなく，早期から必要な情報のみに絞られた効率的な推論が行われていることが伺える．

層間知識更新率： RuleTaker は前段で命題集合が大きく更新されるものの，中段以降は書き換え割合が減少し，推論過程が早期に定常状態へと移行していることが示唆される．これに対し LogicQA は，前段から大きな更新が続くほか，後段においても相対的に高い値が維持される．これは，選択肢の妥当性を保証するための条件探索や候補の切り替えが終盤まで継続し，命題集合が動的に組み替わりやすいタスクの特性を示していると考えられる．

また正誤別の考察では，両タスクに共通して，誤答時に後段層での層間知識更新率が上昇する傾向が確認された．これは，最終段階においても命題集合が大きく入れ替わり続け，推論が安定状態へ十分に収束していないことを示している．特に RuleTaker では，正答時後段の値が 0.091 と極めて低いのに対し，誤答時は 0.215 まで値が増大している．LogicQA においても誤答時の更新率は高まるが，正答時でも一定の値を示すことから，タスク自体が本質的に後段までの試行錯誤を伴いやすい性質を持つと推察される．

5.3 ケーススタディ

前節までの定量的な分析結果を補完するため，本節では LogicQA および RuleTaker の代表的な問題に対する LLM の推論過程を可視化し，定性的な考察を行う．図 4 に LogicQA の問題と得られた動的知識グラフの例を，図 5 に Ruletaker の問題と得られた動的知識グラフをそれぞれ示す．

図 4 より，動的グラフの初期層では，まず “Cantonese some not like chili” といった，文脈から抽出された主要な事実構造が形成されていることが分かる．また中段層に至ると，前提命題と結合可能な論理的橋渡しの候補として，各選択肢に対応する命題がグラフ内に組み込まれ，それらの妥当性が探索される．後段層では，結論導出に寄与しない不要な候補命題が順次削除され，最終的に前提命題，Option C の規則，および結論に関連する最小限の命題集合へと収束する．その結果，最終層において Option C が選択・出力されている．以上のプロセスから，選択肢型の論理タスクにおいて LLM は，前提と結論の間を埋める中間命題を適応的に探索し，後段にかけて必要な論理パスへと情報を絞り込むことで回答を導出していることが見て取れる．

一方，図 5 においては，動的グラフの初期層で “Bob is kind” や “Bob is young” といった入力文に基づく事実命題が形成されていることが分かる．中段層では，関連するルールが活性化されることで，“Bob is kind” と “Bob is young” の同時成立から “Bob is not smart” が導出されるなど，結論の判定に直接関与しない命題も一時的に増加する．これは，与えられた複数のルールを逐次的に適用し，推論候補を拡張している過程を反映している．後段層に入ると，最終的な判断に不要な命題は順次削除され，結論に直結する命題のみが保持されるようになる．最終的には “Bob is kind” を含む最小限の命題集合へと整理され，最終層において “ENTAILMENT” が出力される．以上のプロセスから，結論の根拠が入力に含まれる比較的単純なタスクにおいても，モデルは早期に必要命題を特定するだけでなく，層を通じた探索過程で一時的に派生命題を生成（拡散）し，その後に必要な論理パスへと収束させるという動的な推論プロセスを経ていることが見て取れる．

6 結論

本研究では，Latent-Explorer を用いた論理推論タスクに対する LLM 推論過程の分析を目的とし，タスク構造を反映したソースプロンプトの設計および動的グラフを対象とした多角的な評価指標の提案を行った．また論理推論に関する 2 つのデータセット（LogicQA, RuleTaker）に提案手法を適用し，内部推論の過程を定量的・定性的に評価した．

実験の結果，本手法は「帰結がいつ形成・安定し，推論過程でどの程度探索的に情報が拡散し，層間でいかに書き換わりながら収束するか」という，推論過程の時間的・空間的構造をタスク間や正誤間で比較可能にすることを示した．具体的な知見としては，RuleTaker

問題文: Some Cantonese don't like chili, so some southerners don't like chili.

質問文: Which of the following can guarantee the above argument?

選択肢: (a)Some Cantonese love chili. (b)Some people who like peppers are southerners. (c)All Cantonese are southerners. (d)Some Cantonese like neither peppers nor sweets.

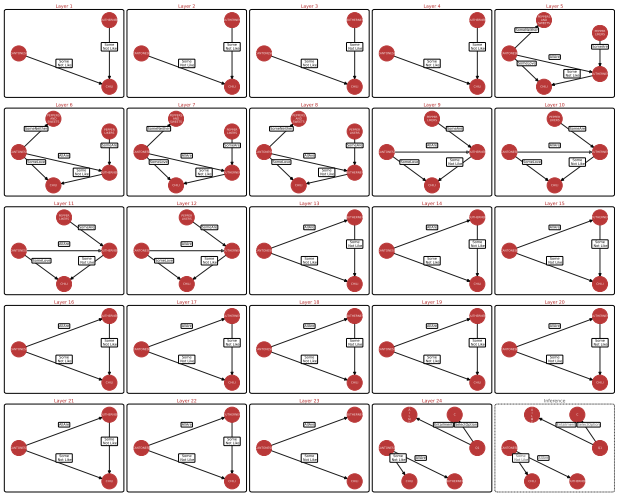

図 4: LogicQA の問題と得られた動的知識グラフの例

問題文: Anne is quiet. Anne is not young. Bob is kind. Bob is young. Dave is rough. Dave is round. Dave is smart. Dave is not young. Fiona is quiet. Fiona is not round. Kind, young things are not smart.

質問文: Bob is kind.

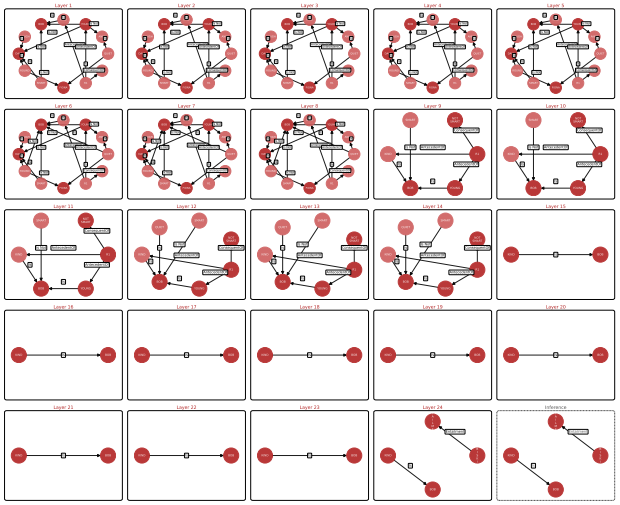

図 5: RuleTaker の問題と得られた動的知識グラフの例

では比較的早期に必要命題が揃うのに対し，LogicQA では未既知の橋渡し探索を反映して，後段まで命題の動的な組み替えが継続するといった，問題の系統性に起因する推論過程の差異を定量的に明らかにした．また，誤答時には最終層付近でも命題集合が不安定なまま推移する現象を確認しており，これはモデルの「迷い」を内部指標から検知できる可能性を示唆している．

今後の課題として，本実験は各 30 問という小規模なサンプルに基づいているため，より大規模かつ多様なデータセットおよび異なるモデル規模への拡張が必要である．また，命題抽出プロセスがモデルの表現揺れや変換規則の影響を少なからず受けることから，観測された推論過程が純粋な推論の変化なのか，あるいは LLM 由来のノイズであるかを厳密に峻別する手法の検討が求められる．今後は，誤答タイプの体系化や詳細な分類を進めることで，知識グラフに基づく内部推論プロセス分析を，LLM の信頼性を担保するための比較・検証基盤へと発展させたいと考えている．

参考文献

[1] X. Wu, Y. Cai, and H.-F. Leung : Abstract-level Deductive Reasoning for Pre-trained Language Models, The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, pp.70–76, 2024.

[2] J. Liu, L. Cui, H. Liu, D. Huang, Y. Wang, and Y. Zhang : LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning, The 29th International Joint Conference on Artificial Intelligence, pp.3622–3628, 2020.

[3] P. Clark, O. Tafjord, and K. Richardson : Transformers as Soft Reasoners over Language, The 29th International Conference on International Joint Conferences on Artificial Intelligence, pp.3882–3890, 2020.

[4] S. Heimersheim and N. Nanda : How to use and interpret activation patching, arXiv:2404.15255, 2024.

[5] M. Bronzini, C. Nicolini, B. Lepri, J. Staiano, and A. Passerini : Unveiling LLMs: The Evolution of Latent Representations in a Dynamic Knowledge Graph, arXiv:2404.03623, 2024.

[6] W. Jiang, H. Shi, L. Yu, Z. Liu, Y. Zhang, Z. Li, and J. Kwok : Forward-Backward Reasoning in Large Language Models for Mathematical Verification, *Findings of Annual Meeting of the Association for Computational Linguistics*, pp.6647–6661, 2024.

ニュースのハード・ソフト分類における判断の齟齬とその要因分析

An Analysis of Factors Underlying Inter-Annotator Disagreement in Hard/Soft News Classification

杉本 麻衣 † 藤代 裕之 ‡ 松下 光範 ‡
Mai Sugimoto Hiroyuki Fujishiro Mitsunori Matsushita

† 関西大学 総合情報学部
Faculty of Informatics, Kansai University

‡ 法政大学 社会学部
Faculty of Social Sciences, Hosei University

Abstract: ニュース報道のハード・ソフト区分は，記事が読者に与える影響を解明する不可欠な指標である．しかし，この区分の分類基準は定義が曖昧であり，一貫性を保つことが困難である．そこで本研究は，判断の段階的な分解と根拠箇所の記録・可視化を可能にするアノテーション支援ツールを開発し，分類の齟齬が生じる原因を調査した．その結果，不一致の主因が，作業者の能力不足ではなく，専門家の暗黙知と言語化された定義とのズレや，記事の記述構造との不適合にあることを明らかにした．

1 はじめに

現代のデジタルニュース空間においては，記事の記述構造に大きな変化が生じている．かつての新聞報道は，最も重要な事実や結論を冒頭のリード文に配置する「逆三角形型」の構造が基本とされ，読者へ知識や情報を効率的に伝達することに重きが置かれていた [12]．しかし，インターネット以降のデジタルニュース空間においては，単なる客観的な事実の伝達にとどまらず，読者の関心や「共感」を強く惹きつける手法が求められるようになっている．より多くの読者を獲得し記事を拡散させるために，事象の背後にある文脈や特定の個人の感情・エピソードといった細部を描き出し，読者に追体験させるような「物語型」の手法の重要性が高まっている [10]．

このようなニュースの伝え方が読者に与える影響を分析するためのアプローチとして，メディア研究においては報道内容の「伝え方」の違いを「ハードニュース」と「ソフトニュース」に分類する枠組みが存在する [8]．しかし，従来の研究では，定義自体が曖昧であり，トピック（政治か芸能・スポーツか）による一次元的な分類に依存していた．そのため，読者に与える効果については一貫した結論が得られていなかった．この限界を克服するため，Reinemann ら [6] はニュース報道を単一の基準で二分するのではなく，「トピック」「フォーカス」「スタイル」という 3 つの次元で尺度化して分類する枠組みを提唱した．大森 [11] によるこの枠組みを用いた実証研究では，「フォーカスのソフト化」は，複雑な政治的問題が個人の生活にどう関わるのかを分かりやすく解説することで「政治の複雑さ」を解消し，政治関心や内的有効性感覚を高めるポジティブな効果を持つことが明らかになった．さらに，「スタイルのソフト化」は，過剰になると政治家等に対する外的有効性感覚を低下させ，政治不信を助長する副作用をもたらすことも示されている．

このように，ニュース報道の多次元的なハード・ソフト区分は，メディアの報じ方が読者の政治意識に与える影響を要因別に解明するための不可欠な指標として位置づけられる．しかし，この指標を用いた客観的な分析を行うためには，実際のニューステキストに分類枠組みを適用し，データとして分類する必要がある．テキストに対して分類枠組みを適用し，ラベルやカテゴリの情報を付与する作業は「アノテーション（あるいはコーディング）」と呼ばれる．Ziems ら [9] が指摘するように，この作業において特定の理論的構成概念をテキストに適用する際，分類基準の定義が曖昧になりやすく，客観的な判断の一貫性を保つことが困難になるという特有の難しさが存在する．また，Krippendorff[4] が述べるように，コーディングは単なる物理的なキーワードの測定ではなく，観測されたテキストを手がかりとして，背後にある

* E-mail: k570068@kansai-u.ac.jp
† E-mail: fujisiro@hosei.ac.jp
‡ E-mail: m_mat@kansai-u.ac.jp

観測できない文脈やニュアンスを読み解く，高度な推論を伴う解釈的なプロセスである．本研究で扱うニュースの「フォーカス」や「スタイル」の判定においても，記事の焦点が社会と個人のどちらに向いているか，あるいは表現にどの程度感情や個人的見解が含まれているかといった，文脈依存的な推論が必要となる．そのため，前後の文脈や作業者の主観によって必然的に解釈の揺らぎが生じる．その結果，作業者が持つ一般的な語彙の理解と，専門家が想定する厳密な定義との間にギャップが生じやすく，その解釈が作業者の主観に委ねられることで判断基準の齟齬が発生しやすい性質を持っている．さらに，「物語型」の記事構造は，客観的な事実と主観的な感情や個人のエピソードが複雑に入り混じるため，多次元的な分類基準を適用する際の判断を一層困難にしていると考えられる．

そこで本研究では，従来の「逆三角形」構造の記事と，現代的な「物語（ナラティブ）」構造の記事の双方に対してアノテーションを実施し，評価の一貫性を保つことが難しい原因を比較・調査した．その検証にあたり，複雑な分類作業を小さな判断ステップに順次分解し，「記事のどの記述を根拠としてその判断を下したのか」という思考プロセスを記録・可視化するアノテーション支援ツールを開発した．このツールを用いて分類作業の実験を行うことで，最終的なラベルのみのアノテーションではわからない「判断のどの段階で」「記事のどのような表現によって」判断の齟齬が生じたのかを詳細に特定することができる．これにより，分類の一貫性を阻害する要因が，作業者の単なる能力不足によるものなのか，専門家が持つ暗黙知と言語化された定義とのズレにあるのか，あるいは現代のネット記事が持つ特有の記述構造との不適合にあるのかを多角的に調査し，より信頼性の高い分類枠組みを構築するための知見を提示する．

2 関連研究

複数の作業者がアノテーションタスクを行う際の一致性は，質的研究や内容分析における関心課題の一つである．Lombard ら [5] は，アノテーションタスクにおいて作業者間で生じる不一致の要因として「基準の曖昧さ」「コードの定義の不十分さ」「作業者の訓練不足」を挙げている．また，Guest ら [3] は，作業仕様書の継続的更新や訓練手続きの明示化が一致率向上に寄与することを示す一方で，作業者にかかる人的負担の大きさも指摘している．

こうした手作業の制約を克服するための支援技術の必要性の高まりに伴い，アノテーションタスクを大規模言語モデル（Large Language Model; LLM）を用いて行うことも検討されている [7]．一部のタスクでは良好な結果が報告されているが，専門的なアノテーション付与タスクを LLM で代替することには課題が残る．Ziems ら [9] は，25 の代表的なタスクを用いて LLM の性能を評価し，その限界を指摘している．これらのタスクは日常的な用語に専門的，あるいは非標準的な定義を適用する必要があり，LLM が事前学習で獲得した一般的な意味論とは異なる非慣習的な言語理解が求められるためである．特定の理論的枠組みに基づいた文脈依存的なニュアンスを正確に捉えることは依然として計算機には難しく，信頼性の高いデータセットを構築するには人間の解釈に基づくアノテーション付与が不可欠である．

また，アノテーションにおける作業者間の齟齬は，排除すべきノイズとして従来は扱われてきたが，Aroyo ら [1] は，その齟齬をタスクの曖昧さや作業者の多様な視点を反映した重要な手がかりであると指摘している．同様に，Cabitza ら [2] は機械学習用のデータセット構築に関わるアノテーション作業において，齟齬を排除するのではなく，判断の多様性を保持し，複数の視点を正解データの構築プロセスに統合するプロセスを提案している．このアプローチを採ることは，予測能力の向上だけでなく，モデルの解釈可能性や公平性向上にも貢献する可能性があるとしている．

3 タスクの定義と判断指標

本研究では，Reinemann らの枠組み [6] をもとに大森が翻訳・整理した指標 [11] を，アノテーションタスクを実施する際の判断基準とした．以下に，大森が提示する各側面の判断指標を示す．

トピック: ニュース項目の政治的関連性

1. 2 つ以上の政治的アクターが登場するか
2. 立法・行政・司法といった意思決定機関が登場するか
3. ニュースの取り上げる主題・問題に対し実現された政策的決定や措置プログラムに言及するか
4. ニュースの取り上げる主題・問題に対し実現された政策的決定や措置プログラムに関係する個人やグループが登場するか

フォーカス: ニュース項目の焦点に着目した分類

1. **「個人ー社会」との関連度 (F1)**：ニュースの内容の帰結が，個人の生活などのミクロな範囲に関連するものか，社会全体の問題に関連するものか

2. **「エピソードーテーマ」フレーム度 (F2)**：ニュース内容が，特定の個人のエピソードに着目するものか，より広範な問題のテーマに着目するものか

スタイル: ニュース報道で用いられる様式

1. **「個人ー非個人」的なリポート度 (S1)**：フォーカス面の「個人ー社会」関連度とは異なり，ニュースにリポーターや解説者，あるいはゲストといった個人の見解が含まれているか
2. **「感情ー非感情」的なリポート度 (S2)**：従来の戦略型フレーム報道やソフトニュースの研究でたびたび注目されてきた，戦いに関連するような言葉や表現を用いているか

本研究では，解釈の揺らぎが生じやすい「フォーカス」と「スタイル」の 2 次元（4 項目）に焦点を当ててアノテーションタスクを設計する．

4 実験

本稿では，判断の齟齬が生じる箇所を詳細に記録可能なアノテーション支援ツールを用いてアノテーション付与の過程を記録し分析する．

4.1 アノテーション支援ツール

本実験で利用するアノテーション支援ツールは，アノテーションタスクの遂行に当たり，作業者の認知プロセスを支援しつつ，作業者間の判断の齟齬が生じる箇所を詳細に把握することを目的としたツールである [13]．一般的なアノテーションでは，作業者の最終ラベルのみが記録されるため，どのような齟齬が生じたのかは把握できても，それがどのような判断や手続きによって生じたのかを特定することが難しい．この課題に対し，本ツールは，作業者が判断の根拠となったテキスト箇所の選択や理由記述を通じて，齟齬の原因と発生箇所の特定を紐づけて捉えられるようにしている．

図 1 に提案ツールのインタフェースを示す．このツールは，記事閲覧エリア（図 1 上部）と作業・判断エリア（図 1 下部）の 2 つから構成されている．記事閲覧エリアには，対象テキストが表示される．作業者はテキスト内の根拠箇所をクリックすることで，ハイライト（青/赤）を入れることができる．また，テキストの区切りが不適切な場合は，動的に文を分割することが可能である．

作業・判断エリアには，現在の作業工程における設問と操作パネルが表示される．複雑な仕様書を一度に提示するのではなく，「主題の特定」→「根拠のハイライト」

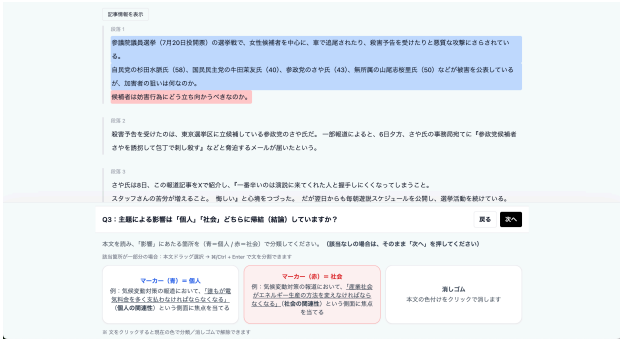

図 1: 提案ツールのインタフェース

→「全体判定」と段階的にタスクを提示することで，作業者の認知的負荷を低減することを企図している．

本ツールでは，これら各ステップでの操作（選択肢，ハイライト箇所，記述内容）および滞在時間をすべてプロセスログとして記録する．このログ機能により，最終ラベルと判断根拠がデータとして保存される．同時に，蓄積されたログは仕様書改善のサイクルを回すための客観的な分析資源として機能し，作業者の迷いや解釈の齟齬を定量的に特定することを可能にする．

4.2 実験準備

アノテーションを付与する対象の記事は，2025 年 7 月 14 日に Yahoo!ニュースに掲載されていた記事の中から，情報の効率的な伝達を目的とした従来の「逆三角形」構造である「ブラジルのペルアスー洞窟国立公園，ユネスコの世界自然遺産に登録される」[*1]（以下，ユネスコ記事）と，「物語」構造である「参政党・さや氏や山尾志桜里氏には殺害予告，国民民主党・牛田氏は車で追尾され… 女性候補への攻撃が相次ぐ理由」[*2]（以下，選挙妨害記事）の 2 記事を用いた．

本実験におけるアノテーションの判断基準は，第 3 章で述べた大森の指標を採用したが，これらは抽象的なアノテーション指針であり，ツール上の具体的な操作手順へ落とし込む必要があった．そこで，新聞記者経験のあるニュース研究者（本稿第 3 著者）の監修の下，ニュース記事 85 件を用いたアノテーションを実施して判断基準の具体化と合意形成を行い，本実験で使用する 2 記事に対する標準解を策定した．具体化された手順を表 1 に示す設問構成として本ツールに実装した．実験参加者は，画面上に順次表示される設問に回答することでアノテーションを進める．

[*1] https://news.yahoo.co.jp/articles/7bdc90774000336a5b4a86aa66fcd65c1ca02e60 (2025/7/14 確認).
[*2] https://news.yahoo.co.jp/articles/8fc4366e34ae12f3611a5eb9acde2116655d20c3 (2025/7/14 確認).

表 1: 大森の基準 [11] に基づき設計された設問

No.	対象側面	設問・タスク内容
Q1	前提	記者が記事で伝えたい「主題」を記述する
Q2	前提	記事が「速報・天気予報」であるか判定する
Q3	**F1**	主題による影響が「個人 (青)」「社会 (赤)」どちらに帰結するか抽出し，全体の傾向を判定する
Q4	前提	記事は「社会的な問題」を扱ったものか判定する
Q5	**F2**	社会的問題の語られ方を「エピソード (青)」「テーマ (赤)」で抽出する
Q6	**F2**	抽出箇所に基づき，記事全体のフレーム (エピソード型/テーマ型) を 3 段階で判定する
Q7	**S1**	リポーター等の「個人の見解」が含まれる文と，その根拠語句を抽出する
Q8	**S1**	抽出箇所に基づき，記事全体のスタイル (個人的/非個人的) を 3 段階で判定する
Q9	**S2**	感情を刺激する表現や，戦いに関連する語句を抽出する
Q10	**S2**	抽出語句に基づき，記事全体のスタイル (感情的/非感情的) を 3 段階で判定する

4.3 評価方法

本実験では，異なる記述構造を持つニュース記事間で，分類枠組みの適用可能性を比較・評価するため，以下のデータ処理および判定を行った．

1. **評価単位**：意味的に完結した一文を最小の評価単位と規定した．これに基づき，システム上で不自然に分割または結合されて提示された箇所については，本来の文構造に合わせて再構成（結合・分割）を行い，評価単位としての整合性を担保した．
2. **ラベル付与の判断基準**：再定義された文に対して，作業者が分割操作等により複数の箇所でラベルを付与していた場合，いずれか一箇所でも付与されていれば，その文に対してラベルが付与されたものとした．

評価指標の選定においては，ニュース記事の特性を考慮した．ニュース記事の 1 文には，「個人の帰結/エピソード」と「社会の帰結/テーマ」の双方が含まれる可能性があるため，これらをどちらが支配的かという排他的な基準で評価することは，情報の欠落を招く恐れがある．そこで本研究では，各カテゴリ（F1 における個人/社会，F2 におけるエピソード/テーマ）の独立性を担保するため，各々について選択されたか否かを個別に評価する 2 値分類のアプローチを採用した．これに基づき，分析の目的と粒度に応じて以下の指標を用いた．

単純一致率（文単位）
作業者間または標準解との間で，ラベルを付与するか否かの判断が一致した割合（κ 係数算出における観測一致率と同義）を算出する．本研究では，文単位での単純一致率をプロセスログと照合することで，判定のパターンや迷いやすい文を抽出する質的分析の手がかりとしても利用した．

κ 係数 （記事単位）
記事内の全単位に対する判断列（ベクトル）を入力とし信頼性係数 κ を算出する．なお，複数属性（青・赤）を統合した全体の評価においては，各色の判定ベクトルを結合した長さ $2N$ のベクトルを対象として κ を算出した．

4.4 実験手続き

アノテーションタスクにおいて生じる判断の齟齬が，単なる「作業手順や定義の曖昧さ」に起因するものか，あるいは「現代のニュース記事が持つ構造的複雑さと分類枠組みの本質的な不適合」に起因するものかを分離して検証するため，2 段階の実験を設計した．まず 1 段階目の実験（実験 1）では，初期状態の仕様書を用いてアノテーションを実施し，提案ツールのプロセスログから齟齬が生じた箇所を特定する．続いて 2 段階目の実験（実験 2）では，実験 1 で明らかになった定義の曖昧さや認知的なバイアスを排除するよう仕様書を修正し，再度アノテーションを実施する．なお，両実験ともに被験者は大学生（実験 1: 9 名, 実験 2: 7 名）とした．各実験において，参加者はツールの事前説明を受けた後，対象記事に対してアノテーションを行うよう指示された．

4.5 実験 1：結果

記事全体での評価を問う設問 (Q2, Q4, Q6, Q8, Q10) における一致率を表 2 に示す．Q2（速報性）や Q4（社会性）において，選挙妨害記事では概ね高い一致を示したが，ユネスコ記事では判断が割れた．Q8/Q10 においては，作業者間一致率は高い一方で，標準解との一致率がほぼゼロであるという乖離が見られた．

表 2: 実験 1：記事全体に対する設問の回答一致率

No.	記事	対 標準解 (正答率)	作業者間 (一致率)
Q2	選挙妨害記事	1.000	1.000
	ユネスコ記事	0.556	0.444
Q4	選挙妨害記事	0.889	0.778
	ユネスコ記事	0.556	0.444
Q6	選挙妨害記事	0.500	0.429
	ユネスコ記事	0.200	0.600
Q8	選挙妨害記事	0.556	0.361
	ユネスコ記事	0.000	0.778
Q10	選挙妨害記事	0.333	0.278
	ユネスコ記事	0.111	0.778

ログおよびアンケート分析に基づき，以下の設計上の課題を特定し，実験 2 へ向けた修正を行った．

- **主題特定の制約強化（Q1）**
  Q1 では「記者が記事で伝えたい主題を説明する」という自由記述形式を採用していたが，抽象度や記述の粒度に作業者間のばらつきが大きく，主題特定そのものが一致率の低下要因となっていた．そこで実験 2 では，ニュース記事に一般的に採用される逆三角形構造[*3]を踏まえ，「第一段落（リード文）から主題を抜き出す」ことを明示的な制約としてインタフェース上に実装した．この制約により，主題特定が作業者の読解力や要約方略に依存しにくい条件を整えることを意図した．さらに，第 1 段落に記事の主題が十分に含まれていない場合には，作業者の誤りとはみなさず，「該当なし」として Q1 の回答を省略できるよう設計した．これにより，本ツールはアノテーション仕様書や作業者の判断だけでなく，ニュース記事自体の構造的な良し悪しを評価する補助的な指標としても機能することが期待される．
- **フィルタリング設問の撤廃（Q2, Q4）**
  「速報性」や「社会性」の判定は，比較対象に依存する相対的なものであり，Q2, Q4 のような二値分類フィルタとして実装するには不適切であることが判明した．曖昧な基準によるフィルタリング設問で後続の分析データが欠損することを防ぐため，実験 2 ではこれらの事前フィルタを廃止し，全データを主要な分析フローへ回す方針とした．

第 3 章で述べた各側面（F1, F2, S1, S2）について，本文中の根拠箇所の抽出結果に基づき，記事全体としての傾向がどれだけ一致していたかを検証する．実験 1 における標準解との一致率，および，作業者間一致率を表 3, 表 4 に示す．全体として一致率は低調であり，特に選挙妨害記事の F1（社会）は 0.093，ユネスコの F2（全体）は 0.040 と，統計的に偶然レベルの一致に留まる項目が散見された．特定された要因に基づき，実験 2 に向けた仕様の修正を行った．

表 3: ユネスコ記事における一致率（実験 1・2 比較）

		実験 1		実験 2	
対象側面	評価ラベル	対 標準解	作業者間	対 標準解	作業者間
F1	個人	0.778	0.580	0.857	0.714
	社会	0.085	0.203	0.301	0.191
	全体	0.058	0.199	0.311	0.197
F2	エピソード	0.139	0.282	-0.247	0.253
	テーマ	0.139	0.171	-0.247	0.253
	全体	0.040	0.126	-0.584	0.290
S1	該当あり	0.111	0.469	0.429	0.352
S2	該当あり	0.066	0.582	-0.007	0.714

表 4: 選挙妨害記事における一致率（実験 1・2 比較）

		実験 1		実験 2	
対象側面	評価ラベル	対 標準解	作業者間	対 標準解	作業者間
F1	個人	0.111	0.016	0.143	0.460
	社会	0.093	0.055	0.011	0.487
	全体	0.050	0.027	-0.003	0.465
F2	エピソード	0.591	0.493	0.045	0.156
	テーマ	0.591	0.493	0.045	0.156
	全体	0.602	0.504	0.173	0.363
S1	該当あり	0.313	0.411	0.209	0.337
S2	該当あり	0.099	0.102	-0.003	0.292

- **F1：用語の認知的な齟齬**
  選挙妨害記事において，記事冒頭の問いかけを社会への帰結として誤ってマーキングする傾向がログから確認された．これは「帰結」という専門用語が，作業者に直感的に理解されていないことを示唆している．
- **F2：定義の境界における迷い**
  「テーマ」の定義に含まれる「専門家の解説」という記述に引きずられ，個人的なエピソードであっても専門家が登場するだけで「テーマ」と分類してしまう誤りが多発した．
- **S2: 対象とスタイルの混同**
  悲惨な事件（殺人予告など）の事実記述に対し，記者の表現自体は中立であるにも関わらず「感情的」と判定するケースが見られた．これは「出来事の性質」と「記事のスタイル」の切り分けが仕様書上で不明確であったことに起因する．

これらの分析に基づき，実験 2 では仕様書の用語変更や，注釈の追加を行った．

[*3] http://www.at-s.com/blogs/nie/study/howto.html（2025/12/12 確認）．

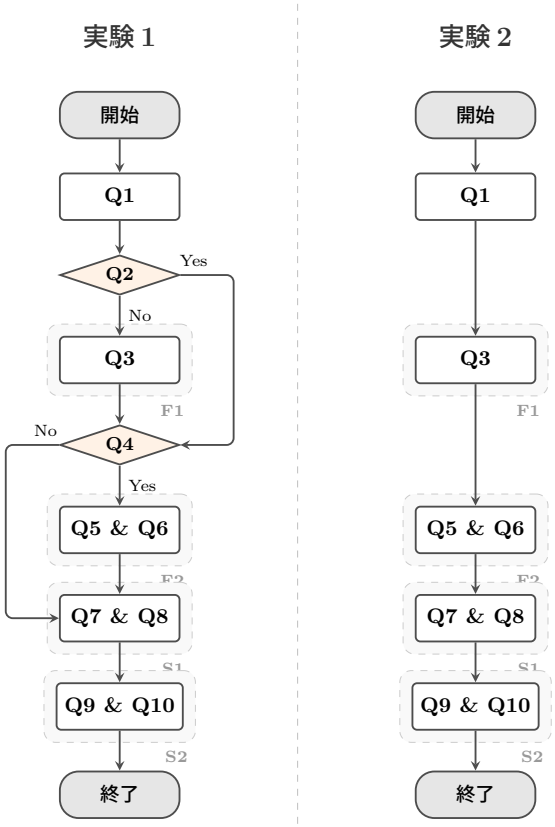

図 2: 実験フローの比較

4.6 実験 2：結果

実験 1 で特定された課題に対し，仕様書およびツール上の注釈を修正するとともに，判断フローを図 2（右）のように修正して実施した実験 2 の結果を比較する．

表 3 および表 4 に示した通り，一部の項目で一致率の向上が確認された一方，F2 や S2 では悪化も確認された．

- **作業者間の一致率の向上**：
  改善が確認された項目として，実験 1 で偶然レベル（$\kappa = 0.027$）と判定された選挙妨害記事の F1（全体）がある．実験 2 ではこれが $\kappa = 0.465$ まで改善が見られた．これは，実験 1 の分析で明らかになった「問い・事実」と「帰結」の混同に対し，「帰結」を「結論」と言い換えたことが，作業者の解釈を統一する上で有効性が示唆された．また，ユネスコの S2（感情）においても，作業者間一致率は 0.582 から 0.714 へと向上した．「出来事の悲惨さと，表現の感情的スタイルを区別する」という注釈を追加したことで，作業者間で判断基準が共有されやすくなったと考えられる．
- **標準解との一致率の向上**：
  ユネスコの F1 においては，「個人」ラベルの正答率が 0.778 から 0.857 へ，「社会」ラベルが 0.085 から 0.301 へと向上し，改善が見られた．一方で，作業者間の一致率は向上したものの，標準解との一致率が必ずしも連動して向上しないケース（例：選挙妨害記事の F1 やユネスコの F2 など）も確認された．これは，仕様書の改善によって「作業者集団の中での解釈」は収束したものの，その解釈が標準解作成者の意図とは異なる方向で収束した可能性を示唆している．

5 考察

実験結果が示すように，アノテーションの不一致は単なるランダムな誤りではなく，仕様書の定義，記事構造，作業者の解釈といった複数の要因が絡んだ構造的問題として発生している．本節では，得られた知見を「記事構造による差異」「標準解との乖離」の 2 点から考察する．

5.1 記事構造による差異

Q1 において「第一段落（リード文）から主題を抽出する」という制約を導入したことで，主題抽出の判断基準が記事構造に明示的に紐づけられ，作業者の読解力や要約能力の差異による影響を大幅に低減できた．実験 2 ではユネスコ記事において全ての作業者がほぼ同一の主題を抽出し，実験 1 で顕著であった主題記述の長さや抽象度に関する記述ゆれが解消された．逆三角形構造のニュース記事では主題が第一段落に集約されるため，この操作は本来，作業者間の一致を高める効果を持つ．しかし，選挙妨害記事においては作業者の回答が三つの小グループに分かれていた．自由記述ではなく第一段落からの抜き出しという操作に統一されているにもかかわらずこの分裂が生じたことは，作業者の読解力や要約能力の差異ではなく，リード文自体が複数の主題候補を含む構造になっていることを示唆する．

客観的な事実伝達を基本とする「逆三角形」構造のユネスコ記事では，スタイル面の作業者間一致率が比較的高かった．対照的に，「物語」構造を持つ選挙妨害記事においては，同項目の一致率が 0.102 に留まった．悲惨な出来事を扱う記事では，事実の描写そのものが強い感情的な文脈を帯びて読者に提示される．そのため，作業者が「客観的な事実の描写」と「記者の感情的な執筆スタイル」を明確に分離できず，結果としてアノテーションの判断が大きく分散したと考えられる．

5.2 作業者間の一致と標準解との乖離

実験 2 の結果において，仕様書の改善により作業者間の一致率は向上したものの，標準解との一致率は必ずしも向上しない，あるいは低下する事例が確認された．作業者同士の解釈は統一されたが，専門家の意図とは異なる方向へ収束したという現象は，アノテーションの品質管理において重要な示唆を与える．

この乖離が生じた主な要因として，標準解作成者が有する「暗黙知」が仕様書に十分に言語化されていなかった点が挙げられる．専門家が判断の際に用いた文脈依存的なニュアンスや判断基準が明文化されていなかったため，仕様書の記述を忠実に参照した作業者との間に認識の齟齬が生じたと考えられる．加えて，標準解そのものが，専門家の無意識のバイアスや感覚的な判断によって，本来定められた定義から逸脱していた可能性もある．

従来のアノテーション手法では，最終的なラベルのみを比較するため，この乖離が「作業者の理解不足」によるものなのか，「標準解の妥当性の欠如」によるものなのかを判別することは困難であった．しかし，本研究で提案したツールは，判断の根拠となるテキスト箇所を記録している．これにより，標準解作成者は，作業者が「なぜその選択をしたのか」という思考プロセスを追跡することが可能となる．

もし，作業者が仕様書の記述を忠実に守った結果として標準解と異なる判断を下しているのであれば，修正すべきは「作業者の判断」ではなく，「仕様書の記述」あるいは「標準解そのもの」であると判定できる．すなわち，本ツールは単に作業者を正解に導くだけでなく，標準解作成者自身が「自分の判断を見直すべきか，仕様書を修正すべきか」を客観的に判定するためのデバッグツールとしても機能する．これは，専門家の意図を仕様書やツール上でどう明確に伝えるかという設計プロセスにおいて，不可欠な機能であると言える．

6 おわりに

本研究では，ニュース報道の多次元的なハード・ソフト区分において分類基準の一貫性を保つことが困難であるという課題に対し，判断の段階的な分解と根拠箇所の記録・可視化を可能にするアノテーション支援ツールを開発し，分類の齟齬が生じる原因を調査した．

ツールから得られたプロセスログと検証実験の分析により，アノテーションにおける判断の不一致の主因が，作業者の単なる能力不足によるものではないことが確認された．専門家が持つ暗黙知と言語化されたマニュアル定義との間に生じるズレや，現代のネット記事が持つ特有の記述構造（物語性）との不適合が，評価の一貫性を阻害する構造的な要因であることを明らかにした．

特に，読者の共感を引き出すために客観的な社会課題の背後に個人の心情やエピソードを意図的に混在させる現代のネット記事（物語構造）では，客観的な事実描写と感情的な表現が融合していると考えられる．

謝辞

本研究は JST RISTEX (課題番号 JPMJRS23L2）の支援を受けた．また，本研究の実施にあたり森野穣氏から示唆を受けた．記して謝意を表す．

参考文献

[1] Aroyo, L. and Welty, C.: Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation, *AI Magazine*, Vol. 36, No. 1, pp. 15–24, DOI: 10.1609/aimag.v36i1.2564 (2015).

[2] Cabitza, F., Campagner, A. and Basile, V.: Toward a Perspectivist Turn in Ground Truthing for Predictive Computing, *Proc. AAAI Conf. on Artificial Intelligence*, Vol. 37, No. 6, pp. 6859–6867, DOI: 10.1609/aaai.v37i6.25840 (2023).

[3] Guest, G., MacQueen, K. M. and Namey, E. E.: *Applied Thematic Analysis*, Sage publications (2011).

[4] Krippendorff, K.: *Content Analysis: An Introduction to Its Methodology*, Sage Publications (2004).

[5] Lombard, M., Snyder-Duch, J. and Bracken, C. C.: Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability, *Human Communication Research*, Vol. 28, No. 4, pp. 587–604, DOI: 10.1111/j.1468-2958.2002.tb00826.x (2006).

[6] Reinemann, C., Stanyer, J., Scherr, S. and Legnante, G.: Hard and soft news: A review of concepts, operationalizations and key findings, *Journalism*, Vol. 13, No. 2, pp. 221–239, DOI: 10.1177/1464884911427803 (2012).

[7] Tan, Z., Li, D., Wang, S., Beigi, A., Jiang, B., Bhattacharjee, A., Karami, M., Li, J., Cheng, L. and Liu, H.: Large Language Models for Data Annotation and Synthesis: A Survey, *Proc. 2024 Conf. on Empirical Methods in Natural Language Process-*

ing, pp. 930–957, DOI: 10.18653/v1/2024.emnlp-main.54 (2024).

[8] Tuchman, G.: Making news by doing work: Routinizing the unexpected, *American Journal of Sociology*, Vol. 79, No. 1, pp. 110–131 (1973).

[9] 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. 90–138, DOI: 10.1162/coli_a_00502 (2024).

[10] 石戸諭: ニュースの未来, 光文社 (2021).

[11] 大森翔子: メディア変革期の政治コミュニケーション: ネット時代は何を変えるのか, 勁草書房 (2023).

[12] 斉藤友彦: 新聞記者がネット記事をバズらせるために考えたこと, 集英社 (2025).

[13] 杉本麻衣, 松下光範, 藤代裕之: 曖昧さを含む仕様書の改善を目的としたアノテーション支援ツールの検討, 情報処理学会研究報告, Vol. 2025-HCI-216, No. 18, pp. 1–7 (2026).

ライトノベルにおける位置情報・意味情報・表紙要素に基づく複合ネタバレ度指標に対する有効性評価

Effectiveness of a Composite Spoiler Metric Based on Positional, Semantic, and Cover Elements in Light Novels

井上大輝[1]　尾崎知伸[1]*
Daiki Inoue[1]　Tomonobu Ozaki[1]

[1] 日本大学
[1] Nihon University

Abstract: This study evaluates the effectiveness of a composite metric focusing on three aspects of positional information, semantic information, and cover elements for the purpose of spoiler detection in light novels. Specifically, we used a manually annotated corpus with spoiler intensity as ground truth data to quantitatively analyze the relationship between each aspect and the spoiler degree. Furthermore, through application to summary generation by generative AI, we verified the utility of the composite metric in practical applications.

1　はじめに

ライトノベルは，個性的なキャラクターやテンポの良いストーリー展開を特徴とする物語小説の一形態であり，読者に深い没入感を伴う読書体験を提供している．その魅力の核心は，キャラクター間の関係性の変容や伏線の回収といったドラマチックな展開にあり，些細な描写が物語全体の方向性に重大な意味を持つことも少なくない．このような特性から，ライトノベルにおけるストーリー情報の取り扱いには細心の注意を要する．未読の読者がネタバレを含む情報に接触した場合，著者が本来意図していた驚きや感情的なインパクトが損なわれる懸念があるためである．その一方で，個々の読者の読書進捗に合わせ，物語の魅力を損なわずに要約や紹介文を生成する技術への需要は高い．ネタバレを適切に制御しつつ物語の核心を伝える要約生成システムが実現すれば，より効果的な情報共有や作品のプロモーションが期待できる．

これらのことを背景に，これまでネタバレ検出に関する研究が精力的に行われている．例えば Wan ら [1] は，書評データセットを用いた文単位のネタバレ検出タスクを対象に，ネタバレの傾向・性質を分析した上で，ユーザと作品のバイアスを考慮した階層的注意ネットワークに基づくモデルを提案した．また Chang ら [2] は，グラフとして表現される文の係り受け関係に GNN を適用することで，文脈に応じた意味識別を可能にする手法を開発した．一方，前田ら [3] は，レビュー内の記述が本文のどの部分に対応するかを推定することで，ネタバレの可能性を評価する枠組みを示した．さらに Tran ら [4] は，ネタバレ検出をレビューと本文との意味的対応問題として再定義している．

本研究では，ライトノベルのネタバレを対象とした，位置情報・意味情報・表紙要素の 3 側面に着目した複合指標 [5] に着目する．また，この指標に基づく実応用に向けての初手として，指標が持つ性質・特徴を明らかにすることを試みる．具体的には，人手によりアノテーションを行ったネタバレ度付きコーパスを正解データとし，各側面とネタバレ度との関連を定量的に分析する．さらに，生成 AI による要約生成への適用を通じ，実応用における有用性を検証する．

本論文の構成は以下のとおりである．2 章で対象とする複合指標を説明する．3 章で指標に対する評価実験を行う．4 章でまとめを行い，今後の課題を述べる．

2　複合ネタバレ度指標

本研究で対象とする複合ネタバレ度指標 [5] は，小説本文を固定長の文書（文書区間）に分け，位置情報・意味情報・表紙要素の 3 側面から各スコアを算出・総合することで最終的なネタバレ度を決定する．すなわち，小説 n における第 i 番目の文書区間 $c_i^{(n)}$ に対して位置スコア S_{pos}，意味スコア S_{sem}，表紙スコア S_{cov} を算

*連絡先：日本大学 文理学部 情報科学科
〒 156-8550 東京都 世田谷区 桜上水 3-25-40
E-mail: ozaki.tomonobu@nihon-u.ac.jp

出し，重み w_{pos}，w_{sem}，w_{cov} による加重和である

$$S(c_i^{(n)}) = \frac{w_{\mathrm{pos}} S_{\mathrm{pos}} + w_{\mathrm{sem}} S_{\mathrm{sem}} + w_{\mathrm{cov}} S_{\mathrm{cov}}}{w_{\mathrm{pos}} + w_{\mathrm{sem}} + w_{\mathrm{cov}}}$$

を，$c_i^{(n)}$ に対する複合ネタバレ度とする．

2.1 位置スコア

位置スコアは，物語が進行するにつれてネタバレがより顕著になるという仮定に基づく．小説 n における文書区間総数を $B^{(n)}$ としたとき，$B^{(n)} > 1$ を前提に，$c_i^{(n)}$ に対する位置スコア S_{pos} は次のように定義される．

$$S_{\mathrm{pos}}(c_i^{(n)}) = i \,/\, (\, B^{(n)} - 1\,)$$

2.2 意味スコア

意味スコア S_{sem} は，w_{smy} と w_{tpl} を重みとする，要約スコア S_{smy} とトリプルスコア S_{tpl} の加重和である．

$$S_{\mathrm{sem}}(c_i^{(n)}) = \frac{w_{\mathrm{smy}} S_{\mathrm{smy}}(c_i^{(n)}) + w_{\mathrm{tpl}} S_{\mathrm{tpl}}(c_i^{(n)})}{w_{\mathrm{smy}} + w_{\mathrm{tpl}}}$$

要約スコア S_{smy} は，GPT-4o により推定される，前後の文脈を考慮したネタバレ強度を表すスコアである．以下のように定義され，その値域は $[0.0, 1.0]$ である．

$$S_{\mathrm{smy}}(c_i^{(n)}) = \\ LLM\left(D\left(\{c_j^{(n)}\}_{j=0}^{i-1}\right), c_i^{(n)}, D\left(\{c_k^{(n)}\}_{k=i+1}^{B^{(n)}-1}\right)\right)$$

ここで LLM は，GPT-4o によるスコア推定を，また D は，GPT-4o mini によって得られる連続文書区間の要約テキストをそれぞれ表す．

一方，トリプルスコア S_{tpl} は，$c_i^{(n)}$ に含まれるトリプルの全体集合 $\mathcal{T}_i^n$ から導出されるスコアであり，以下のように定義される．

$$S_{\mathrm{tpl}}(c_i^{(n)}) = \frac{1}{|\,\mathrm{Top}_K(\mathcal{T}_i^n)\,|} \sum_{t' \in \mathrm{Top}_K(\mathcal{T}_i^n)} S_{\mathrm{tmp}}(t')$$

ここで Top_K は，$\mathcal{T}_i^n$ に含まれる，時間スコア S_{tmp} 上位 K 件からなる集合である．

2.3 時間スコア

トリプル t に対する時間スコア S_{tmp} は，w_{pst} と w_{fut} を重みとする，過去スコア S_{pst} と未来スコア S_{fut} の加重和である．

$$S_{\mathrm{tmp}}(t) = \frac{w_{\mathrm{pst}} S_{\mathrm{pst}}(t) + w_{\mathrm{fut}} S_{\mathrm{fut}}(t)}{w_{\mathrm{pst}} + w_{\mathrm{fut}}}$$

過去スコア　トリプル $t \in \mathcal{T}_i^n$ に対する過去スコア S_{pst} は，t 単体のネタバレ危険性を表す内容スコア $S_{\mathrm{cnt}}(t)$ と，先行するトリプル集合 $\mathcal{PT}_i^n = \bigcup_{j=0}^{i-1} \mathcal{T}_j^n$ に対する新規性スコア S_{nvl} との積，すなわち

$$S_{\mathrm{pst}}(t) = S_{\mathrm{cnt}}(t) \times S_{\mathrm{nvl}}(t)$$

と定義される．ここで内容スコア $S_{\mathrm{cnt}}(t)$ は，8 種のネタバレを対象とした事前学習済み BERT モデルによる予測値 $p_y^t\,(y \in \{1, \cdots, 8\})$ と閾値 θ_c を用い，

$$S_{\mathrm{cnt}}(t) = 1 - \prod_{y \in \{1,\ldots,8\} \wedge p_y^t \geq \theta_c} (1 - p_y^t)$$

と定義される．一方，新規性スコア S_{nvl} は，ベクトル間類似度を基礎とし，パラメタ θ_n と α_n を用いて

$$S_{\mathrm{nvl}}(t) = \sigma\left(\left(\theta_n - \max_{t' \in \mathcal{PT}_i^n} \cos(\mathrm{enc}(t), \mathrm{enc}(t'))\right) \times \alpha_n\right)$$

と定義される．なお，σ はシグモイド関数，enc はエンコーダによるトリプルのベクトル表現，cos はベクトル間コサイン類似度である．

未来スコア　トリプル $t \in \mathcal{T}_i^n$ に対する未来スコア S_{fut} は，後続するトリプル集合 $\mathcal{FT}_i^n = \bigcup_{j=i+1}^{B^{(n)}-1} \mathcal{T}_j^n$ との平均類似度 avg と，類似トリプル集合 sim_triples を用い

$$S_{\mathrm{fut}}(t) = \mathrm{avg}(t) \times \frac{\log(\,1 + |\,\mathrm{sim\_triples}(t)\,|\,)}{\log(\,1 + |\,\mathcal{FT}_i^n\,|\,)} \times \alpha_f$$

と定義される．なお α_f はパラメタである．また，パラメタ θ_f を用い，平均類似度と類似トリプル集合は，それぞれ以下のように定義される．

$$\mathrm{avg}(t) = \frac{1}{|\,\mathcal{FT}_i^n\,|} \sum_{t' \in \mathcal{FT}_i^n} \cos(\mathrm{enc}(t), \mathrm{enc}(t'))$$

$$\mathrm{sim\_triples}(t) = \{t' \in \mathcal{FT}_i^n \mid \cos(\mathrm{enc}(t), \mathrm{enc}(t')) \geq \theta_f\}$$

2.4 表紙スコア

表紙スコア S_{cov} は，表紙単語群（表紙に関する単語集合）$\mathcal{C}^n$ を基準とする，制御パラメータ α_s および閾値 θ_s を伴う飽和加重和であり，

$$S_{\mathrm{cov}}(c_i^{(n)}) = 1 - \prod_{cw \in \mathcal{C}^n \wedge \cos(\mathrm{enc}(c_i^{(n)}), \mathrm{enc}(cw)) \geq \theta_s} \left(1 - \cos(\mathrm{enc}(c_i^{(n)}), \mathrm{enc}(cw))^{\alpha_s}\right)$$

と定義される．なお $\mathcal{C}^n$ の抽出は，表紙イラストとタイトルを対象に，GPT-4o/4o-mini によって行われる．

3 評価実験

複合ネタバレ度指標の有効性と実用性を検証するため，以下の 3 つの実験を行う[1]．なお正解データは，実

[1] パラメタ設定：$w_{\mathrm{smy}} = w_{\mathrm{tpl}} = 1, K = 5, w_{\mathrm{pst}} = w_{\mathrm{fut}} = 1, \theta_c = 0.5, \theta_n = 0.8, \alpha_n = 1.2, \alpha_f = 1.0, \theta_f = 0.8, \alpha_s = 1.5, \theta_s = 0.5$

表 1: スコア間のピアソン相関行列

	正解	S_{pos}	S_{sem}	S_{cov}
S_{pos}	0.268	—	0.219	−0.157
S_{sem}	0.300	0.219	—	0.110
S_{cov}	0.147	−0.157	0.110	—

験協力者 6 名がそれぞれライトノベル 1 冊（平均文字数約 13 万 ±3,000 字）を精読し，各文書区間に対して 6 段階（0 ∼ 5）でネタバレ度を付与することで構築した．

実験 1： 位置，意味，表紙の各スコアが，人間が感じるネタバレ度と整合しているかを検証する．

実験 2： 線形計画法を用い，各スコアの最適な統合方法（重み）を決定する．

実験 3： あらすじ生成タスクに複合ネタバレ度指標を適用し，生成物の品質や安全性を評価する．

3.1 スコアに対する相関分析

各ネタバレ度スコアと正解データとの相関係数を表 1 に示す．得られた結果に基づき，以下の 3 点について考察する．

第一に，すべてのスコアが正解データに対して正の相関を示した．特に意味スコア S_{sem} は 0.300 と最も高い相関を示しており，テキストの意味情報がネタバレ判定に強く関係していることが確認できる．

第二に，表紙要素の役割についてである．表紙スコア S_{cov} と正解データとの相関係数は 0.147 と弱い．この結果は，表紙イラストは読者にとって既知の情報でありネタバレには該当しない，という既知情報としての性質と，表紙イラストはクライマックスの場面や重要人物を描写する傾向が強く，重要なネタバレを含んでいる，という重要情報としての性質とが混在していることを示唆しており，互いに相殺しあった結果として相関が現れなかったものと推察される．

第三に，スコア間の独立性である．特に S_{pos} と S_{cov} の相関係数は −0.157 と負の値を示した．これは，表紙情報が物語の冒頭などの導入部に出現しやすいという直観的な傾向と一致する．このように，互いに異なる時間的内容的傾向を持つ各指標を適切に組み合わせることで，単一の指標では捉えきれないネタバレ箇所を相補的に検出できる可能性が高い．

3.2 統合重みの最適化

各スコアのネタバレに対する貢献を検証するため，線形計画法を用い，最適な統合重み（$w_{\text{pos}}, w_{\text{sem}}, w_{\text{cov}}$）を導出する．なお表紙スコア S_{cov} に関しては，既知情報・重要情報の両性質が考えられるため，新たにスコア

$$\tilde{S}_{\text{cov}}(c_i^{(n)}) = \begin{cases} 1 - S_{\text{cov}}(c_i^{(n)}) & \text{（表紙を既知情報と見做す）} \\ S_{\text{cov}}(c_i^{(n)}) & \text{（表紙を重要情報と見做す）} \end{cases}$$

を導入し，複合ネタバレ度を

$$S(c_i^{(n)}) = \frac{w_{\text{pos}} S_{\text{pos}} + w_{\text{sem}} S_{\text{sem}} + w_{\text{cov}} \tilde{S}_{\text{cov}}}{w_{\text{pos}} + w_{\text{sem}} + w_{\text{cov}}}$$

と再定義した上で最適化を行う．

具体的には，文書の全体集合を N，文書区間 $c_i^{(n)}$ に対する正解データを $G(c_i^{(n)})$ とし，制約

$$w_{\text{pos}} \geq 0, w_{\text{sem}} \geq 0, w_{\text{cov}} \geq 0, w_{\text{pos}} + w_{\text{sem}} + w_{\text{cov}} = 1$$

のもとで，以下の 3 つの目的関数をそれぞれ最小化する．またその際，表紙を既知・重要のどちらの情報と見做すのかを定め，両者を比較する．

目的関数 O_1： 文書区間の絶対誤差を直接最小化する．

$$\sum_{n \in N,\, 0 \leq i \leq B^{(n)}-1} \left| S(c_i^{(n)}) - G(c_i^{(n)}) \right|$$

目的関数 O_2： 全文書区間ペア $(c_i^{(n)}, c_m^{(m)})$ について，S の差と G の差の絶対誤差を最小化する．

$$\sum_{n \in N,\, 0 \leq i \leq B^{(n)}-1} \sum_{m \in N,\, 0 \leq j \leq B^{(m)}-1} \left| \left(S(c_i^{(n)}) - S(c_j^{(m)}) \right) - \left(G(c_i^{(n)}) - G(c_j^{(m)}) \right) \right|$$

目的関数 O_3： 同一文書内における文書区間ペアについて，S の差と G の差の絶対誤差を最小化する．

$$\sum_{n \in N} \sum_{0 \leq i < j \leq B^{(n)}-1} \left| \left(S(c_i^{(n)}) - S(c_j^{(n)}) \right) - \left(G(c_i^{(n)}) - G(c_j^{(n)}) \right) \right|$$

表 2: 最適重みパラメタ

	既知情報			重要情報		
	w_{pos}	w_{sem}	w_{cov}	w_{pos}	w_{sem}	w_{cov}
O_1	0.000	0.965	0.035	0.000	1.000	0.000
O_2	0.114	0.886	0.000	0.141	0.766	0.093
O_3	0.112	0.888	0.000	0.138	0.760	0.102

得られた重みパラメタを表 2 にまとめる．結果から，以下の 2 つの知見が伺える．

第一に，単純な絶対誤差に基づくの最適化における限界である．目的関数 O_1 では，表紙スコア $\tilde{S}_{\text{cov}}$ の設定（既知／重要）に関わらず，位置スコアの重み w_{pos} が 0.000 となり，位置情報が評価に寄与しない結果となっ

た．特に設定（重要）では，意味スコアの重み w_{sem} が 1.000 に収束し，モデルが単一の指標と等価的になってしまっている．これは，スコアの絶対誤差を直接最小化する手法では，正解データと最も高い相関を持つ意味スコアのみに最適化が集中してしまい，物語の進行に伴う相対的な変化を示す位置情報の特性を十分に反映できないことを示している．

第二に，表紙スコアの設定による有効性の差異である．O_2 および O_3 において表紙要素を「既知情報」として扱った場合，w_{pos} は約 0.11 まで回復したものの，w_{cov} は 0.000 となり，表紙スコアは再度無効化された．これに対し，表紙要素を「重要情報」として扱った場合，w_{cov} に約 0.09 から 0.10 の有意な重みが割り当てられた．この結果は，表紙をネタバレを増進させる要素として定義することが，多角的なネタバレ度評価モデルを構築する上で不可欠であることを裏付けている．

3.3 要約生成

複合ネタバレ度指標の実用性を検証するため，大規模言語モデル（GPT-5）を用いたネタバレ配慮型あらすじ生成を題材とした評価実験を行った．具体的には，前節の実験結果に従い，設定 $\tilde{S}_{\text{cov}}(c_i^{(n)}) = S_{\text{cov}}(c_i^{(n)})$，$w_{\text{pos}} = 0.138$，$w_{\text{sem}} = 0.760$，$w_{\text{cov}} = 0.102$ を用い，以下の各方法で GPT-5 にあらすじを生成させ，それらを比較する．

方法 M_1： 小説全文を入力し，プロンプトでの言語的な指示に依存して生成する方法．ネタバレ指標は利用しない．

方法 M_2： 各文書区間の先頭に，メタデータとしてネタバレ度スコアを付与した上で，小説全文を入力する方法．プロンプトにおいて，高スコアの箇所は物語の核心であるため，あらすじには使用しないよう指示を与える．

方法 M_3： 予めネタバレ度スコアが低い文書区間を抽出・排除し，総文字数が約 3 万文字になるように調整したテキストのみを入力する方法．

あらすじに対する評価は，3 名の被験者によるブラインドテスト形式で実施した．各被験者に対し，各方法でそれぞれ 2 つずつ生成した 6 件のあらすじ $M_i^j, (1 \leq i \leq 3, j \in \{a, b\})$ と，出版社による公式あらすじ P の計 7 つを提示し，ネタバレのなさ（安全性）と紹介文としての魅力（品質）の観点から総合的に優れていると感じた順に順位付けを求めた．

結果を表 3 に示す．得られた順位傾向より，以下の 2 つの知見が確認できる．

表 3: あらすじに対するランキング

評価者	1 位	2 位	3 位	4 位	5 位	6 位	7 位
A	P	M_2^b	M_3^b	M_2^a	M_1^a	M_3^a	M_1^b
B	P	M_2^b	M_2^a	M_1^b	M_1^a	M_3^b	M_3^a
C	M_1^a	M_2^a	P	M_2^b	M_1^b	M_3^b	M_3^a

第一に，機械的なフィルタリングに基づく方法 M_3 の評価が極めて低い点である．特に評価者 B および C において，M_3^a と M_3^b が最下位となっている．これは，ネタバレ度スコアが高い文書区間，すなわち物語の核心や山場を単純に削除した結果，あらすじとして読者の興味を惹くフックや重要な導入部分までもが欠落し，文脈が不自然となったためである．この結果は，ネタバレ度スコアが物語の魅力的な要素を適切に捉えていること，つまりネタバレ個所の検出性能が高いことを逆説的に示している．

第二に，方法 M_2 の有効性である．M_2 によるあらすじは，すべての評価者において安定して上位を獲得した．方法 M_1 と比較しても，手法 M_2 の方が「物語の核心を保護しつつも具体的」という，安全面と品質面の高度なトレードオフを成立させている。これは，GPT-5 がネタバレ度スコアを情報の重要度や秘匿性に関するメタデータとして解釈し，全体の文脈を維持しながらも高スコア区間の情報漏洩のみを適応的に抑制した結果であると考えられる．

4 まとめ

本研究では，ライトノベル 6 冊を対象に，複合ネタバレ度指標に対する評価を行った．その結果，以下の成果が得られた．第一に，ネタバレ検出における表紙要素の重要性を明らかにした点である．従来，表紙イラストは読者が最初に接する情報であり，既知の要素としてネタバレには該当しないと解釈される傾向にあった．しかし本研究により，ライトノベルの表紙には物語のクライマックスや重要人物が描写されていることが多く，表紙との類似性が高いシーンほどネタバレ度も高いという相関が確認された．

第二に，複合ネタバレ度指標の最適重みを特定した点である．実験の結果，同一作品内における文書区間同士の相対的な差異を基準とする方法が最も有効であることが判明した．これは，作品ごとに評価基準やネタバレの性質が異なるため，同一作品内での相対的な盛り上がりを捉えるアプローチが，より正確な特定に寄与するためである．

第三に，大規模言語モデルを用いた応用可能性を示した点である．算出されたネタバレ度をメタデータとしてモデルに提示することで，物語の核心を保護しつ

つ，読者の関心を惹く魅力的なあらすじが生成可能であることを確認した．単なる機械的な削除では文脈の不自然さを招くが，スコアを情報の重要度や秘匿性の指標として活用することで，生成物の品質維持とネタバレ防止の両立に成功した．

今後の課題として次の 3 点が挙げられる．第一に，評価規模の拡大と客観性の向上である．本研究における評価は限られた被験者数による試行的な検証にとどまっている．今後はより大規模な被験者を対象とした調査を行い，ネタバレ度指標の汎用性と実用性を統計的に検証する必要がある．

第二に，他ジャンルへの適応である．例えば，ミステリーやサスペンスといったジャンルでは，表紙イラストが読者を誘導するミスリードとして機能する場合も想定される．ジャンル固有の表現上の特性を考慮し，ネタバレ度算出のアルゴリズムを適応的に調整していくことが求められる．

最後に，スコア統合モデルの高度化である．各指標の線形和ではなく，指標間の相互作用や非線形な関係性を考慮した，より複雑なモデルを導入する余地がある．高度な機械学習手法を取り入れることで，人間の複雑なネタバレ感覚により近い判定モデルへの発展が期待される．

参考文献

[1] M. Wan, R. Misra, N. Nakashole, and J. McAuley : Fine-grained spoiler detection from large-scale review corpora, In *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, pp.2605–2610, 2019.

[2] B. Chang, I. Lee, H. Kim, and J. Kang : “Killing Me” is not a spoiler: Spoiler detection model using graph neural networks with dependency relation-aware attention mechanism, In *Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pp.3613–3617, 2021.

[3] 前田 恭佑，土方 嘉徳，中村 聡史，酒田 信親 : ストーリー文書を用いたレビュー文書のネタバレ判定，システム制御情報学会論文誌，32(3):87–100, 2019.

[4] R. Tran, C. Xu, and J. McAuley : Spoiler detection as semantic text matching, In *Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp.6109–6113, 2023.

[5] D. Inoue and T. Ozaki : A Composite Scoring Method for Spoiler Detection in Light Novels, *Proc.* of the 2026 11th International Conference on Intelligent Information Technology, 2026.

COM-PASS: 物語の共通点・相違点の収集システム

COM-PASS: A System for Collecting Similarities and Differences in Stories

塩田 隼士 †,*　藤川 雄翔 ‡　松下 光範 †
Hayato Shiota　Taketo Fujikawa　Mitsunori Matsushita

† 関西大学総合情報学部
Faculty of Informatics, Kansai University

‡ 関西大学大学院総合情報学研究科
Graduate School of Informatics, Kansai University

Abstract: 本稿では，物語作品同士のストーリーに関する共通点・相違点を収集するシステム「COM-PASS」について提案する．本システムでは，物語を特徴付ける「メインキャラクタ」「目標・結末」「世界観」を記述の枠組みとして設定し，ユーザの比較による作品間の情報を構造的に収集することを可能にする．収集されたデータを分析した結果，ユーザは類似するジャンルやキャラクタの特徴を共通点として記述し，それらを詳細化するかたちで相違点を記述する傾向が確認された．

1　はじめに

現在，閲覧可能なコミックやアニメ，小説などの物語作品は日々増加を続けており，その数は膨大となっている．それらの物語作品の中から自身の嗜好に合致する作品を発見するために，これらのコンテンツに関わる既存の配信サービスでは複数作品に当てはまる共通点に着目した情報アクセス手法が利用されている．例えば，honto[*1]といった電子書籍販売サイトや Wikipedia[*2]などの Web 百科事典では，書誌情報（e.g., タイトル，著者名）やジャンル（e.g., 少年コミック，ファンタジー）などのメタ情報をクエリとした作品検索が可能である．しかし，こうしたメタ情報による作品検索の場合，該当する作品が数多く存在するため，ユーザが自身の嗜好に合致する作品を効率的に発見することは難しい．また，作品そのものではなく，物語に登場するキャラクタを検索することができるサービスも存在する．例えば，キャラ属性王国[*3]では既知のキャラクタと共通の属性を持つ未知のキャラクタを発見できる．しかし，本サービスはキャラクタの類似性を提示するに留まっており，物語作品のストーリーを重視した情報アクセスは難しい．

こうした現状のサービスの課題を解決するため，藤川らはストーリーを重視した効率的な情報アクセスの実現を目指し，物語のコンセプトに着目した研究を行っている [6]．物語作品のコンセプトとは，ストーリーの土台となるアイディアのことである [2][4]．ストーリーはそのコンセプトに沿ってキャラクタが行動することにより進行する．藤川らは，「敵対する組織に属する男女二人が恋に落ちる話」のように，複数の物語作品に共通するストーリーを端的に表現した文章が物語のコンセプトを表すものであるとし，これをコンセプト文と定義してその収集や自動生成を試みている．しかし，コンセプト文は複数の物語作品に当てはまることから，現状では同一のコンセプト文を有する作品群を提示するに留まっている．同一のコンセプト文に属する作品群であっても，個々の作品はストーリーの細部や展開に違いが存在する．この違いを明示できれば，ユーザは自らの嗜好に合致する物語作品をより効率的に発見できるようになると考えられる．このような情報アクセスを実現するには，各物語作品のストーリーにおける共通点を説明するデータに加え，共通点を有する作品同士の相違点を説明するデータが必要となる．本稿ではこのようなデータをユーザから収集するためのシステム「COM-PASS」を提案し，物語のストーリーに関する共通点および相違点のデータセットの構築を試みる．

* 連絡先：関西大学総合情報学部
〒569-1095 大阪府高槻市霊仙寺町 2-1-1
E-mail: k237166@kansai-u.ac.jp

*1 https://honto.jp/ （2025/2/15 確認）.
*2 https://ja.wikipedia.org/wiki/（2025/2/15 確認）.
*3 https://chara-zokusei.jp/（ 2025/2/15 確認）.

2 関連研究

物語内容の構造化やそれによる内容把握の支援を試みた研究が存在する．Propp はロシアの魔法民話を対象に物語の分割基準として 31 の機能を定義し，それらの連鎖によって物語が構成されていることを明らかにした [1]．Snyder は，映画のストーリーをその中核要素に基づいて 10 のタイプに分類した [3]．野村らは，コミックを書誌情報，物理的構造，知的内容の 3 視点から捉えるメタデータモデルを提案している [5]．宮川らは，登場人物間の関係性について，鑑賞者が読み進める時間である「物語言説」と作中の時間が流れる「物語内容」の 2 つの時間軸を対応付け，その可視化を試みた [7]．

藤川らは，コンセプトに基づいた物語作品への情報アクセスの実現を目指し，コンセプト文に関する研究を行っている [6]．物語作品の鑑賞者を対象としたクラウドソーシングを通じて物語作品を要約した短文の収集を行い，それらが類似するか，調査を行った．その結果，鑑賞者が記述した物語作品を要約した短文は物語のコンセプトの説明として利用可能であることが示唆された．また，物語作品間の共通点を説明するコンセプト文には，キャラクタに関する情報およびキャラクタのストーリー上で主軸となる行動の情報が含まれることを明らかにしている．藤川らは人手によりコンセプト文を収集するコストを削減するため，既存のあらすじ文からコンセプト文を自動生成する手法を検討している．類似する複数の物語作品のあらすじ文に共通して出現する単語に着目し，大規模言語モデルを用いることでコンセプト文を生成する手法を提案した．

コンセプト文に関する研究では，作品間の共通点の提示に主眼が置かれており，同一のコンセプトを共有する作品同士の差異を把握することは困難である．ストーリーの差異を抽出するためには，比較の基準となる共通点の存在が不可欠であるが，この両者を同時に扱う枠組みは十分に検討されていない．そこで本稿では，共通点と相違点を併せて収集することにより，コンセプト文では捉えきれなかったストーリーの差異を明らかにすることを目指す．ここで扱う共通点および相違点は，複数作品の比較を通じて初めて顕在化する相対的な情報であるため，既存のあらすじ文や要約文を蓄積したデータセットをそのまま利用することは難しい．このことから，物語作品の鑑賞者による比較記述から共通点および相違点を収集するというアプローチを採る．情報の収集に際しては，内容把握を容易にするための構造化を図る一方で，鑑賞者が行う比較の思考を阻害しない記述形式が求められる．そこで，鑑賞者の着眼点を反映した記述の枠組みを設定することで，その思考過程を妨げることなく，ストーリーに関する構造化されたデータ収集を行う．

3 デザイン指針

3.1 収集データの定義

本稿では，共通点を「複数の物語作品間で共有される内容を説明する文」，相違点を，「共通点を持つ作品間における内容の差異を説明する文」と定義する．ここで，共通点および相違点には，その作品でしか使われない固有表現（e.g., キャラクタの名前，場所の名前）は含めないものとする．これは，作品の固有表現がユーザにとって，作品の内容理解のノイズになる懸念があるためである．また，収集されるデータは「敵キャラクタの服装」といった情報ではなく，ストーリーとの関連性の高い情報に絞ることとする．

3.2 記述の枠組み

ストーリーとの関連性が高く，かつ構造化された共通点・相違点のデータをユーザから効率的に収集するために，本稿では「メインキャラクタ」「目標・結末」「世界観」の 3 項目を記述の枠組みとして設定した．以下に，設定項目の定義および採用理由について述べる．

「メインキャラクタ」は，主人公やライバル，ヒロインなどのストーリーの中心的存在となる登場人物の性格や他者との関係などの情報を収集する項目であり，「目標・結末」は，メインキャラクタが物語上で目指す状態やそれに至るための行動に関する情報を収集する項目である．これらの項目は，物語のストーリーとの関連性が高い情報を収集するために収集する．先述のように，藤川らはコンセプト文にはキャラクタに関する情報およびキャラクタのストーリー上で主軸となる行動の情報が含まれることを示唆している [6]．本システムではこの知見に基づき，ストーリーを進行させる中心人物を記述する「メインキャラクタ」と，そのキャラクタの到達点や行動を記述する「目標・結末」の 2 項目を収集対象として採用する．

また，「世界観」は，舞台設定や登場人物を取り巻く環境についての情報を収集する項目である．「世界観」はユーザが共通点を想起する際の手がかりとして機能するという仮説から設定した．ユーザは「異世界」といった一般ではジャンルとして利用可能な情報や，「勝負に勝ち続けなければならない世界」といった作品全体を俯瞰した情報を共通点として認識していると考えた．

4 実装

本稿で提案する COM-PASS は，多様な利用者の協力を得ることを企図して，ブラウザで操作・閲覧が可能な Web システムとして開発した．本システムは大きく分けてデータ投稿画面（図 1）とデータ可視化画面（図 2）の 2 つの画面から構成される．

まずデータ投稿画面において，ユーザは本画面で任意の物語作品を 2 つ選択する．比較する 2 作品を決定した後，各項目を共通点・相違点のいずれとして入力するかを選択して記述する．「この内容で登録する」ボタンを押下することで内容がデータベースに保存される．

つづいて，データの可視化画面について述べる．この画面では，投稿されたデータが共通点を通じてネットワーク構造で表示される．任意の作品を選択することで，その作品と直接共通点を持つ作品とそこから共通点を介して繋がりのある全ての作品が強調して表示される．これと同時にサイドバーが表示され，選択された作品と直接比較された作品の共通点および相違点の記述を閲覧することができる（図 2）．可視化画面には，「新たに 2 作品を登録」などのボタンが存在する．このボタンを押下することで図 1 に示したような投稿画面を表示させることができる．

5 データ収集

本システムを Web 上で公開した．この際，ユーザには記述に固有表現を含めないこと，冗長にならないよう各項目を 100 文字以内で記述することを求めた．

25 日間の公開中に，105 件のデータを収集した．登録された作品数は 123 作品であった．同一の作品の組み合わせが 2 件収集されたため，総エッジ数は 103，1 作品あたり 1.675（最大 8，最小 1）であった．収集された作品をマンガペディア[*4]を用いてジャンル分けした．本サイトで分類できなかった場合は Wikipedia，ピッコマ[*5]の順で利用した．作品のジャンルは，バトル（24 件），ラブコメ（13 件），青春（4 件）など多様であった．

5.1 共通点を想起する手がかりの考察

収集された同一ジャンル間の比較データについて，ユーザの記述が共通点・相違点のいずれであったかに着目して分類した結果を表 1 に示す．同様に，非同一ジャンルの場合の分類結果を表 2 に示す．これらの表から，「メインキャラクタ」が共通点として記述された件数は，同一ジャンルでは 23 件，非同一ジャンルでは 36 件であることがわかる．同様に，「世界観」の場合は同一ジャンルでは 30 件，非同一ジャンルでは 25 件であることが確認された．この結果から，ユーザは同一ジャンルの場合「世界観」を，同一でない場合は「メインキャラクタ」を共通点として記述する傾向が確認された．

本データ収集で特に同一ジャンル同士で比較される傾向が強かったのは「推理・ミステリ」ジャンル（7 件）であり，その中で最も多く比較対象に用いられた作品は，『名探偵コナン』(8 件）であった．表 3 に，本作品と比較

図 1: 投稿画面．図は『ドラえもん』のみを選択済み．

図 2: 作品を選択した際の可視化画面．図は『ドラえもん』を選択した場合の結果．

[*4] https://mangapedia.com/ （2025/2/15 確認）．
[*5] https://piccoma.com/web/ （2025/2/15 確認）．

表 1: 同一ジャンル（48 件）の記述パタン別件数

パタン	メイン キャラクタ	目標・結末	世界観	件数
1	共通	共通	相違	8
2	共通	相違	共通	8
3	共通	相違	相違	7
4	相違	共通	共通	14
5	相違	共通	相違	3
6	相違	相違	共通	8

表 2: 非同一ジャンル（57 件）の記述パタン別件数

パタン	メイン キャラクタ	目標・結末	世界観	件数
1	共通	共通	相違	12
2	共通	相違	共通	11
3	共通	相違	相違	13
4	相違	共通	共通	8
5	相違	共通	相違	7
6	相違	相違	共通	6

された推理・ミステリジャンル作品の記述とパタンを示す．「世界観」を共通点とするデータは 3 件確認された．本項目には「事件を解決していく」「リアリティのある世界観」といった記述がされていた．「メインキャラクタ」や「目標・結末」にはキャラクタの所属や行動，行動の動機に関する記述がされていた．このことから，ユーザは類似するジャンルを比較対象とし，「事件を解決していく」といったジャンルの説明を「世界観」の共通点として記述する傾向があることが確認された．同一ジャンルを比較対象として挙げる傾向は，スポーツを題材とした作品でも確認された．これらの作品は，「バレーボール」ジャンルの『ハイキュー！！』や，「サッカー」ジャンルの『イナズマイレブン』などの 14 作品が登録され，14 件の比較記述が得られた．これらのうち，10 件がスポーツ作品同士の組み合わせであり，同一ジャンルは 7 件であった．そのうち 5 件がパタン 4 に分類された．このパタンでの記述は，共通点として「世界観」に「バレー」「バスケ」といった競技名を記述し，「目標・結末」に「全国制覇」などの目標を記述する傾向が確認された．

「メインキャラクタ」を共通点として記述する傾向は，『僕のヒーローアカデミア』（7 件）で最も顕著であった（表 4）．これらのデータでは，全ての記述において「メインキャラクタ」が共通点として挙げられた．本項目では，「能力を持たない」「臆病」といった，主人公の弱点への言及が多く確認された．

以上のことから，ユーザは類似するジャンルやキャラクタの属性を手がかりとして比較対象となる物語作品を選定し，これらの説明をそれぞれ「世界観」「メインキャラクタ」で記述する傾向が確認された．

5.2 相違点の記述傾向の考察

ユーザが類似するジャンルやメインキャラクタ共通点を記述する一方で，どのような相違点を記述し，ストーリーの差異を具体化しているのか考察する．表 1 の同一ジャンル間の比較において，パタン 4（14 件）が比較的多い．表 2 の非同一ジャンル間比較において「メインキャラクタ」を共通点とするパタン（パタン 1，2，3）におけるそれぞれの件数は項目によらず概ね同数であった．以上のことから，ユーザは共通点以外の要素を幅広く相違点として選択し，作品間の差異を記述していることが確認された．

収集された相違点に着目し，記述された内容について述べる『名探偵コナン』との比較で収集された相違点に着目する．（表 3）．「メインキャラクタ」では，「高校生」に対して「医者」といった所属に関する違いや，「自身の頭脳だけで事件を解決する」に対して「死体の声が聞こえ，ヒントをもらえる」といった，事件の解決方法の差異が記述されていた．「目標・結末」では，「元の身体に戻る」に対して「事件を解決する」といった，キャラクタの目標の差異が記述されていた．スポーツを題材とした作品において相違点とされる傾向にある「メインキャラクタ」では，「バスケ初心者」と「バスケ経験者」，「傲慢な高身長スパイカー」と「明るい性格の低身長スパイカー」といった，キャラクタの役割や性格，競技の熟達度の違いを説明する記述が確認された．

『僕のヒーローアカデミア』を比較対象とした記述（表 4）における「目標・結末」に着目すると，同作は「ヒーローを目指す」という旨の記述が 5 件確認された．これに対し，比較対象の記述は，「平穏な日常を守る」「世界征服を目指す」など，多様な目標が相違点として挙げられた．「世界観」では，同作において「特異体質を持っていることが当たり前」といった，特異な体質を持つことが一般的であることに言及する記述が 5 件確認された．本項目は，比較対象の世界観が同様の能力社会である場合には共通点として記述され，「魔法世界」や「日常」である場合には相違点として対比されている．同作の比較記述においては，「能力社会における無能力者」という設定を前提に置き，そこから派生する「ヒーローを目指す」といった行動や主要なストーリーの展開を記述

表 3: 名探偵コナン [A] と比較された推理・ミステリジャンル作品 [B] の記述とパタン

比較対象 [B]	メインキャラクタ	目標・結末	世界観	パタン
鴨乃橋ロンの禁断推理	[共通] 他の人に解決を手伝ってもらう	[共通] 悪の組織に追求していく	[A] 色んな人が殺人を起こしていく [B] 特定の組織が全て裏で動いて事件が起こる	1
四ッ谷先輩の怪談。	[A] 学生が探偵を務め，刑事事件を解決する [B] 成人している可能性が高い学生	[共通] 謎を求め続ける	[共通] 現代が舞台	4
異常死体解剖ファイル	[A] 自身の頭脳だけで事件を解決する [B] 死体の声が聞こえ，ヒントをもらえる	[A] 道具などを駆使し，事件を解決する [B] 能力を使わずに事件を解決する	[共通] 現代で起こる事件を解決していく	6
天久鷹央の事件カルテ	[A] 高校生の名探偵。体が小さくなっていた [B] 職業が医者だが探偵の顔も持つ	[A] 元の身体に戻ること [B] 事件を解決すること	[共通] リアリティのある世界観	6

表 4: 僕のヒーローアカデミア [A] と比較された作品 [B] の記述とパタン（一部）

比較対象 [B]	メインキャラクタ	目標・結末	世界観	パタン
ブラッククローバー	[共通] 能力を持たない少年	[共通] 仲間とともに困難を乗り越えていく	[A] 現代 [B] ほぼ全員が魔法を使う世界	1
魔入りました！入間くん	[共通] 優しく臆病だが勇気をもって立ち向かう	[A] ヒーローになるため頑張る [B] 魔界で生き抜く術を身につける	[共通] 皆が何か特別な能力を持っていてそんな世界の学校が舞台	2
マッシュル-MASHLE-	[共通] 無能力	[A] 最高のヒーロー [B] 平穏な日常を守る	[A] 特異体質をもっていることが当たり前の社会 [B] 魔法を使えないと殺される世界	3
戦隊大失格	[共通] 特段力のない主人公	[A] 最高のヒーローを目指す [B] 世界征服を目指す	[A] 特殊な力を持つことが普遍的な世界 [B] 戦隊モノのヒーローと悪役が実在する世界	3
からかい上手の高木さん	[共通] おっちょこちょいだけど，強い情熱を持つ	[A] 一流のヒーローになることを目指す [B] 普段の変わらない日常を送る	[A] ヒーローの世界 [B] 普段の日常の一部	3

する傾向が顕著に見られた．

以上の結果から，ユーザは既知の作品から類似するジャンルやキャラクタを有する作品を想起した上で，それらの説明を共通点で記述し，キャラクタの特徴や行動に関する詳細な差異を相違点として言語化する傾向があることが明らかになった．特に，「主人公が無能力」といった，物語の根幹をなす要素を共通点として設定した場合，それに基づいた目標や世界観の差異を記述することで，ストーリーの構造的な違いが明示的に収集された．

5.3 記述項目の整合性に関する考察

設定した枠組みの整合性について考察する．『名探偵コナン』と『異常死体解剖ファイル』の比較記述（表 3）では，全項目において「事件を解決する」という記述が含まれた．こうした記述の重複が生じる背景には，物語要素間の不可分な連関がある．物語におけるキャラクタの属性や行動，および舞台設定は一連の構造体として機能しており，それらを厳密に分離することは困難である．例えば「事件を解決する」という要素は，キャラクタの専門性を示す属性であると同時に，目標や結末を表す行動でもある．さらに，この要素は日常的に事件が発生して解決されるという世界観を表す記述でもある．

本システムにおいて発生した記述の揺れは，こうした要素間の連関だけでなく，物語には複数の解釈が存在するという性質にも起因すると考えられる．作品の内容をどの項目として記述するかは，ユーザが作品のどこに重きを置いているかという着眼点を反映している．例えば，同じ「事件解決」を「目標・結末」と捉えるユーザは，事件の発生や解決という展開を重視する一方で，「世界観」と捉えるユーザは，メインキャラクタの属性や行動動機を重視している可能性がある．今後は，設定した記述項目とユーザの着眼点との間にどのような規則性が存在するのかについての調査を検討する．

5.4 収集データの活用に関する今後の展望

本システムで収集されたデータは共通点・相違点を用いた情報アクセスに利用することと，共通点・相違点の自動生成のための学習データとして利用することの 2 点が考えられる．前者は，ユーザが既知の物語作品と共通点を有する未知の作品群へアクセスし，さらに作品間の具体的な相違点を把握可能な手法である．本手法は，設定した 3 項目のうち任意の項目を共通点として固定した上で，作品探索が可能であるという特徴を持つ．例えば,「メインキャラクタ」を共通点として固定すれば，クエリ作品と類似した属性を持つキャラクタが登場する作品群が提示される．その上で，各作品における「世界観」の設定や，キャラクタの行動やその動機の差異（目標・結末）に着目して比較を行うことができる．後者は，収集された共通点・相違点の記述を正解データとし，それらに対応する 2 作品のあらすじ文を比較・照合させる手法である．提案システムで得られた構造化データを教師データとして用いることで，あらすじ文内のどの記述が作品同士を比較する際の指標に該当するのかを同定するモデルの構築が可能となる．

6 おわりに

本稿では，物語のストーリーに関する共通点および相違点を収集するシステム「COM-PASS」を開発した．提案システムでは,「メインキャラクタ」「目標・結末」「世界観」の 3 項目を記述の枠組みとして設定することで，ストーリーとの関連性が高く，構造化されたデータを収集可能とした．今後は収集データをストーリー重視の物語作品への情報アクセスのため利用するほか，物語創作でのアイディア支援への応用についても検討していく．

謝辞

本研究の実施にあたり，JSPS 科研費 24K15255 の支援を受けた．記して謝意を表す．

参考文献

[1] ウラジーミル・プロップ（著）, 北岡誠司, 福田美智子（訳）: 昔話の形態学, 水声社 (1987).
[2] カール・イグレシアス（著）, 島内哲朗（訳）: 「感情」 から書く脚本術 心を奪って釘づけにする物語の書き方, フィルムアート社 (2016).
[3] ブレイク・スナイダー（著）, 廣木明子（訳）: 10 のストーリー・タイプから学ぶ脚本術 SAVE THE CAT の法則を使いたおす！, フィルムアート社 (2014).
[4] ラリー・ブルックス（著）, シカ・マッケンジー（訳）: 工学的ストーリー創作入門 売れる物語を書くために必要な 6 つの要素, フィルムアート社 (2018).
[5] 野村聡美, 両角彩子, 永森光晴, 杉本重雄: マンガのためのメタデータモデルを目指したマンガのアーキテクチャの分析, ディジタル図書館, Vol. 36, p. 3–14 (2009).
[6] 藤川雄翔, 松下光範, 山西良典: あらすじ文に含まれる物語内容の共通要素に着目したコンセプト文の生成, 情報処理学会エンタテインメントコンピューティングシンポジウム 2025 論文集, pp. 10–19 (2025).
[7] 宮川栞奈, 藤川雄翔, 松下光範, 山西良典: 物語展開に伴う登場人物間の関係性変化の可視化, 人工知能学会論文誌, Vol. 40, No. 5, p. MO25–B_1–12 (2025).