日本語文書における生成 AI 由来の構成的特徴の検出とその改善支援

— 対比コーパスによるルール較正とローカル LLM 構造判定の統合 —

長井 英之(一般社団法人オープンデータラボ) / 2026 年 7 月 9 日

要旨

生成 AI による日本語文章の「AI らしさ」を検出し、人間による改善を支援する手法を提案・実装・評価した。AI 生成テキストの真偽判定には理論的限界が示されているため、本手法は真偽判定ではなく改善提案(サジェスト)に目的を限定し、検出ルールを人間可読な設定ファイルとして保持する。正解既知の AI 生成文書 50 本と人間文書 15 本からなる対比コーパスを構築して検証した結果、(1)通説に基づく語彙・リズム系の仮説ルールの相当数が人間文書への誤検知として棄却されること、(2)Markdown 記法の残骸への依存を除くと従来手法の弁別力はほぼ消失すること、(3)文書全体の構成的特徴(三段構成・メタ言及・律儀な結論)をローカル LLM で判定する方式が有効であり、記号を出力しないモデル(ChatGPT 系)への検出力を大幅に改善することを示した。最終的に AI 群スコア中央値 76.5、人間群中央値 2.2、人間群最大値 25.6 の分離を得た。あわせて、LLM 判定プロンプトにおける判定項目リストの削減が残存項目の判定挙動を変化させる現象を報告し、運用上の対処を示す。

キーワード: AI 生成テキスト、日本語、スタイル検出、ローカル LLM、キャリブレーション、説明可能性

1. はじめに

生成 AI の普及に伴い、AI が作成・処理した文章が実務文書に混入する場面が増えている。本研究の目的は、文章の「AI らしさ」を指摘し、人間による改善(リライト)を支援することである。

前提として重要なのは、「このテキストは AI が書いたか否か」の真偽判定を目的としない点である。AI 生成テキストの信頼できる検出は、パラフレーズ等の単純な回避に対して脆弱であり、十分に高性能な言語モデルに対しては理論上最良の検出器でもランダム分類器をわずかに上回る程度に留まることが示されている[1]。この理論的限界を踏まえ、本手法は真偽判定の精度競争には参入せず、「どこが AI っぽく見えるか、直すならどこか」を提示するサジェストに

徹する。スコアも「AI 生成確率」ではなく「AI っぽさ指摘密度」と位置づけ、リライト作業の進捗指標として機能させる。

本研究の貢献は次の 3 点である。第一に、日本語を対象とした対比コーパス方式の検証基盤を構築し、通説に基づくルールの妥当性を定量評価した。第二に、表層記号(Markdown 残骸)への依存を排し、文書の構成的特徴をローカル LLM で判定する方式を統合して有効性を示した。第三に、LLM 判定プロンプトの項目構成に関する再現性のある技術的知見を報告した。

2. 関連研究

AI 生成テキスト検出の理論的限界については Sadasivan ら[1]が、幅広い検出器(watermark 系・ニューラルネット系・zero-shot 系・retrieval 系)のパラフレーズ攻撃に対する脆弱性と、検出性能の理論上の上限を示している。本研究の「検出ではなくサジェスト」という方針は、この結果を設計の出発点とする。

集団レベルの語彙シフト分析としては Kobak ら[2]が、1,500 万件超の PubMed アブストラクトを対象に、LLM 登場前の頻度から予測される値との差分により「超過語彙」を定義する手法を示した。本研究の対比コーパス方式は、この手法の考え方を小規模な日本語コーパスに応用したものと位置づけられる。ただし英語圏で報告される特徴語(delve 等)は日本語にそのまま移植できず、日本語における同種の定量検証は調査した範囲で見当たらない。

実務者コミュニティでは、AI らしさの徴候として語彙・定型句・文末表現の単調さ等が経験的に指摘されているが、いずれも伝聞・印象に基づくもので、正解既知のコーパスによる裏付けを持たない。本研究の結果(4.3 節)は、これらの通説の相当部分が実測で棄却されることを示す。

3. 提案手法

3.1 全体構成

本手法は、(a)前処理としての Markdown 記号除去、(b)人間可読な設定ファイル(JSON)に基づくルールベース検出(語彙 3 階層・リズム・統計)、(c)ローカル LLM による構造判定、の 3 層で構成される。各検出の指摘に付与された重みをカテゴリ別に合成し、文章量で正規化した上

で飽和関数(tanh)により 0〜100 のスコアへ写像する。検出ルールを学習済みモデルの重みではなく設定ファイルとして保持するのは、判定根拠の説明可能性と、LLM の世代交代への追従容易性(再学習不要、ルールの追記・削除のみ)を優先した意図的な選択である。

Markdown 記法の残骸(**・#等)は、チャット AI の画面ではレンダリングされ記号として見えないため、利用者が気づかずコピペすることで文書に残る。検出の手がかりとしては強力だが、表示上の痕跡にすぎず、記号を除去された文章には無力である(4.2 節で定量的に示す)。そこで本手法では記号を前処理で除去し、除去の発生自体は低い重みの加点要素に格下げした上で、主判定を除去後のプレーンテキストに対して行う。

3.2 構造判定

構造判定とは、文書全体の「組み立て方の癖」をローカル LLM(qwen3.5:9b)に読ませて認定させる処理である。三段構成のような段落間の役割関係は、正規表現などの文字列パターンでは形式化できないため、この判定のみを LLM に委ねた。

ただし LLM には「AI が書いたか」という総合判断はさせていない。「三段構成が有るか」「メタ言及表現が有るか」といった個別の事実認定を true/false で答えさせるだけであり、それをどう点数化するかは人間が管理するルールファイル側で決めている。判定の主導権を設定ファイルに残すこの分担は、説明可能性の原則と整合する。また、求めているのが創作や推論ではなく読解と分類であるため、9B パラメータ級の小型モデルで安定して動作する(検証時の応答形式エラーは 0 件、判定 1 回あたり約 5 秒、外部 API への文書送信なし)。判定に失敗した場合はルールベースのみのスコアで処理を完了させ、その旨を結果に明示する。

4. 評価

4.1 実験設定: 対比コーパス

日本語の「AI っぽさ」について正解ラベル付きの公開コーパスは存在しない。そこで、正解が既知のサンプルを自作して対比する方式を採った。AI 群として、主要チャット AI 3 系統(ChatGPT・Gemini・Claude、いずれもデフォルト設定)およびローカル LLM 2 系統に共通のお題で書かせた計 50 本。人間群として、生成 AI 登場以前に書かれた第三者文書(青空文庫の随筆、2019 年の国会会議録)と自社の実務文書、計 15 本。両群に全ルールを適用し、ルールごと

の発火率を群間比較することで、各ルールが「AI 特有の特徴」を捉えているか「人間の正常な書き方への誤検知」かを定量判定した。

なお、当初は自社文書のみで検証する構成を検討したが、それでは「AI 一般の特徴」ではなく「特定個人の文体の癖」を検出する器になってしまう。第三者文書の追加によって初めてこの区別が可能になった。

4.2 表層記号への依存の定量化

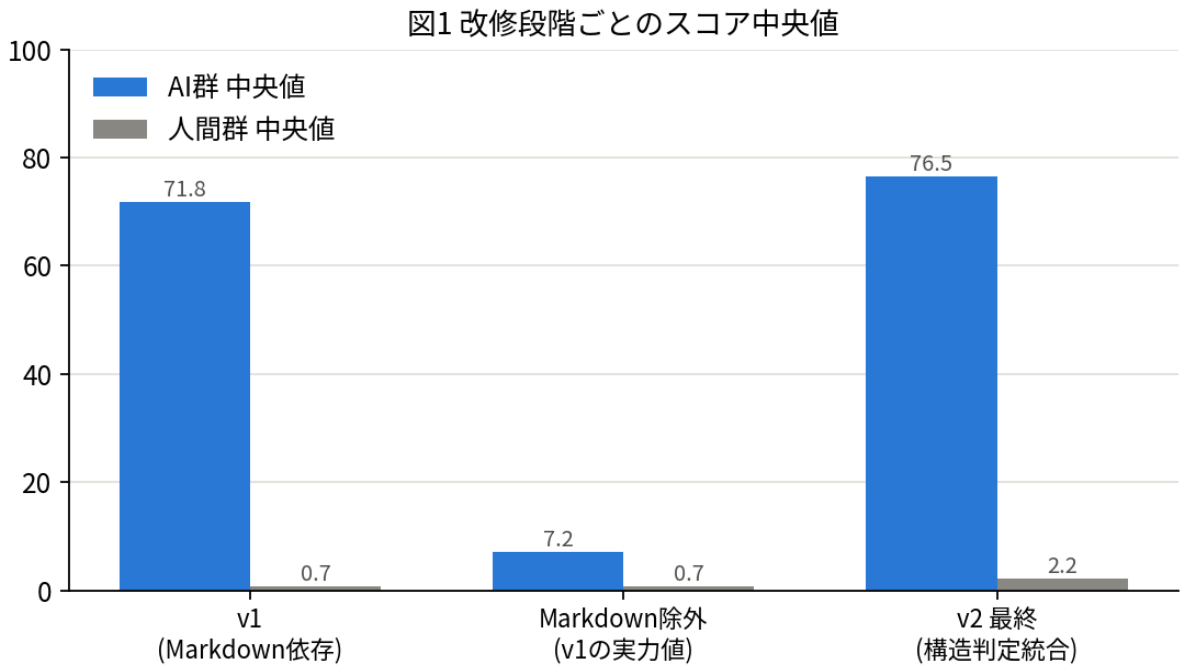
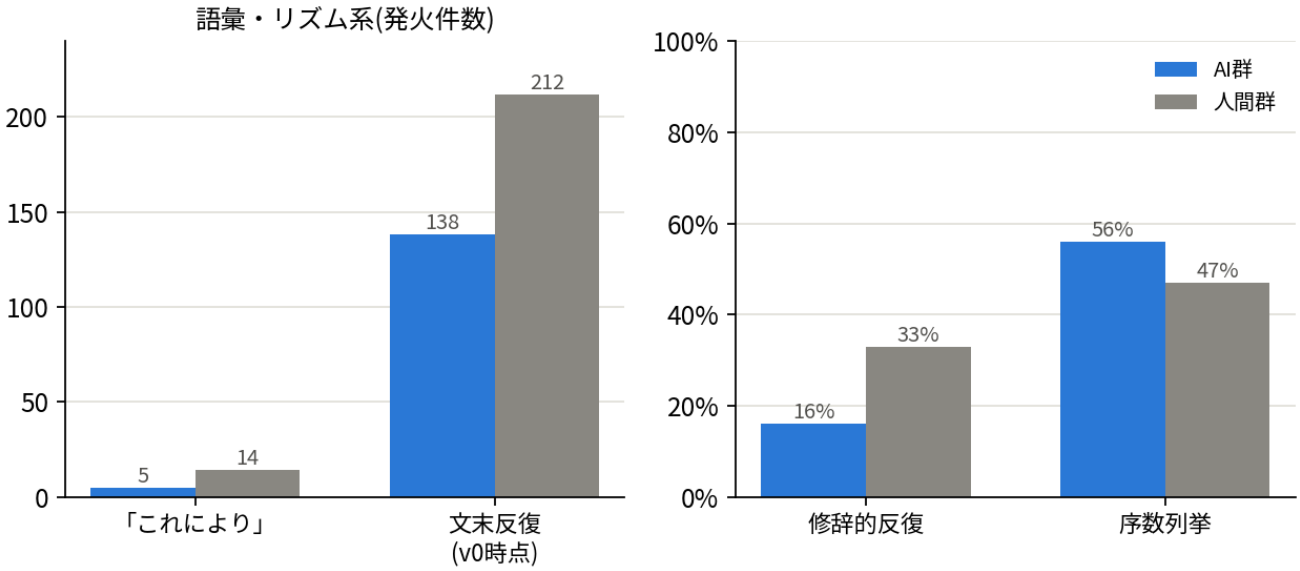


図 1 に改修段階ごとのスコア中央値を示す。改修前(v1)の分離は Markdown 残骸検出への依存で成立しており、当該カテゴリを除外して再集計すると AI 群中央値は 71.8 から 7.2 まで低下し、人間群との分離がほぼ消失する(図 1 中央)。従来手法の実力値は実質的に「記号探し」であったことを意味し、構成的特徴による補完の必要性を裏付ける。

4.3 仮説ルールの棄却

図2 実測で棄却された仮説ルールの例(人間側で同等以上に発火)



通説やネット上の情報から採用した仮説ルールのうち、4 つが実測で棄却された(図 2)。「これにより」の多用(実務者間で「人間はほぼ使わない」とされていた)は、実測では人間側のほうが多く、AI の特徴ではなく個人の書き癖だった。同一語尾(です・ます)の連続は、議事録・答弁・講演録など丁寧体の日本語では人間でも自然に発生する。同型構文の修辞的反復は、随筆の修辞技法や答弁の定型構文を誤検知し、むしろ人間側で高発火した。序数による列挙は実務文書の正常な作法であり、群間の弁別力がなかった。印象ベースのルールは高い確率で人間の正当な文章への誤指摘となる。仮説→実測→棄却/採用のサイクルを検証基盤として持つこと自体が、この種のツールの品質を規定する。

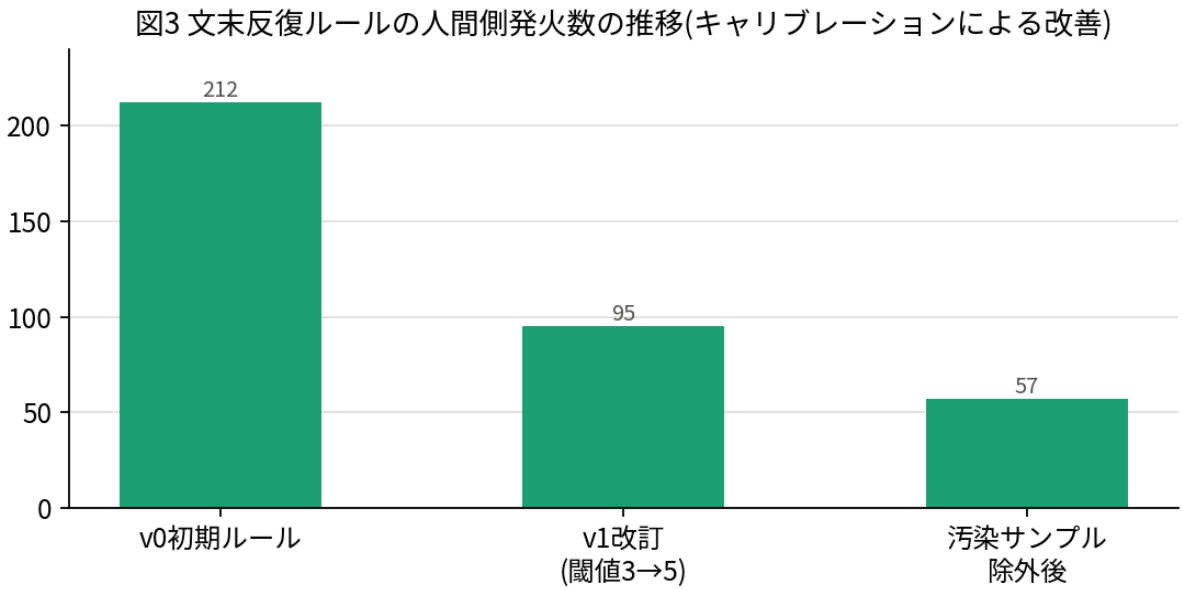
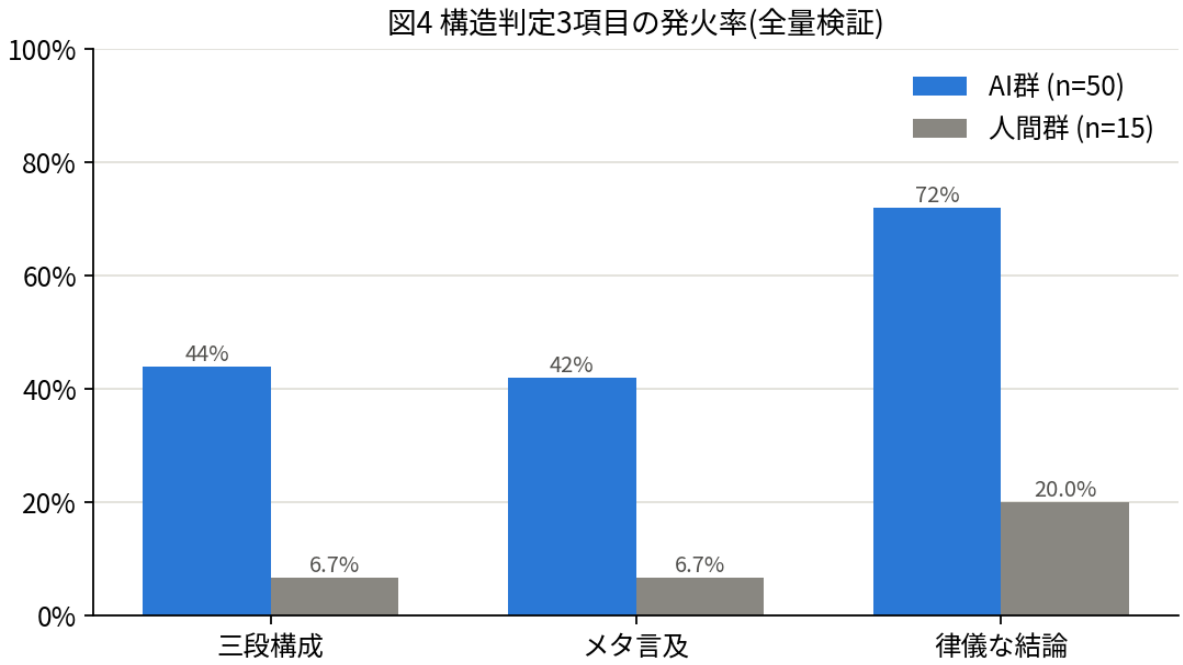
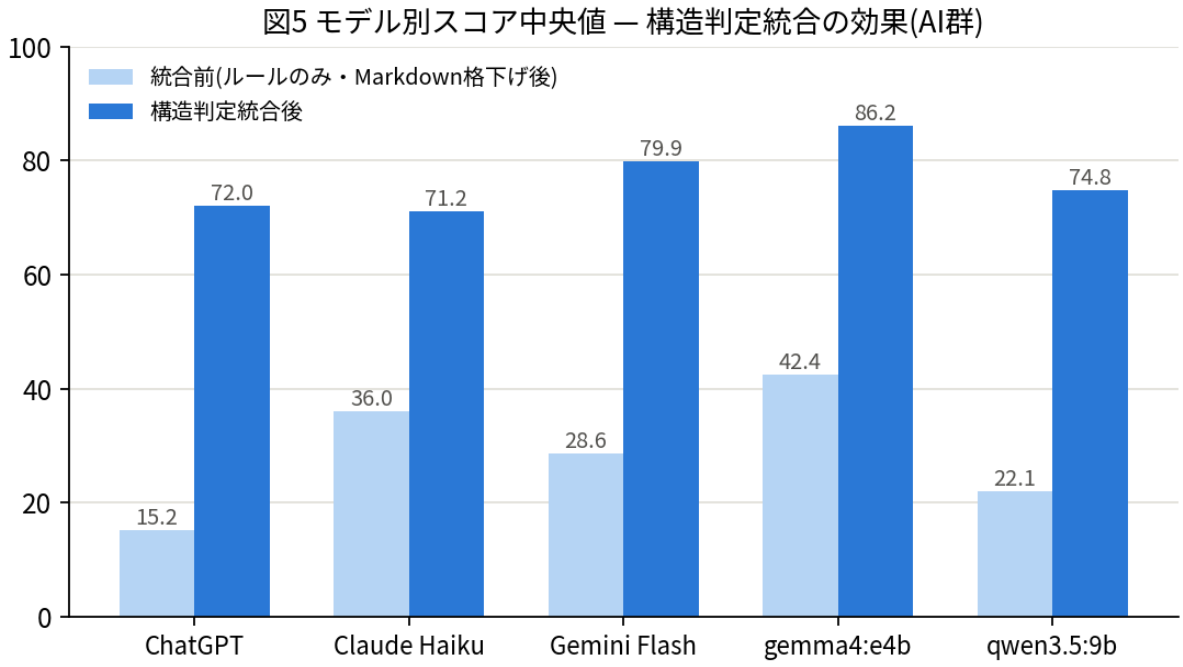


図 3 は文末反復ルールについてこのサイクルが機能した推移である。あわせて、検証中に人間群サンプルの 1 本が外れ値スコアを示し、現物確認の結果、文書後半に AI 下書きの貼り付け(Markdown 記号残存)が混入していたことが判明した。作成者自身が忘れていた混入を検出器が正しく発見した実例であり、同時に「人間作成」という自己申告ラベルすら汚染されうることを示す(6 章の運用設計に反映)。

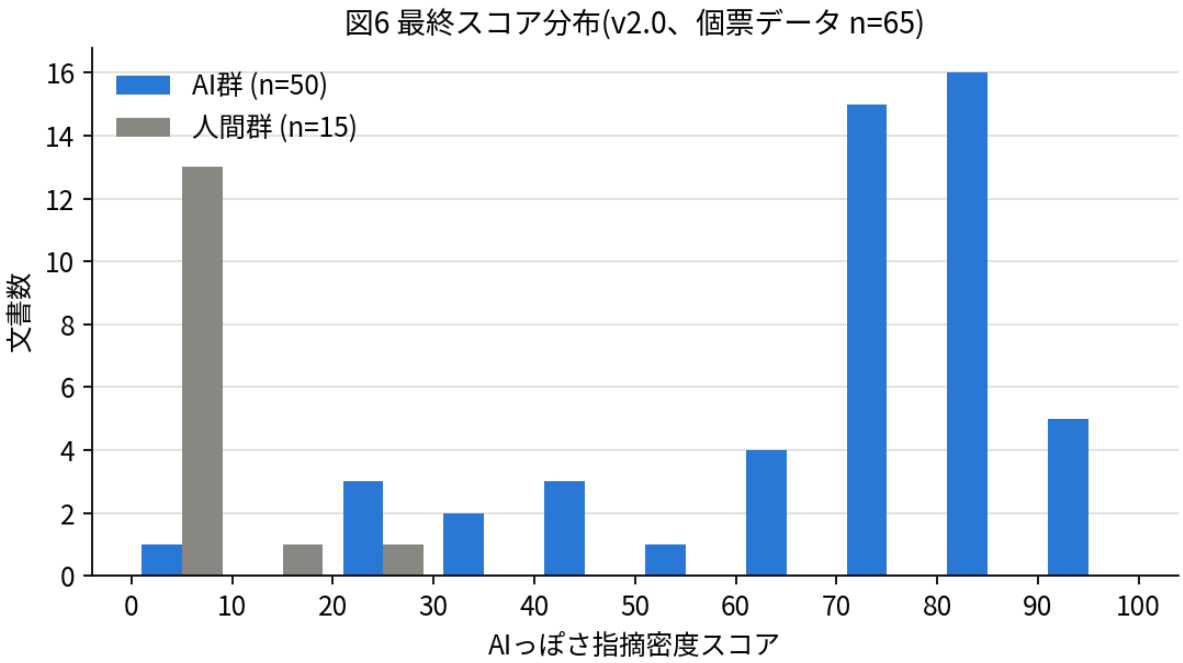
4.4 構造判定の有効性と最終性能



全量検証(AI 群 50 本・人間群 15 本)で有効性が確認された構造判定項目は図 4 の 3 つである。三段構成(導入→本論→結論の型どおりの構成)は AI 群 44%/人間群 6.7%、メタ言及(「まとめると」「本稿では」等、文章が自身の構成に言及する表現)は AI 群 42%/人間群 6.7%、律儀な結論(最終段落を全体要約と一般論・呼びかけで締める癖)は AI 群 72%/人間群 20%で発火した。律儀な結論は感度が最も高いが偽陽性率も相対的に高いため、スコアへの寄与は低めに重み付けした。3 項目の論理和で AI 群の 80%をカバーする。



モデル別の効果を図 5 に示す。Markdown 記法をほとんど出力しないためルールベースで最も検出困難だった ChatGPT 系のスコアが、構造判定の統合により中央値 15.2 から 72.0 まで引き上がり、他モデルと同水準になった。「見た目の記号」ではなく「構成の癖」を捉える方式の妥当性を示す結果である。

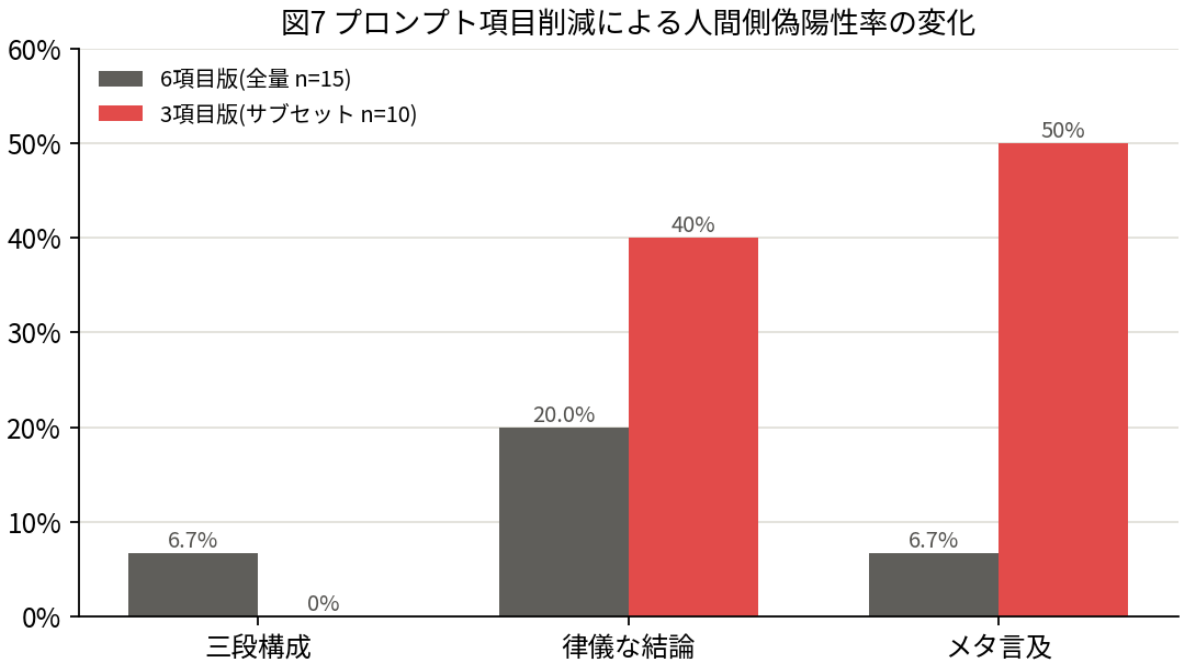


最終的なスコア分布を図 6 に示す。AI 群中央値 76.5、人間群中央値 2.2、人間群最大値 25.6 であり、実用上十分な分離を得た(図 6 は確認計測パスの個票 n=65。LLM 判定の揺らぎにより中央値は計測ごとに数ポイント変動するが、分離構造は安定している)。なお図 6 にはスコア 0 点の AI 文書が 1 本存在する。全ルール・全構造判定をすり抜けた個体であり、本手法の限界(すべての AI 文章を捕捉できるわけではない)を示すものとして記録する。

5. 考察

5.1 LLM 判定プロンプトにおける項目構成の影響

実装過程で得られた再現性のある知見を報告する。構造判定は当初 6 項目で検証し、うち 3 項目を採用・3 項目を廃棄と決定した。実装時にプロンプトを採用 3 項目のみに整理したところ、同一文書に対する残存項目の判定が反転する現象が発生した(図 7。メタ言及の人間側偽陽性が 6.7%から 50%に急増)。



LLM は項目リスト全体を文脈として判定基準を相対的に構成しているため、項目の削除は「無害な整理」ではなく判定挙動の変更となる。対処として、検証済みの 6 項目版プロンプトを一切変更せず使用し、廃棄項目は判定結果を受け取るがスコア重み 0 とする方式を採った(将来の再評価材料も継続的に蓄積される)。検証済みプロンプトからの項目削減には再検証が必須である。

5.2 本手法の限界と、意図的に踏み込まない領域

本手法から「AI らしくない文章」を積極的に定義することは行わない。理由は二つある。第一に、検証設計上それを測っていない。「AI は三段構成を多用する」の対偶は成立せず、AI 群の過半は三段構成なしであり低スコアの個体も存在する。第二に、原理的な問題として、「AI らしくなさ」を仕様として定義できるなら、その仕様どおりに書くよう AI に指示すれば無効化される。検出をやめてサジェストに振ったのと同じ理由で、人間らしさの判定も競争軸にできない。本研究で主張するのは「AI の構成上の癖の検出」までであり、その癖の不在は人間らしさの証明ではない。

また、本評価の AI 群は無修正の AI 出力のみで構成している。正解ラベルの純度と再現性を最優先した設計判断であるが、実運用では人間の修正が加わったテキストが多いと考えられ、本評価の分離性能はその意味で上限値である。修正度合いを段階づけたコーパスでの評価は今後

の課題である。

また、検証コーパスは丁寧体(です・ます調)の文書に偏っており、常体(だ・である調)の書き言葉での検出力は未検証である。方向性の仮説として、AI らしさの根本原因は表層パターンではなく「主張の提案と妥当性判断まで委任した」ことにあり、人間らしさは固有の体験・具体的根拠・独自の優先順位の密度として現れる、という整理を持っているが、これを表層特徴から推定する方法は未確立であり、今後の検証課題である。

6. むすび: 継続的キャリブレーションへ

本研究では、真偽判定の理論的限界を前提に改善支援へ目的を限定した「AI らしさ」検出手法を構築し、対比コーパスによる実測でルールを較正した。今後の品質向上策として、利用者フィードバックの仕組みを設計済みである(実装は次期)。結果画面に「AI 作成だった/人間作成だった/わからない」の 3 択を設け、回答を正解ラベル付きサンプルとして蓄積する。「わからない」は蓄積対象外とし、出所不明文書によるラベル汚染を防ぐ。

これは機械学習における「学習」(モデルパラメータの自動更新)ではない。蓄積データの用途は検出ルールの妥当性検証と較正であり、更新対象は人間可読なルール定義ファイルである。蓄積サンプルによる定期的な再キャリブレーション(全ルールの両群発火率比較)を根拠に閾値・重みを改訂することで、4.3 節と同一のサイクルが運用の中で回り続ける。説明可能性を保ったまま、実際の利用ドメインの文体分布へコーパスが近づく利点もある。安全弁として利用者の申告ラベルは無条件に信用せず、隔離領域への保存と人間のレビューを経て正式コーパスへ昇格させる(4.3 節の混入実例が根拠である)。

参考文献

[1] V. Sadasivan, A. Kaur, S. Balasubramanian, W. Wang, S. Feizi, "Can AI-Generated Text be Reliably Detected?", arXiv:2303.11156, 2023/2024.

[2] D. Kobak, R. González-Márquez, E.-Á. Horvát, J. Lause, "Delving into LLM-assisted writing in biomedical publications through excess vocabulary", Science Advances (arXiv:2406.07016), 2024.