卒業論文

Binoculars を用いた日本語ニュース記事における AI 生成文と人間の文章の特徴比較と判別

小出 彩加

2026 年 2 月 6 日

岐阜大学 工学部 電気電子・情報工学科 情報コース
鈴木研究室

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

小出 彩加

指導教員：
鈴木 優 教授

Binoculars を用いた日本語ニュース記事における AI 生成文と人間の文章の特徴比較と判別*

小出 彩加

内容梗概

本研究では，従来は LLM 生成文と人間生成文を判別するための最終的な判断の材料として使われていた Binoculars スコアを一次の特徴量として再解釈し，機械学習モデルに判別境界を学習させることによって，スコアの分布の不安定性に対処し安定した文章判別を行うことを目的とする．

LLM(Large Language Model，大規模言語モデル) の性能向上により自然な文章が生成されるようになってきたが，LLM 生成文は確かな情報をもとに生成されているとは限らないため，LLM 生成文と人間の文章を判別することは重要であると考えられる．本研究では特に，情報の信頼性が重要視されるニュース記事の領域に着目する．

LLM 生成文と人間生成文を見分けるための指標として perplexity がある．perplexity は文章の意外さという観点から一定精度の文章判別が可能だが，使用するモデルの種類によって値が変動してしまうという問題点を持つため，汎用的に LLM 生成文と人間生成文を見分けることは困難である．そこで，スコア算出に使用するモデルを一つではなく二つにすることで perplexity のモデル依存性に対処している Binoculars スコアという指標がある．しかし，Binoculars スコアもモデルの出力をもとに算出される値である以上，スコア分布に不安定性が生じる可能性があるため，固定の閾値による文章判別は困難である．

そこで本研究では，Binoculars スコアを一次の特徴量として機械学習モデルに判別境界を学習させる二値分類器を提案することで，Binoculars スコアの不安定性に対処し，高精度の文章判別の実現を目指した．

実験の結果，Binoculars スコアの算出に用いる二つのモデルの選択や組み合わ

*岐阜大学 工学部 電気電子・情報工学科 情報コース 卒業論文, 学籍番号: 1223033071, 2026 年 2 月 6 日.

せ，判別対象の変化によってスコアが変動することが明らかになり，従来のようにスコア単体を用いて LLM 生成文と人間生成文の判別を行うことが困難であると示された．また，Binoculars スコアには文章の長さや文字の繰り返しといった特徴に対応する性質があることが観察され，文法や意味的齟齬といった特徴を捉えきれない傾向が示された．さらに，提案手法の二値分類器は比較的古いモデルによる LLM 生成文よりも，比較的新しいモデルによる LLM 生成文と人間生成文を大きく分離する傾向があることが示され，今後新しいモデルが現れても対応できる可能性が示唆された．最終的な研究成果としては提案手法はある程度の有用性を示したが，判別対象は日本語ニュース記事という領域に限られる．本研究における提案手法では機械学習モデルに Binoculars スコア以外の特徴量を追加することも可能なため，今後新たな特徴量を追加することで様々な文章の特徴を加味して LLM 生成文と人間生成文の判別を行うことが可能であると考えられる．

キーワード

perplexity，Binoculars，LLM，LLM 生成文，機械学習モデル

目次

図目次

表目次

第 1 章　はじめに

本研究の目的は，従来は LLM 生成文と人間生成文を判別するための最終的な判断の材料として使われていた Binoculars スコアを一次の特徴量として再解釈し，機械学習モデルに判別境界を学習させることによってスコアの分布の不安定性に対処し安定した文章判別を行うことである．

近年の LLM の性能向上により，LLM 生成文は語彙や文法などが自然な文章になってきている．LLM 生成文と人間生成文を判別するための手法は既に存在するが，LLM はそれぞれ異なる語彙や確率分布を持つため，判別できる対象は特定のモデルによる文章や特定の話題の文章に限られる．例えば Gutiérrez-Megías らによる研究 [1] では，perplexity という判別指標が判別対象を作成するモデルや判定を行うモデルによって影響を受けることを述べている．このように，多様なモデルによる文章や多様な話題の文章を網羅的に判別できる手法は現在確立されていないと言える．

いくつかの領域においては LLM 生成文と人間生成文を判別することは重要な意味を持つ．例えば，学校の課題レポートについて，LLM 生成文を生徒自身が書いたと偽って提出されるケースがあるため，正しい評価をするためには人間生成文か否かを見分ける必要がある．また，インターネットの商品レビューについても，人間の使用者が実際に商品を使用してみた感想を知る目的で読む人が多いため，実際の体験に基づかない LLM 生成文を見分けることが重要である．さらに，ニュース記事は情報源が確かであることが求められるが，LLM 生成文は確かな情報源に基づいて書かれていない場合もあるため，情報の信頼性の一つの尺度として LLM 生成文か否かを見分けることは重要である．このように，文章を作成したのが人間か LLM かを判別することには意義があるので，文章判別のための手段が必要であると考えた．

LLM 生成文と人間生成文を判別するために，これまで様々な研究が行われている．例えば perplexity や Binoculars スコアという指標に基づいて文章判別を行う手法が存在する．これらの指標は文章判別において一定の有用性を示しているが，スコア算出に用いるモデルによってスコアの分布が変動するという問題のため，判別性能には限界がある．

perplexity とはモデルにとっての文章の意外さを表す指標であり，文章判別にある程度有用であるとされている．ただし，perplexity にはスコアが判定モデルに依存するという特性がある．この特性により，同じ文章に対しても判定モデルによって異なる値が示されるため，汎用的に LLM 生成文と人間生成文を判別することが困難である．このようなモデル依存性の問題は生成モデルの多様化に対応できないという側面も持つ．そのため perplexity 単体では安定した文章判別ができるとは言えない．

perplexity のモデル依存性という問題点に対し，スコア算出に用いるモデルを一つではなく二つ用いることで対処を試みる Binoculars スコアが提案されている．Binoculars スコアでは，算出に用いるモデルはそれぞれ performer と observer という役割を担う．それぞれのモデルの出力を用いることで，意外さを表す値とモデル差を表す値の比を取ってモデル依存性を弱めるような設計である．しかし Binoculars スコアもモデルの出力を用いている以上，算出に用いるモデルの選択や組み合わせによって分布が変動する可能性が考えられる．スコアの分布が変動的である場合，スコアの絶対値から文章判別を行ったり，スコアの確率的解釈によって文章判別を行うことは難しい．

本研究では，従来は LLM 生成文と人間生成文を判別するための最終的な判断材料として使われていた Binoculars スコアを，一次の特徴量として再解釈することによって，固定の閾値による判別では対応できない Binoculars スコアの分布の不安定性に対処し，高精度な文章判別を実現することを目指した．具体的には，Binoculars スコアを一次の特徴量として機械学習モデルに判別境界を学習させることで，データの分布に基づいた柔軟な文章判別を行う二値分類器を作成した．

評価実験では，Binoculars スコアの性質の確認と，Binoculars スコアを一次の特徴量として機械学習モデルに判別境界を学習させた二値分類器が，高精度な文章判別ができるかどうかを検証することを目的とし，段階的な検証を行った．

実験の結果，Binoculars スコアの算出に用いるモデルの役割である performer と observer のうち，observer モデルによりリリース日が新しくパラメータ数の多いモデルを配置することと，比較的新しいモデルを用いて作成された文章が含まれたデータセットを用いることという条件を設けることで，二値分類器は一定の判別精度を示すことが明らかになった．他にも，Binoculars スコアは文章の長さや同じ

文字の繰り返しといった特徴に応じて変動し，文法や意味的な齟齬とは関係しない可能性が高いことが示唆された．

本研究の貢献は以下の通りである．

- 従来 LLM 生成文と人間生成文を判別するための最終的な判断の材料として使われていた Binoculars スコアを一次の特徴量として再解釈し，機械学習モデルに判別境界を学習させる手法を提案した．
- Binoculars スコアが算出に用いるモデルの選択や組み合わせ，判別対象の変化によって分布が変動するという性質を持つことを示し，固定の閾値による従来手法の判別には限界があることを明らかにした．
- 機械学習モデルによる判別境界の推定というアプローチが，日本語ニュース記事という領域における文章判別に有効である可能性を示した．

本論文の構成は以下の通りである．2 章では，基本的事項について述べる．3 章では，関連研究について述べる．4 章では，提案手法について述べる．5 章では，評価実験について述べる．最後に 6 章では本論文のまとめと今後の展望について述べる．

第 2 章　基本的事項

2.1　LLM 関連

2.1.1　LLM

LLM[2] とは Large Language Model の略であり，日本語では大規模言語モデルとも称される．代表的な例には ChatGPT* や Gemini† が挙げられる．文章を生成したり予測したりすることのできるモデルである．

モデルは，自身が学習した確率分布に基づいて単語を生成したり予測したりする．例えば，「今日は」に続く言葉として「晴れ」が 0.7，「運動会」が 0.2，「ございます」が 0.1，というような確率で出力されることを予測した場合，最も高い確率で予測された「晴れ」が続くであろうと考えられる．

具体的な処理としては，文章をトークンという単位に分割したものをモデルに入力すると，モデルはそのトークンを文章中の位置情報も含めた一つのベクトルへと変換する．次に transformer という仕組みによって文脈が考慮された上で，softmax 関数という関数によってトークンが次に出力される確率が出力される．

文章を途中まで与えられたとき，次に来る確率が最も高い単語を選んで繋げていくという LLM の仕組み上，情報源が明確でない文章や，文法が間違っている文章，話のつじつまが合わない文章が生成される可能性がある点に注意する必要がある．

2.1.2　トークナイザー

トークナイザーとは，文章をトークンに分割するための仕組みである．トークンとは，LLM が文章を処理できるように分割した単位であり，単語と一致する場合が多いが，必ず一致するとは限らない [3]．例えば，「こんにちは元気ですか」という文章をトークンに分割する場合，「こんにちは」「元気」「です」「か」のように分割される場合もあれば，「こんにちは」「元」「気」「ですか」のように分割される場合もある．分割の方法はトークナイザーによって異なる．

*https://chatgpt.com/

†https://gemini.google.com/

トークナイザーは，LLM が効率よく学習を行えるような分割方法を選択する．LLM の学習とは，文章を途中まで与えられたとき，次のトークンをどのくらい高確率で当てられたかが計算され，予測の誤りが最小になるようにモデル内のパラメータが更新されていくことである．文章の分割方法が不適切であれば，各トークンについての学習機会が少なくなり，モデルが正確な予測をできるようになりづらいため，トークナイザーは学習が適切に行われるような分割方法を採用する方が好ましい．

加えて，分割する数によってもトレードオフの関係が存在する．分割数が少なければ，処理回数が減ったり，LLM の内部で扱う埋め込み行列のサイズが小さくなったりして計算量が少なくて済む．分割数が多ければ，未知語が発生しにくく，極端なことを言えば 1 文字ずつに分ければ未知語はほぼ発生しない．

トークナイザーは，以上のような LLM の学習効率や分割数のトレードオフを考慮して分割方法を決定する．

2.1.3 perplexity

perplexity[4] は，LLM 生成文と人間生成文の判別において重要な指標である．モデルが文章中の次のトークンをどれだけ予測しづらいかを表す値であり，計算式は以下のように表される．N はトークン数，$P(w_i)$ は位置 i より前の全てのトークンが与えられていたときの，次に w_i というトークンが出力されるとモデルが予測した確率である．

$$perplexity = \exp(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i))$$

モデルが文章中の次のトークンを予測する際には，それより前の文脈をもとに，自身が持つ語彙の中の全てのトークンについて次に出力される確率を計算し，確率分布を生成する．この確率分布が平坦であれば予測が難しく，特定のトークンに確率が集中していれば予測が容易ということになる．perplexity は，この確率分布の中から正解トークンの出力を予測できた確率を文章中の全トークンについて抜き出し，文章中のトークン数で平均を取ることで，文章の長さに依存しない指標を実現

している．

また，計算式から分かるように，正解トークンに割り当てる確率が低いほど perplexity の値は大きくなり，モデルが自信をもって推測できなかったことを示す．

2.1.4 Binoculars

Binoculars[5] は，Hans らによって研究された，LLM 生成文と人間生成文を判別するための手法である．前項の perplexity では，判別をするモデルが何であるかによって値が上下してしまうという問題点があった．しかし Hans らの研究では，この問題点を解決するために Binoculars スコアという指標を考案している．Binoculars スコアでは，あるモデルから見た文章の意外さを表す値と，別のモデルとの挙動の差を表す値との比をとることで，モデル依存性を抑えることを試みている．モデル同士の挙動の差を計算するために，Binoculars スコアの算出には二つのモデルを用いるが，これらのモデルはそれぞれ performer と observer という役割を担う．先述した文章の意外さを表す値を算出するのに用いるのが performer モデルであり，モデルの挙動の差を表す値を算出するのには performer モデルと observer モデルの両方を用いる．

Binoculars スコアの算出手順は以下の通りである．([5] より式を引用)

1. モデルを二つ用意する．以降 M_1 を performer モデル，M_2 を observer モデルとして扱う．
2. 入力文章 s をトークナイザー T によってトークンに分解し，リスト $T(s)$ に格納する．Y は文章中の各位置におけるモデルの次のトークン予測を表す二次元配列であり，i は文章中のトークンの位置，j はモデルが持つ語彙 V 中のインデックスを示す．モデルが x_0 から x_{i-1} までのトークンを与えられたとき，v_j の出力を予測した確率が Y_{ij} である．

$$M(T(s)) = M(\overrightarrow{x}) = Y$$

$$Y_{ij} = P(v_j | x_{0:i-1}) \ for\ all\ j \in V$$

3. performer モデルから見た入力文章の意外さを表す log-perplexity を求める．log-perplexity は，文章中の各位置でモデルが正解トークンを予測でき

た確率の対数尤度を，トークン数で平均したものである．トークン数で平均するのは，文章の長さへの依存を抑えるためである．

$$\text{logPPL}_M(s) = -\frac{1}{L}\sum_{i=1}^{L}\log P(Y_{ix_i}),$$

$$\text{where } \overrightarrow{x} = T(s),\ Y = M(\overrightarrow{x}),$$

$$\text{and L} = \text{number of tokens in } s.$$

4. performer モデルの予測と observer モデルの予測の差を表す log-cross-perplexity を求める．log-cross-perplexity は，M_1 が文章中の位置 i で出力した予測確率分布を正解とみなしたときの M_2 の位置 i での予測確率分布とのクロスエントロピーである．これも log-perplexity と同じくトークン数で平均することで文章の長さへの依存を抑えている．

$$\text{logX PPL}_{M_1,M_2}(s) = -\frac{1}{L}\sum_{i=1}^{L} M_1(s)_i \cdot log(M_2(s)_i)$$

5. log-perplexity を log-cross-perplexity で割ることで Binoculars スコアを算出する．log-perplexity は performer モデルから見た文章の意外さであり，performer モデル固有の確率分布に強く依存する．log-cross-perplexity は performer モデルを基準とした observer モデルとの予測確率分布の差である．両者ともに performer モデル固有の確率分布に由来する影響が含まれているため，これらの比を取ることによって，performer モデルによる影響を弱めたスコアを算出することができる．結果としてこのような手順で，判別をするモデルに依存しにくい，文章そのものの性質に焦点を当てたスコアを構成することができる．

$$Binoculars_{M_1,M_2}(s) = \frac{\text{logPPL}_{M_1}(s)}{\text{logX PPL}_{M_1,M_2}(s)}$$

performer モデルと observer モデルにどのようなモデルの組み合わせを置くかによって，Binoculars スコアの分布は変動する．

2.2 統計

2.2.1 t 検定

t 検定は，二つのデータ群の平均値の間に統計的に意味のある差が存在するかどうかを調べる分析手法である [6]．本研究では，いくつかの種類がある t 検定の中でも，二つの独立したデータ群に用いる対応のない 2 標本 t 検定を用いる．本研究では Binoculars スコアという LLM 生成文と人間生成文で分散が大きく異なる指標を使うため，対応のない 2 標本 t 検定の中でも等分散性を仮定しない Welch の t 検定を用いる．

Welch の t 検定では，二つのデータ群の平均に有意な差はないという帰無仮説を設け，この帰無仮説が棄却される確率を算出することで有意差があるかどうかを判断する．まず，以下の式で示される t 値を算出する．ここで，$\overline{x}_1$ は集団 1 の平均，$\overline{x}_2$ は集団 2 の平均，s_1^2 は集団 1 の分散，s_2^2 は集団 2 の分散，n_1 は集団 1 のサンプル数，n_2 は集団 2 のサンプル数である．

$$t_0 = \frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

次に，帰無仮説が正しいと仮定した時に取り得る t 値全体を表す確率変数を T とする．このとき，観測された t_0 以上に極端なデータが得られる確率，すなわち p 値を，以下の計算式で算出する．

$$p = 2 \times P(T \geq |t_0|)$$

この p 値が，通常 1% や 5% と定められる有意水準より小さかった場合，観測された平均との差は偶然によるものとは考えにくいため，二つのデータ群の間に有意な差が存在すると判断できる．逆に，p 値が有意水準以上である場合，有意差があるとは言えない．

2.3 分類モデル

2.3.1 SVM

SVM[7] は，Support Vector Machine の略であり，データを分類するための判別境界を学習する機械学習手法である．データを最もよく分離する境界線とその境界線にもっとも近いデータ点とのユークリッド距離 (マージン) を最大化するように学習することで，最適な境界線の引き方を求める．

境界線の式は以下のように表される．ここで，x は入力されたデータ，y はクラスラベル，w は重み，b はバイアスを示す．

$$y = g(x) = \begin{cases} 1 & (f(x) \geq 0) \\ -1 & (f(x) < 0) \end{cases}$$

$$f(x) = w^\top x + b$$

また，境界線の最近傍点を x_i としたとき，マージンは以下のように表される．

$$M = \frac{\left| w^\top x_i + b \right|}{\|w\|}$$

SVM ではこのマージンを最大化することを考える．計算式より，i 番目のデータが正確に分けられていれば，i 番目のデータ x_i とその正解ラベル y_i を用いて，$y_i f(x_i) > 0$ という式が成り立つはずである．ここで，w と b を同じ定数倍しても M は変わらないため，M の分子を 1 と置く．すると $M = \frac{1}{\|w\|}$ の最大値を求めることは $\|w\|$ の最小値を求めることと同義となる．この最小化問題を解決することで w と b が最適化され，汎化性能の高い境界線を推測することができる．

2.3.2 ロジスティック回帰

ロジスティック回帰 [8] は，データを二つのクラスに分類するために用いられる機械学習手法である．データをクラス 0 とクラス 1 の二つに分類する際，各データがクラス 1 に分類される確率が 0.5 以上ならクラス 1 に，0.5 未満ならクラス 0 に分類するという，確率に基づいたアプローチをとる．

計算式は以下の通りである．ここで，p はデータがクラス 1 に分類される確率，w は重み，x はデータの説明変数である．

$$y = \begin{cases} 1 & (p \geq 0.5) \\ 0 & (p < 0.5) \end{cases}$$

$$p = \frac{1}{1 + \exp(-w^\top x)}$$

ロジスティック回帰において w を最適化するには，正解ラベルを含む教師データを使い，損失を最小化するための学習を行う．損失は以下の式で表される．なお，N はトレーニングデータの個数であり，x_i は i 番目のデータの説明変数，y_i は i 番目のデータの正解クラスである．

$$-\log L = -\sum_{i=1}^{N} [y_i \log p_i + (1 - y_i) \log(1 - p_i)]$$

上の式で表せる損失を，勾配降下法等を用いて最小化することで適切な w を求め，適した境界線を決定することに繋がる．

2.4 評価指標

2.4.1 Accuracy

Accuracy は，二値分類において分類器がどのくらいデータを正しく分けられたかを示す正解率である [9]．計算式は以下のように示される．TP は True Positive，つまり分類器の予測がクラス 1 で正解もクラス 1 だった場合を表す．FP は False Positive，つまり分類器の予測がクラス 1 で正解がクラス 0 だった場合を表す．FN は False Negative，つまり予測がクラス 0 で正解がクラス 1 だった場合を表す．TN は True Negative，つまり予測がクラス 0 で正解もクラス 0 だった場合を表す．

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2.4.2 AUC

AUC[10] とは Area Under the Curve の略であり，Curve とは ROC 曲線のことを指す．ROC 曲線は，横軸に FP の割合を示す FPR，縦軸に TP の割合を示す TPR をとり，二値分類における判別境界 (閾値) を変化させていったときの推移をプロットしたグラフである．AUC の値はこの ROC 曲線の下側面積と等しい．AUC では閾値に依存しない評価ができることが特徴であり，ランダムに選んだ正例が，ランダムで選んだ負例よりも高いスコアを持つ確率と解釈することもできる．

AUC，FPR，TPR の計算式は以下の通りである．

$$AUC = P(\text{正例} > \text{負例}), 0 \leq AUC \leq 1$$

$$FPR = \frac{FP}{FP + TN}$$

$$TPR = \frac{TP}{TP + FN}$$

第 3 章　関連研究

LLM 生成文と人間生成文の判別に関する研究が既にいくつか存在している．

perplexity は，LLM 生成文と人間生成文を見分けるための代表的な指標の一つである．文章を作成する際，LLM は自身の持つ確率分布に基づいて，次に来る確率が最も高いトークンを選択し連ねていくのに対し，人間は特定の確率分布に基づいて言葉を選択するわけではない．このような文章の作り方の違いから，意外さという観点を数値化した perplexity には一定の判別能力があると考えられている．

しかし，perplexity には，判別を行うモデルが何であるかによって値が変化してしまうという問題がある [1]．perplexity における意外さというのはあくまで判別するモデルにとっての意外さであり，同じ文章を対象とした場合でも，あるモデルにとっては意外性が高かったが他のモデルにとっては意外性が低かった，というような事象が起こり得るためである．

perplexity のモデル依存性という問題に対処するために考案された，Binoculars スコアという指標も存在する．Binoculars スコアは，算出に用いるモデルを一つではなく二つにして，文章の意外さを表す値とモデル差を表す値の比を取ることで，モデル依存性をある程度抑えた指標である．

Hans ら [5] は，この Binoculars スコアという指標を考案することで perplexity のモデル依存性の問題への対処を試みている．Hans らは perplexity について，モデル依存性の他に文章作成の際のプロンプトによるスコアの変動も問題として挙げており，Binoculars スコアを用いればモデル依存性やプロンプト依存性をある程度低減できることを示している．また，ゼロショットでの高精度な LLM 生成文と人間生成文の判別の実現にも有効性を示している．

Hans らの研究では Binoculars スコアについて，LLM 生成文と人間生成文の判別のための指標として一定の有用性が示されている．perplexity のモデル依存性に対処し，相対的に安定した指標を提示した点がこの研究の功績である．

一方で，この先行研究には問題点も存在する．Binoculars スコアに用いる二つのモデルの選択方法や組み合わせ方について，先行研究では，パフォーマンスが近いモデルを使用したほうが良いという以上に明確な言及がない．Binoculars スコアはモデルの出力から算出される値であるため，使用するモデルや組み合わせを固定

しなければ，スコアそのものやスコアの分布が変動してしまう可能性が高い．スコアの分布が不安定である場合，固定の閾値によって，あるいはスコアを確率的に解釈して LLM 生成文と人間生成文を判別することが困難である．にもかかわらず，従来 Binoculars スコアは LLM 生成文と人間生成文を判別するための最終的な判断材料として用いられている．

スコア算出に用いるモデルについての言及が無いことで，モデルの組み合わせによってスコアの分布がどう変化するのかが予測できないという問題も生じている．新しいモデルが登場した時にどのように使用すれば有用か，あるいは使用の方法によらず有用性が期待できないのかが不明であるという懸念も存在する．Binoculars スコアを算出するのに用いるモデルの選択方法や組み合わせ方について厳密に定められていないことが，スコアの分布の様子を事前に見積もることができないことにつながり，結果として実用的な閾値設定が困難になっていると考えられる．

また，Binoculars スコアは英語の文章に対して考案された指標であり，ほかの言語において実用可能な指標であるかは不透明であった．岩田らは，Hans らの研究をもとに，Binoculars スコアの算出に日本語モデルを用いることで，日本語の文章に対しても有効性があることを論じている [11]．言語体系の異なる文章においても Binoculars スコアが有用であると示し，研究例の少ない日本語の文章における LLM 生成文と人間生成文の判別に言及したことがこの研究の成果であると言える．

一方で，この研究においても使用するモデルの選択や組み合わせ方について明確な言及がなく，モデルの選択や組み合わせ方によってスコアがどのような影響を受けるのかが不透明である．論文中では文章の話題や文字数によって判別精度が変化する可能性にも言及されているが，これらの問題に対する十分な検討は行われていない．

本研究では，Binoculars スコアの有用性に着目しつつも，LLM 生成文と人間生成文の高精度な判別を実現するには，Binoculars スコアの持つ問題点に対処することが重要であると考えた．これまでに述べた問題として，モデル依存性やモデルの組み合わせ方によるスコアへの影響，スコアの分布の不安定性や固定の閾値を設けるのが困難であることといったことが挙げられる．これらの問題点の中でも，本研究では特に，使用モデルの選択や組み合わせによるスコアの分布の不安定性に対処する必要があると判断した．そこで本研究では，Binoculars スコアを一次の特徴量

として再解釈し，SVM やロジスティック回帰といった機械学習モデルに判別境界を学習させる．機械学習モデルの学習によって，固定の閾値を定めるのではなく，データの分布に基づいた最適な判別境界を自動推定することができると考えたためである．Binoculars スコアの算出に用いるモデルの選択や組み合わせによってスコアの分布が変動した場合でも，トレーニングデータに基づく判別境界が再学習されるため，柔軟な対応が可能となる．

従来の研究では，Binoculars スコアは LLM 生成文と人間生成文を判別するための最終的な判断材料として直接用いられていた．それに対して，本研究では Binoculars スコアを一次の特徴量として再解釈する．このアプローチによって，従来の研究の中で十分な検討がされていなかったモデル依存性やスコアの分布の不安定性に対処し，データに適応した判別境界を自動推定することが可能になり，結果として LLM 生成文と人間生成文の高精度な判別が実現できると考える．

第 4 章　提案手法

本研究では，日本語ニュース記事における LLM 生成文と人間生成文を精度良く判別するため，Binoculars スコアを一次元の特徴量と見なし，教師あり機械学習モデルを用いてデータに柔軟に対応した判別境界を推定する手法を提案する．

文章判別において，モデルにとっての文章の意外さを表す指標である perplexity はある程度有用であることが知られている．しかし，perplexity には判定するモデルに対する依存性が強いという問題がある．Binoculars スコアはこの問題を抑制した指標であり，先行研究において一定の有用性が示されている．

しかし，Binoculars スコアが perplexity をもとにした指標である以上，完全にモデル依存性を無くすことは困難であると考える．そこで本研究においては，Binoculars スコアを算出するのに用いるモデルの組み合わせ方によってスコア分布が変動するという仮定を置き，5 章で仮定を検証するための実験を行う．また，Binoculars スコアによってどのような文章が LLM 生成文と判断されるのかについても，Binoculars スコアが perplexity をもとにしていることから，perplexity

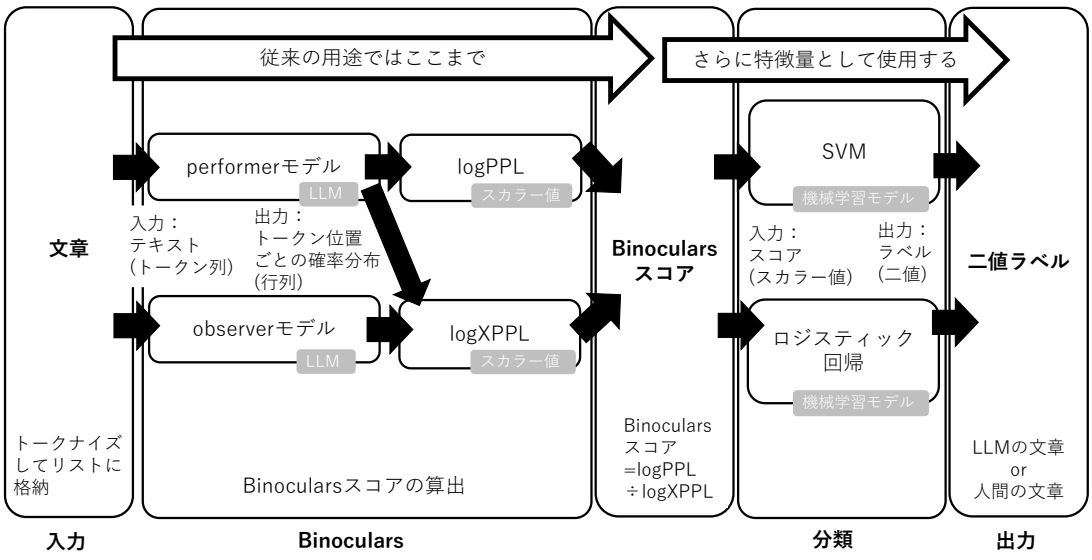

図 4.1　提案手法の流れ

と同様にモデルにとっての文章の意外さという観点で文章判別を行うという仮定を置き，5 章で仮定を検証するための実験を行う．

Binoculars スコアの分布が不安定であるという仮定は，perplexity の特性をもとにしている．perplexity は，モデルにとって入力文章がどのくらい意外だったかを数値化した指標である．文章を作成するとき，LLM は自身の持つ確率分布に基づいて文章の続きとして出力する確率が一番高いトークンを選択するのに対し，人間は特定の確率分布に基づいて言葉を選ぶわけではない．この違いから，モデルは人間生成文をより意外であると解釈する傾向にあり，結果として perplexity の値は通常 LLM 生成文に低い値，人間生成文に高い値を示す．Binoculars スコアは，算出にモデルを二つ用いることでこの perplexity の持つモデル依存性を抑えた指標ではあるが，モデルの出力を用いていることには変わりない．そのため，同じ文章に対してモデルごとに出力の高低関係は抑えられる一方，分布自体の高低や分布の幅に不安定性が出るのではないかと考えた．

また，Binoculars スコアが，perplexity と同様にモデルにとっての文章の意外さという観点で文章判別を行うという仮定も，perplexity の特性に基づいている．モデルにとって意外であるとは，モデル自身の持つ確率分布からすると文章の続きに来る確率が低いトークンが出力されているということである．Binoculars スコアでは perplexity のモデル依存性を抑える操作を加えているが，意外さを測るという特性を打ち消すような操作はしていないため，Binoculars スコアもモデルにとっての文章の意外さという観点で文章判別を行うと予想する．ただし，具体的に文章のどのような特徴がモデルにとって意外であると判断されるのかは不明なため，5 章での実験に伴い，実際に文章を観察して人の目から見た LLM 生成文の特徴を探す．

なお，本研究においては日本語ニュース記事を判定対象としている．丁寧な文体か砕けた文体か，使っている言語は何か等によって，モデルにとっての LLM 生成文らしさ，あるいは人間生成文らしさが変化する可能性がある．しかし，本研究では言語の違いや文体の丁寧さの違いについては検討しないため，あくまで日本語ニュース記事に対する提案手法が文章判別に有効であるかを見る研究であることに留意する．

本研究では Binoculars スコアの有用性を認識しつつも，高精度な判別の実現のためにはスコア分布の不安定性に対処することが重要であると考えた．スコア分布

が不安定である場合，スコア単体を見て対象の文章が LLM 生成文か否かを判断したり，スコアを LLM 生成文である確率として解釈したりすることは困難であるためである．そこで本研究では，Binoculars スコアを従来のように判別のための最終的な判断材料として扱うのではなく，一次元の特徴量として再解釈する．そして，正解ラベルに基づいて誤分類を最小化する判別境界の学習を機械学習モデルに行わせることで，スコアの絶対値による判断や確率的解釈に依存しない，データに柔軟に対応した文章判別が可能な二値分類器の実現を目指す．機械学習モデルを用いることで，Binoculars スコア単体に基づく判別にとどまらず，将来的にその他の特徴量を追加可能な特徴量空間に基づいて判別境界を設定する枠組みを提供することに本研究の独自性がある．

本研究における文章判別とは，入力された文章 s に対して，以下の二値ラベル y を出力することである．y は一次元の特徴量 x をもとに推定された判別関数 $f(x)$ の符号によって決定される．ここで，$f(x) = 0$ で表される判別境界を満たす x を閾値 θ とする．

$$y = \begin{cases} 1 & (f(x) \geq 0) \\ 0 & (f(x) < 0) \end{cases}$$

$$f(x) = wx + b$$

$$\theta = -\frac{b}{w}$$

$$x = Binoculars(s)$$

$y = 1$ は入力文章がモデルによって作成されたこと，$y = 0$ は入力文章が人間によって作成されたことを意味する．$f(x)$ は Binoculars スコアと正解ラベルからなる学習データの集合に基づき，教師あり機械学習によって誤分類を最小化するように推定された判別関数である．w は重み，b はバイアスを表し，特徴量が一次元である空間においては判別境界 $f(x) = 0$ を満たす x が θ となる．$Binoculars(s)$ は入力文章 s について算出された Binoculars スコアを示す．

提案手法の処理の流れは図 4.1 に示す．この処理では，まず文章をトークン列として入力すると，Binoculars スコアが算出され，次にそのスコアを入力特徴量として判別境界が推定され，最後にその判別境界を用いて LLM 生成文か人間生成文か

を示す二値ラベルが出力される．Binoculars の詳しい算出方法は 2.1.4 節，閾値の決め方やその目的は 4.2 節で述べる．

5 章では，Binoculars スコアを一次元の特徴量として用い，機械学習により判別境界を推定するという手法が，LLM 生成文と人間生成文の判別に有効であるという仮説について，その妥当性を検証するための実験を行う．本手法は，前提条件として Binoculars スコアにおける LLM 生成文と人間生成文の間に平均値や分布の差が存在することを想定しており，その差を一次元の特徴量として機械学習モデルが捉えることができた場合には本手法の有効性があると考えられる．一方で，Binoculars スコアにおける LLM 生成文と人間生成文の間に統計的に有意な差が無い場合や，機械学習モデルが Binoculars スコアの分布の不安定性にうまく対処できない場合には，本手法において十分な判別性能が得られない可能性があると考えられる．

4.1 Binoculars の構造とモデルの役割

Binoculars スコアの算出には，performer モデルと observer モデルという役割の二つのモデルを使用する．Binoculars スコアはモデルにとっての文章の意外さを表す log-perplexity という値を，モデル間の差を表す log-cross-perplexity という値で割ることで求められるが， log-perplexity を求めるのに performer モデルを，log-cross-perplexity を求めるのに performer モデルと observer モデルを用いる．Binoculars スコアの詳しい計算式は 2.1.4 節で述べた通りである．

本研究では，performer モデルと observer モデルに何を選択するかを重要視している．この二つのモデルの出力がスコアに直結するためである．performer モデルが高性能である場合，LLM 生成文と人間生成文の違いを正確に捉え，分離したスコアを示す可能性が考えられる．一方，performer モデルが低性能である場合，LLM 生成文と人間生成文の違いを捉えきれず，判別境界を設定しづらいスコア分布を示す可能性が考えられる．observer モデルが高性能な場合，performer モデルの出力に対して，performer モデル特有の出力であるか一般的な出力であるかを評価しやすく，log-cross-perplexity が log-perplexity を適切に正規化でき，結果として文章判別に有用なスコアを示しやすい可能性がある．observer モデルが低性能

な場合，performer モデルの出力を理解しきれず，文章判別において意味のない差を拾ってしまう可能性が考えられる．5 章では，performer モデルと observer モデルそれぞれのスコアに対する影響を調べるため，スコア算出に用いるモデルの組み合わせを変えて出力を比較する実験を行う．

4.2 閾値推定

本研究では，文章判別における判別境界の推定には SVM とロジスティック回帰の 2 種類の機械学習手法を用いる．提案手法の頑健性を検証するため，複数の手法を用いて文章判別を行い，結果を比較する．

SVM とロジスティック回帰は両方とも教師あり機械学習モデルである．先述したように，本研究では Binoculars スコアは算出に使用するモデルの組み合わせによって分布が変動してしまう不安定性を持つと仮定している．そのため，直感的，あるいは平均値など統計的な閾値を定めて高精度の文章判別を行うことは困難であると考える．一方，機械学習を用いた場合，正解ラベルを用いて判別境界が最適化され，直接的に誤分類を減らすことに繋がる．

また，SVM とロジスティック回帰は，固定的な閾値に基づく判別とは異なり，学習データに基づいて判別境界を最適化する柔軟な判別モデルである．本研究では扱う特徴量が一次元に留まるが，提案手法の二値分類器は将来的に新たな特徴量を追加することのできる拡張性を持つ．Binoculars スコアの分布の不安定性や提案手法の拡張性を考慮し，本研究では，多次元の特徴量から学習を行い柔軟な文章判別を行うことのできる SVM やロジスティック回帰を採用した．

機械学習モデルによる判別境界の学習と評価のため，提案手法の二値分類器に入力するデータをトレーニングデータとテストデータに分けて行う．それぞれのデータは，判定する対象の文章の Binoculars スコアである．トレーニングデータを使った学習では，入力されたスコアをもとに，文章が LLM 生成文か人間生成文かの二値ラベルを出力し，その出力と正解ラベルがなるべく一致するように，判別境界を学習していく．テストデータを使った評価では，学習の結果求められた閾値を使ってテストデータの判別を行い，文章が LLM 生成文か人間生成文かの二値ラベルを出力し，正解ラベルと比較して分類性能を評価する．

第 5 章　評価実験

本実験の総合的な目的は，Binoculars スコアを一次の特徴量として機械学習モデルに判別境界を学習させた二値分類器が，LLM 生成文と人間生成文を高精度に判別できるかどうかを検証することである．そのために，四つの段階に分けた実験を行う．実験 1〜3 は二値分類器に大きく影響する Binoculars スコアの性質を確認するための探索的実験であり，実験 4 は実験 1〜3 の結果を踏まえて実際に二値分類器を作成し判別性能の評価を行う主実験として位置付ける．

実験 1 では，Binoculars スコアの算出にかかわるモデルの選択方法による Binoculars スコアへの影響を観察する．実験 2 では実験 1 で分かったことが様々なモデルを使った場合にも言えるのか汎用性の検証を行う．実験 3 では判別対象の変化に伴う Binoculars スコアの観察を行う．実験 4 では Binoculars スコアを一次の特徴量として機械学習モデルで判別境界を推定した時の判別性能の検証を行う．

各段階での目的や手順は実験 1〜4 の各節にて示す．

5.1　使用モデル

本実験で用いたモデルの一覧は以下の通りである．3 章でも取り上げたように，先行研究より，日本語の文章に対しては日本語モデルを使用することが有効であると判明している．そのため日本語モデルを中心に，パラメータ数やリリース日の異なる計 15 個のモデルを選択した．これらのモデルはデータセットの作成や Binoculars スコアの算出に用いられている．

- rinna/japanese-gpt2-xsmall*
- rinna/japanese-gpt2-small†
- rinna/japanese-gpt2-medium‡
- rinna/japanese-gpt-1b§

*https://huggingface.co/rinna/japanese-gpt2-xsmall
†https://huggingface.co/rinna/japanese-gpt2-small
‡https://huggingface.co/rinna/japanese-gpt2-medium
§https://huggingface.co/rinna/japanese-gpt-1b

- rinna/japanese-gpt-neox-3.6b[¶]
- rinna/bilingual-gpt-neox-4b[‖]
- rinna/youri-7b[**]
- meta-llama/Llama-2-7b-hf[††]
- deepseek-ai/deepseek-llm-7b-chat[‡‡]
- deepseek-ai/deepseek-llm-67b-chat[§§]
- Qwen/Qwen3-Embedding-8B[¶¶]
- Qwen/Qwen3-14B[***]
- google/gemma-3-12b-it[†††]
- google/gemma-3-27b-it[‡‡‡]
- openai/gpt-oss-20b[§§§]

5.2 使用データ

今回実験に使用するデータセットは 2 種類ある．

一つ目は日本語フェイクニュースデータセット[¶¶¶]である．内容はウィキニュース日本語版の記事と GPT-2 日本語版のモデルが作成した記事の混合で，ラベルによって完全に LLM が作成したもの，部分的に LLM が作成したもの，完全に人間が作成したものの 3 種類に仕分けられている．本研究では完全に LLM が作成した記事 4671 件と，完全に人間が作成した記事 3685 件を使用するものとする．以降このデータセットをデータセット 1 と呼称する．

[¶] https://huggingface.co/rinna/japanese-gpt-neox-3.6b-instruction-ppo
[‖] https://huggingface.co/rinna/bilingual-gpt-neox-4b
[**] https://huggingface.co/rinna/youri-7b
[††] https://huggingface.co/meta-llama/Llama-2-7b-hf
[‡‡] https://huggingface.co/deepseek-ai/deepseek-llm-7b-chat
[§§] https://huggingface.co/deepseek-ai/deepseek-llm-67b-chat
[¶¶] https://huggingface.co/Qwen/Qwen3-Embedding-8B
[***] https://huggingface.co/Qwen/Qwen3-14B
[†††] https://huggingface.co/google/gemma-3-12b-it
[‡‡‡] https://huggingface.co/google/gemma-3-27b-it
[§§§] https://huggingface.co/openai/gpt-oss-20b
[¶¶¶] https://github.com/tanreinama/Japanese-Fakenews-Dataset

二つ目は半自作のデータセットである．人間が作成した記事は，先述の日本語フェイクニュースデータセットからそのまま用いる．LLM が作成した記事については，モデルに「数百文字程度で○○についてのニュース記事を 1 つ作ってください」のようなプロンプトを与えて作成させた．○○の部分にはスポーツ，天気，社会，国の四つの分野から以下のような話題をあてはめた．

- スポーツ
  - 野球，サッカー，バスケットボール，水泳，卓球，陸上，スケート，バレーボール，ソフトボール，剣道
- 天気
  - 春の天気，夏の天気，秋の天気，冬の天気，梅雨の天気，台風，地震，竜巻，津波
- 社会
  - 宗教，政治，戦争，テロ，難民，選挙，芸能，病気，環境問題，産業
- 国
  - 日本，中国，ロシア，アメリカ，韓国，タイ，ベトナム，フィリピン，イギリス，イタリア

作成に使用したモデルは Deepseek-r1:70b，gemma3:27b，llama4，gpt-oss:20b，qwen3:8b である．この五つのモデルにそれぞれ上記の話題を全て与えている．この二つ目のデータセットでは，人間生成文は同じく 4671 件，LLM 生成文は 2000 件となっている．以降このデータセットをデータセット 2 と呼称する．

5.3 実験 1

実験 1 では，Binoculars スコアにおける，スコアの算出にかかわるモデルの影響を調べる．そのために，Binoculars スコアの算出に用いる二つのモデルの選択方法や組み合わせ方を変化させ，算出された LLM 生成文と人間生成文の Binoculars スコアの乖離具合の変動を観察する．

Binoculars スコアは log-perplexity を log-cross-perplexity で割ることで求められる．詳しい算出手順は 2.1.4 に示す． log-perplexity は performer モデルの出

力から算出される値であり，performer モデルから見た入力文章の意外さを表す．log-cross-perplexity は performer モデルと observer モデル両方の出力から算出される値であり，performer モデルと observer モデルの挙動の差を表す．両者とも performer モデルの影響を強く受けているため，これらの比をとることで，意外さという値に含まれるモデル依存性を弱めることになるのである．

performer モデルの影響を含む二つの値の比を取るという構造から，Binoculars スコアの算出において, 最終的には performer モデルによる影響よりも observer モデルによる影響が強く残る可能性があると考える．そのため，本実験では observer モデルの変化によるスコアへの影響について特に注目しながら実験を行う．

5.3.1 実験 1 の手順

実験 1 においては，使用するモデルを japanese-gpt2-xsmall，japanese-gpt2-medium, japanese-gpt-1b という 3 種類に限定し，文章作成, performer, observer の三つの役割に組み合わせを変えながらあてはめる．なお，モデルの数や判別対象の文章のサンプル数も限られていることから，実験 1 から得られる結果はあくまで直感的に LLM 生成文と人間生成文のスコアの乖離を観察することに焦点を置いたものであり，統計的な意味を見出すことは今回の目的にない点に留意する．

手順は以下の通りである．

1. 表 5.1 に従いモデルを選択する．
2. 文章作成モデルを用いて 3 件の文章を作成し，同じ話題について人間が 3 件の文章を作成する．
3. 作成された文章に対する Binoculars スコアを算出する．
4. LLM 生成文のスコアと人間生成文のスコアでそれぞれ平均と分散を記録する．

文章作成，performer，observer の三つの役割を担うモデルが，すべて等しい場合，一つだけ異なる場合，すべて異なる場合といった観点で条件を設定し，表 5.1 に示す計 6 通りの組み合わせ方について実験を行う．

表 5.1: 実験 1 モデル組み合わせ表

	文章作成	performer	observer
1	japanese-gpt2-medium	japanese-gpt2-medium	japanese-gpt2-medium
2	japanese-gpt2-medium	japanese-gpt-1b	japanese-gpt-1b
3	japanese-gpt2-medium	japanese-gpt2-medium	japanese-gpt-1b
4	japanese-gpt2-medium	japanese-gpt-1b	japanese-gpt2-medium
5	japanese-gpt2-xsmall	japanese-gpt-1b	japanese-gpt2-medium
6	japanese-gpt2-medium	japanese-gpt-1b	japanese-gpt2-xsmall

5.3.2 実験 1 の結果と考察

実験 1 の結果からは，Binoculars スコアの算出において，performer モデルよりも observer モデルにパラメータ数の大きいモデルを配置した方が，LLM 生成文と人間生成文のスコアの乖離が大きくなるということが示された．

実験 1 の結果を以下の表 5.2 に示す．LLM 生成文と人間生成文それぞれの平均，分散の列に示す数字は，Binoculars スコアを意味している．

表 5.2: 実験 1 結果表

	人間平均	LLM 平均	人間分散	LLM 分散
1	2.5325	1.8666	0.0360	0.0117
2	2.3654	1.5138	0.0604	0.0579
3	0.7979	0.9655	0.0014	0.0108
4	0.7147	0.7630	0.0018	0.0288
5	0.7147	0.5587	0.0018	0.0865
6	0.7192	0.6754	0.0028	0.0199

表 5.2 において，平均差が大きい順にケースを並び替えると $6 < 4 < 5 < 3 < 1 < 2$ となる．ここで，実験結果の傾向を分析するための観点として，実験 1 に使用した三つのモデルのリリース日及びパラメータ数を以下に示す．

- gpt2-xsmall
  - リリース日：2021 年 8 月 25 日
  - パラメータ数：43.7M
- gpt2-medium
  - リリース日：2021 年 4 月 7 日 (アップデート：2021 年 8 月 25 日)
  - パラメータ数：0.4B
- gpt-1b
  - リリース日：2022 年 1 月 26 日
  - パラメータ数：1B

平均差が大きかった 2 や 1 のケースでは，performer モデルと observer モデルが同一である．平均差が次に大きい 3 のケースでは，パラメータ数がより大きくリリース日がより新しい gpt-1b が observer モデルに配置されている．その次に平均差が大きい 5 のケースは 3 のケースに対して performer モデルと observer モデルが逆転している．平均差が小さい 4 や 6 のケースでは，パラメータ数がより大きくリリース日がより新しい gpt-1b が performer モデルに配置されている．

以上の結果から，パラメータ数の大きさやリリース日の新しさが平均差の大きさに関係している可能性が考えられる．しかし，本実験においてはパラメータの大きさの順とリリース日の新しさの順が同一であったため，明確にどちらがどの程度実験結果に影響しているか切り分けて分析することは困難である．また，先述したようにこの実験ではサンプル数が少なく，加えて分散差も大きいことから，単純な平均差順で傾向を測りきることは不可能であることに注意する．

また，Binoculars スコアは算出途中に performer モデルと observer モデルの挙動の差を表す値を計算するため，performer モデルと observer モデルが同一である場合は Binoculars スコアの利点を生かすことになっていないと考えた．よって，その他のケースの傾向を鑑み，performer モデルよりも observer モデルの方にパラメータ数が大きいまたはリリース日が新しいモデルを置いた場合に，LLM 生成文と人間生成文のスコアの乖離が最も大きくなるという仮説を立てる．ただし，先述したようにパラメータ数の大きさとリリース日の新しさをそれぞれ独立に分析できないため，本研究に限り，パラメータ数の大きさ及びリリース日の新しさをモデルの特性を表す複合的な性質として考える．次の実験 2 では，この仮説に基づいて

performer モデルと observer モデルの組み合わせを決定していく．

さらに，4 と 5 のケースの比較から，文章作成モデルによっても Binoculars スコアが変化することが示唆される．判定対象の変化による Binoculars スコアの変化の統計的な観測は，後の実験 3 にて行うこととする．

5.4 実験 2

実験 2 では，実験 1 の結果を受け，performer モデルよりも observer モデルの方がパラメータ数が大きくリリース日が新しい場合に LLM 生成文と人間生成文の Binoculars スコアの乖離が大きいという仮説が，どのモデルにおいても成り立つかを検証する．そのため，observer モデルの方がパラメータ数が大きくリリース日が新しくなるように，様々なリリース時期のモデルを使用して Binoculars スコアの算出を行い，LLM 生成文と人間生成文の Binoculars スコアに統計的な有意差があると言えるかどうかを見る．

5.4.1 実験 2 の手順

実験 2 の手順は以下の通りである．LLM 生成文と人間生成文の Binoculars スコアに統計的な有意差があるかどうかを検証するため，実験 2 では t 検定を用いる．t 検定は 2.2.1 節で述べたように，二つのデータ群の間に有意差が存在すると言えるのかを確認する手法である．実験 1 の結果から，LLM 生成文と人間生成文の Binoculars スコアは等分散性が言えないことが分かっているため，非等分散の場合でも使える Welch の t 検定を用いる．なお，有意水準について，本実験は複数の統計的検定を行う多重比較にあたる．多重比較では一度の検定に比べ，偶然有意差を持ってしまう可能性が高くなるため，有意水準を下げる必要がある．本実験では多重検定のうち Bonferroni 法を採用し，有意水準 5% を実験回数である 71 で割り，0.07% とする．

1. 表 5.3 に従いモデルを選択する．
2. データセット 1 内の各文章について Binoculars スコアの算出を行う．
3. LLM 生成文と人間生成文でそれぞれ Binoculars スコアの平均と分散を算

出する．

4. 得られた平均，分散，そして文章の個数を使って，t 検定を行う．有意差の有無を記録する．

実験 2 では，performer モデルよりも observer モデルの方がパラメータ数が大きい場合に LLM 生成文と人間生成文の Binoculars スコアの乖離が大きいということについてなるべく網羅的に検証するため，5.3 に示す 71 通りの組み合わせについて実験を行う．

表 5.3: 実験 2 モデル組み合わせ表

	performer	observer
1	japanese-gpt2-xsmall	japanese-gpt2-small
2	japanese-gpt2-xsmall	japanese-gpt2-medium
3	japanese-gpt2-xsmall	japanese-gpt-1b
4	japanese-gpt2-xsmall	japanese-gpt-neox-3.6b
5	japanese-gpt2-xsmall	bilingual-gpt-neox-4b
6	japanese-gpt2-xsmall	youri-7b
7	japanese-gpt2-xsmall	Llama-2-7b-hf
8	japanese-gpt2-xsmall	gpt-oss-20b
9	japanese-gpt2-xsmall	deepseek-llm-67b-chat
10	japanese-gpt2-xsmall	Qwen3-Embedding-8B
11	japanese-gpt2-xsmall	gemma-3-12b-it
12	japanese-gpt2-xsmall	gemma-3-27b-it
13	japanese-gpt2-small	japanese-gpt2-medium
14	japanese-gpt2-small	japanese-gpt-1b
15	japanese-gpt2-small	japanese-gpt-neox-3.6b
16	japanese-gpt2-small	bilingual-gpt-neox-4b
17	japanese-gpt2-small	youri-7b
18	japanese-gpt2-small	Llama-2-7b-hf
19	japanese-gpt2-small	deepseek-llm-67b-chat

20	japanese-gpt2-small	Qwen3-Embedding-8B
21	japanese-gpt2-small	gpt-oss-20b
22	japanese-gpt2-small	gemma-3-12b-it
23	japanese-gpt2-small	gemma-3-27b-it
24	japanese-gpt2-medium	japanese-gpt-1b
25	japanese-gpt2-medium	japanese-gpt-neox-3.6b
26	japanese-gpt2-medium	bilingual-gpt-neox-4b
27	japanese-gpt2-medium	youri-7b
28	japanese-gpt2-medium	Llama-2-7b-hf
29	japanese-gpt2-medium	deepseek-llm-67b-chat
30	japanese-gpt2-medium	Qwen3-Embedding-8B
31	japanese-gpt2-medium	gpt-oss-20b
32	japanese-gpt2-medium	gemma-3-12b-it
33	japanese-gpt2-medium	gemma-3-27b-it
34	japanese-gpt-1b	japanese-gpt-neox-3.6b
35	japanese-gpt-1b	bilingual-gpt-neox-4b
36	japanese-gpt-1b	youri-7b
37	japanese-gpt-1b	Llama-2-7b-hf
38	japanese-gpt-1b	deepseek-llm-67b-chat
39	japanese-gpt-1b	Qwen3-Embedding-8B
40	japanese-gpt-1b	gpt-oss-20b
41	japanese-gpt-1b	gemma-3-12b-it
42	japanese-gpt-1b	gemma-3-27b-it
43	japanese-gpt-neox-3.6b	bilingual-gpt-neox-4b
44	japanese-gpt-neox-3.6b	youri-7b
45	japanese-gpt-neox-3.6b	Llama-2-7b-hf
46	japanese-gpt-neox-3.6b	deepseek-llm-67b-chat
47	japanese-gpt-neox-3.6b	gpt-oss-20b
48	japanese-gpt-neox-3.6b	gemma-3-12b-it

49	bilingual-gpt-neox-4b	youri-7b
50	bilingual-gpt-neox-4b	Llama-2-7b-hf
51	bilingual-gpt-neox-4b	deepseek-llm-67b-chat
52	bilingual-gpt-neox-4b	Qwen3-Embedding-8B
53	bilingual-gpt-neox-4b	gpt-oss-20b
54	bilingual-gpt-neox-4b	gemma-3-12b-it
55	bilingual-gpt-neox-4b	gemma-3-27b-it
56	youri-7b	Llama-2-7b-hf
57	youri-7b	deepseek-llm-67b-chat
58	youri-7b	Qwen3-Embedding-8B
59	youri-7b	gpt-oss-20b
60	youri-7b	gemma-3-12b-it
61	youri-7b	gemma-3-27b-it
62	Llama-2-7b-hf	deepseek-llm-67b-chat
63	Llama-2-7b-hf	Qwen3-Embedding-8B
64	Llama-2-7b-hf	gemma-3-12b-it
65	Llama-2-7b-hf	gemma-3-27b-it
66	gpt-oss-20b	deepseek-llm-67b-chat
67	gpt-oss-20b	Qwen3-Embedding-8B
68	gpt-oss-20b	gemma-3-12b-it
69	gpt-oss-20b	gemma-3-27b-it
70	deepseek-llm-67b-chat	Qwen3-Embedding-8B
71	gemma-3-12b-it	deepseek-llm-67b-chat

5.4.2 実験 2 の結果と考察

実験 2 の結果からは，performer モデルよりも observer モデルの方がパラメータ数が大きくリリース日が新しい場合，多くのモデルの組み合わせにおいて，LLM

生成文と人間生成文の Binoculars スコアに統計的な有意差が生じるということが示された.

実験 2 の結果を以下の表 5.4 に示す．なお，平均や分散の列はそれぞれ Binoculars スコアについての値を表している.

表 5.4: 実験 2 結果表

	人間平均	LLM 平均	人間分散	LLM 分散	p 値	有意差
1	2.0985	2.1858	0.0585	0.1744	1.0211e-32	有
2	2.0096	2.1320	0.0463	0.1580	9.9932e-71	有
3	1.0335	1.0651	0.0085	0.0278	6.485e-28	有
4	0.7990	0.8152	0.0060	0.0165	1.08346e-12	有
5	1.0298	1.0692	0.0036	0.0253	9.751e-54	有
6	1.0260	1.0515	0.0024	0.0170	1.993e-34	有
7	0.7496	0.7704	0.0253	0.0404	1.3194e-07	有
8	1.1338	1.1957	0.0040	0.0377	6.7345e-90	有
9	1.1499	1.2280	0.0047	0.0325	2.698059e-154	有
10	0.7826	0.7841	0.0006	0.0012	2.062068e-02	無
11	0.7756	0.7969	0.0153	0.0248	4.690865e-12	有
12	0.7314	0.7681	0.0068	0.0189	6.229075e-51	有
13	2.4989	2.4022	0.1485	0.2467	1.6694e-23	有
14	1.0530	1.0715	0.0149	0.0314	1.7741e-08	有
15	0.8025	0.8047	0.0034	0.0096	0.202391	無
16	1.0358	1.0564	0.0032	0.0184	7.7225e-21	有
17	1.0243	1.0322	0.0036	0.0134	5.6711e-05	有
18	0.7262	0.7328	0.0343	0.0447	1.2879e-01	無
19	1.0961	1.1352	0.0060	0.0245	1.480221e-49	有
20	0.8041	0.7901	0.0007	0.0013	8.432488e-91	有
21	1.0765	1.1061	0.0050	0.0298	3.1284e-26	有
22	0.7503	0.7598	0.0154	0.0223	1.504212e-03	無

23	0.7562	0.7772	0.0072	0.0153	5.216822e-20	有
24	1.0844	1.0901	0.0124	0.0291	0.065782	無
25	0.8317	0.8272	0.0054	0.0178	0.0501392	無
26	1.0710	1.0755	0.0030	0.0053	0.00127108	無
27	1.0706	1.0799	0.0031	0.0277	3.54501e-04	有
28	0.7669	0.7701	0.0313	0.0465	0.45628	無
29	1.1670	1.2075	0.0047	0.0257	1.256058e-53	有
30	0.8128	0.7925	0.0007	0.0013	6.470616e-184	有
31	1.1413	1.1714	0.0029	0.0408	2.6726e-22	有
32	0.7934	0.7991	0.0161	0.0248	6.695738e-02	無
33	0.7714	0.7829	0.0074	0.0196	3.956010e-06	有
34	0.7873	0.7599	0.0327	0.0375	2.8251e-11	有
35	0.8586	0.8503	0.0521	0.0737	0.129185	無
36	0.7321	0.7027	0.0319	0.0288	2.50232e-14	有
37	0.5441	0.5154	0.0272	0.0256	1.3945e-15	有
38	0.7815	0.7675	0.0671	0.0878	0.02139345	無
39	0.9083	0.8571	0.0007	0.0014	0.000000e+00	有
40	0.6616	0.6435	0.0502	0.0755	9.1567e-04	無
41	0.5088	0.4961	0.0353	0.0444	3.657741e-03	無
42	0.8185	0.8119	0.0028	0.0128	4.217811e-04	有
43	1.0913	1.0402	0.0056	0.0312	1.3178e-69	有
44	1.1865	1.1517	0.0074	0.0460	7.772e-24	有
45	1.1606	1.1313	0.0077	0.0461	2.9365e-17	有
46	1.1559	1.1274	0.0059	0.0266	7.836119e-26	有
47	1.1395	1.1089	0.0041	0.0235	1.2468e-34	有
48	0.6819	0.6539	0.0108	0.0150	2.198755e-29	有
49	1.0141	0.9639	0.0014	0.0071	1.155e-150	有
50	0.9048	0.8759	0.0020	0.0040	2.3385e-127	有
51	1.0814	1.0745	0.0031	0.0249	5.496689e-03	無

52	0.7863	0.7359	0.0011	0.0019	0.000000e+00	有
53	1.0892	1.0818	0.0018	0.0273	3.2903e-03	無
54	0.6985	0.7000	0.0131	0.0153	0.5660251	無
55	0.7194	0.6872	0.0017	0.0076	2.308072e-107	有
56	6.4910	4.3956	14.1974	7.8988	1.0656e-164	有
57	1.1651	1.1609	0.0075	0.0197	0.0930697	無
58	0.9088	0.8373	0.0008	0.0015	0.000000e+00	有
59	1.0517	1.0338	0.0028	0.0122	2.5964e-22	有
60	0.9882	0.9747	0.0220	0.0298	1.231940e-04	有
61	1.0217	1.0070	0.0081	0.0185	3.281563e-09	有
62	1.1052	1.1333	0.0072	0.0195	1.247358e-29	有
63	0.8610	0.8113	0.0008	0.0015	0.000000e+00	有
64	0.9576	0.9784	0.0175	0.0261	1.038389e-10	有
65	0.9695	0.9824	0.0076	0.0181	1.227555e-07	有
66	0.7723	0.7786	0.0021	0.0049	7.516230e-07	無
67	0.6321	0.6232	0.0011	0.0012	0.000000e+00	有
68	0.6199	0.6351	0.0051	0.0075	1.780590e-18	有
69	0.6635	0.6793	0.0020	0.0043	1.328376e-38	有
70	0.8650	0.8228	0.0007	0.0015	0.000000e+00	有
71	0.8681	0.7661	0.0011	0.0028	0.000000e+00	有

71 個のケースの中で，有意差があると言えるケースは 53 個，有意差があると言えないケースは 18 個であった．つまり，本実験下においては有意差があったケースは全体の 74.6% であった．LLM 生成文と人間生成文の Binoculars スコアで，平均値の差は 0.0015〜2.0954 の範囲を取る．平均差だけを見ればそこまで大きな差は無いように感じられるが，分散の差が大きいために，有意差を取るケースが多いことが考えられる．さらに，有意差があるケースにおいては p 値も 0.000000e+00 といった非常に小さい値を取っている場合が複数見受けられる．

以上の結果から，performer モデルより observer モデルのパラメータ数が大き

くリリース日が新しい場合には，多くのモデルの組み合わせにおいて，LLM 生成文と人間生成文の Binoculars スコアの間に統計的な有意差が安定して存在する可能性が高いといえる．次の実験 3 では，実験 1 の考察から立てた仮説が正しいものとして performer モデルと observer モデルの組み合わせを固定し，判別対象の変化によるスコアへの影響を観察していく．

5.5 実験 3

実験 3 では，判別対象の変化による Binoculars スコアへの影響を観察する．そのため，performer モデルと observer モデルの組み合わせは固定し，判定対象となる文章のうち LLM 生成文が比較的古いモデルによって生成されているデータセット 1 と，比較的新しいモデルによって生成されたデータセット 2 を用いてそれぞれ Binoculars スコアの算出を行う．なお，performer モデルと observer モデルの組み合わせについては，実験 1 や実験 2 から得られた結果と考察より，observer モデルとしてよりパラメータ数が大きくリリース日が新しいモデルを配置する．

直感的に，古いモデルよりも新しいモデルの方がより人間に近い文章を書くことができるはずであり，新しいモデルによる LLM 生成文と人間生成文を人の目で区別するのは困難である．Binoculars スコアを一次の特徴量として機械学習モデルに判別境界を学習させる二値分類器において，古いモデルによる LLM 生成文と人間生成文は見分けられるが新しいモデルによる LLM 生成文と人間生成文は見分けられない，という事象が発生する場合，これは二値分類器の明確な弱点となる．逆に，新しいモデルによる LLM 生成文と人間生成文を Binoculars スコアによって見分けられるのであれば，Binoculars スコアを特徴量とする二値分類器も新しいモデルによる LLM 生成文に対応できる力を持つということになるため，二値分類器の強みになるといえる．実験 3 では二値分類器の性能に大きく影響する Binoculars スコアについて，判定対象の LLM 生成文の新旧による影響を観察する．

また，実験 3 においてはデータセットに焦点を当てているため，実験後実際にスコア付けされた文章を観察し，Binoculars スコアと文章の特徴の相関がどこにあるかを探る．

5.5.1 実験 3 の手順

手順は以下の通りである．データセット 1 とデータセット 2 それぞれについて同じ実験を行う．

1. 表 5.5 に従いモデルを選択する．
2. データセット内の各文章についてそれぞれ Binoculars スコアの算出を行う．
3. LLM 生成文と人間生成文それぞれの Binoculars スコアの平均と分散を算出する．
4. 得られた平均，分散，そして文章の個数を使って，t 検定を行う．有意差の有無を記録する．

なお，t 検定については実験 2 と同じく Bonferroni 法を採用し，有意水準は 5% ÷ 3=1.67% とする．

表 5.5: 実験 3 モデル組み合わせ表

	performer	observer
1	Llama-2-7b-hf	gemma-3-27b-it
2	gpt-oss-20b	deepseek-llm-67b-chat
3	gpt-oss-20b	gemma-3-27b-it

5.5.2 実験 3 の結果と考察

実験 3 の結果からは，比較的新しいモデルによる LLM 生成文の方が，比較的古いモデルによる LLM 生成文よりも，人間生成文との Binoculars スコアの乖離が大きいことが示された．つまり，Binoculars スコアを一次の特徴量として使用する二値分類器にも，新しいモデルによる LLM 生成文と人間生成文を見分ける能力が一定程度期待できると言える．また，Binoculars スコアは文の長さや同じ単語の繰り返しといった文章の特徴に応じて上下する傾向が示され，文章の文法的，意味的な特徴には反応できない可能性が示唆された．

実験 3 の結果を以下の表 5.6 に示す．簡略のため，p 値と有意差の列における (1) をデータセット 1，(2) をデータセット 2 として表記する．

表 5.6: 実験 3 結果表

	p 値 (1)	p 値 (2)	有意差 (1)	有意差 (2)
1	1.227555e-07	2.678176e-06	有	有
2	7.516230e-07	0.000000e+00	有	有
3	1.328376e-38	0.000000e+00	有	有

どのケースも有意差があり，LLM 生成文と人間生成文の Binoculars スコアの間に統計的に十分な乖離があると言える．表の 2 と 3 のケースにおいてデータセット 2 を用いた方の p 値が小さいことから，少なくとも本実験においては，新しいモデルによる LLM 生成文の方が概ね人間生成文と乖離した Binoculars スコアを示すと言える．

Binoculars スコアにおいて，比較的新しいモデルによる LLM 生成文と人間生成文の間に大きな差があるということは，Binoculars スコアを一次の特徴量とした二値分類器は，比較的古いモデルによる LLM 生成文よりもむしろ比較的新しいモデルによる LLM 生成文と人間生成文を見分けることが得意である可能性が高い．人の目では見分けることが難しい文章を高精度に判別することができるならば，この二値分類器には大きな利点があると言える．

実験 3 の結果は，LLM 生成文を作るモデルの新旧によってもたらされたが，実際に文章の特徴はどのように変化しているのだろうか．ここで，Binoculars スコアと文章の特徴の相関を探すため，実際にデータセット 1 に含まれる文章とその Binoculars スコアを観察してみる．

例えば，以下の文章は 2 のケースの場合に，Binoculars スコアが最も高かった文章である．

日経新聞や産経新聞によると、4区では前原誠司大学が区間賞を獲得。青山学院大学も、青山学院大学駅伝研究会代表の中で、中西圭亮大学の1年先輩である中西学大学のエース・吉本賢一選手が区間記録（6時間2分43秒、前年大会1位のタイムで80秒更新）をマークした。4区で2位、5区で3位、5区で2位、5区で4位、5区で3位、5区で3位、5区で2位、5区で1位、5区で3位、5区で3位、5区で3位、5区で1位、5区で3位、5区で3位、5区で1位、5区で3位、5区で4位、5区で2位、5区で3位、5区で1位、5区で4位、5区で3位、5区で1位、5区で2位、5区で3位、5区で3位、5区で3位、5区で1位、5区で4位、5区で1位、5区で2位、5区で3位、5区で4位、5区で1位、5区で3位、5区で2位、5区で3位、5区で3位、5区で3位、5区で3位、5区で1位、5区で1位、5区で2位、5区で3位、5区で1位、5区で3位、5区で4位、5区で1位、5区で3位、5区で9位、5区で1位、5区で2位、5区で3位、5区で1位、5区で3位、5区で2位、5区で4位、5区で2位、5区で3位、5区で2位、5区で1位、5区で3位、5区で2位、6区で3位、5区で4位、4区で1位、5区で4位、5区で2位、5区で3位。

例文 1

上の例文 1 は LLM 生成文である．明らかに不自然に同じような文字列が繰り返されており，直感的にも LLM 生成文であると判断しやすい文章だと言える．

また，同じ 2 のケースの場合で，Binoculars スコアが最も低い文章も見てみる．

東京新聞（以下東京）によると、白い祭器の後ろで白い装束の老人が「赤い衣（いしろ）」の頭を垂れる「あばれみこし」を踊った。その踊りを見ながら、老人は「この世の神様、仏様たちがお力をお貸しくださっていただいてこちらまで来てくださった。神様の御御……」…
「「「「「「「…………」」」」「ふふっ、ふふふっ。ふふふ。ふふふふふふふふふふふふふふふふっ」

例文 2

上の例文 2 も LLM 生成文である．これも先程と同じく同じ文字の繰り返しを含む．

他の文章もいくつか見たところ，このような繰り返しの特徴を持つ文章は Binoculars スコアが極端に高いか低いかのどちらかになりやすいようだ．さらに他の文章を観察したところ，外国語や記号が多く用いられている文章やごく短い文章も Binoculars スコアが極端な値を取る場合が多いことが分かった．

なぜこのような繰り返しを含む文章が生成されてしまうのかを考えてみる．第 2.1.1 節で述べたように，モデルは次に来る確率が高い単語を連ねて文章を作る．「4 区で 2 位」「5 区で 3 位」のような列記があると，次も同様の要素が続く確率が高いとモデルが判断してしまうことが考えられる．そのような判断が続いてしまい，結果異常な長さの繰り返しが発生したのではないかと推測する．「御御御」も，一度同じ文字が続いたのだから次も同じ文字が来るだろうという判断からくる長い繰り返しだと考えられる．

さらに，Binoculars スコアが中程度の文章も例示する．下の例文 3 が LLM 生成文，例文 4 が人間生成文で，スコアは同値であった．

東亜日報日本語版によると、ITmediaによれば、2006年12月28日の記事数は2万5175件と、2006年11月の記事数の2倍以上に達している。また、インターネット上でのウェブサイト訪問数は1920年代から2000年の18年間におよそ2倍以下に減っている。記事数においてはITmediaによれば、日本の2006年12月を含めてインターネットを通じて、1日あたり1699記事となり、210億件を超えているという分析がある。記事の多さについて、ITmediaによれば、「2011年に入ってから、2012年には15倍にまで増える」とされている。記事数の多さについては、Girlfriend誌やKADOKAWA誌などの雑誌でも注目されている。IGNのような誌誌名による記事の多さについては、「記事数を増やせば、記事数が変わってくるのかな」とも言われている。インターネットが普及する中で、インターネット上での動画の閲覧回数や投稿やコメント、Facebookを含むSNSなどのアプリケーション・チャットツールの利用の多いこと、スマートフォンやタブレットなどでのコミュニケーションの少なさ、ユーザーとの直接的なコミュニケーションの少なさなどが、スマートフォンとの差別化として懸念されていることは言うまでもない。ただし、Apple iPad、iPhone、Apple Twitter Life & RecruitsといったIGNITE（AOL）と提携しているApple Booksの利用者、すなわち、Apple Best Display Art's booksの利用者にとっては、インターネットやその他のコミュニケーションツールで、より広くコンテンツを提供できることが期待される。

例文 3

台湾南部の高雄市の市街地で、現地時間7月31日深夜から翌未明[注釈 1]にかけて大規模な爆発が複数回起こった。爆発は、3平方キロメートルほどの広範囲で起こり、爆発により道路の陥没や車の横転、路上が炎上するといった被害がでている。また現地の消防当局によると、25人(内、4人は警官・消防隊員)が死亡し271人が負傷した。台湾政府は、住民の救出や避難のため軍などを投入している。爆発に至った経緯として現地メディアが報道したところによると、現場近くに所在している石油化学工場[注釈 2]から接続されてあるパイプラインより気体の漏れが発生、気体が可燃性の化学物質であったため引火・爆発に至ったとみて当局が調査中であると、現地で報道がなされている。現地メディアは目撃者の話として、「爆発が起きる前にガスの匂いがしていて、消防が原因を調べていた。負傷者の中には、消防隊員も含まれている。」と伝えている。一部情報では、爆発前に化学物質の気体の漏れが検出されていたとも伝えられている。現場から1キロメートル程離れたマンションに住む日本人男性は、「午前0時頃に大きく建物が揺れ、地震かと思った」と述べている。爆発が起きた場所は、高雄駅南東で日本人も居住している地域。また、同市には日本人は約1300人が在留しており、日系企業170社も進出している。日本・台湾間の窓口機関である交流協会は、日本人が被害に遭っていないか確認中としているが、現在のところ負傷したとの情報はないとしている。

例文 4

例文 3 を見ると，主語の消失や文のねじれ等おかしいところはあるものの，スコアが極端な文章に比べれば違和感は少ないといえる．例文 3 と例文 4 では，文章の長さや，文末に伝聞の言葉が多いという特徴が共通しているのが分かる．

スコアが同じでも，LLM 生成文はニュース記事として信頼できるとは言い難い．なぜなら文法や意味の部分での文章の違和感が払拭できていないためである．このことから Binoculars スコアでは，あくまで文章の長さや記号，数字の乱用の有無等に基づいて判断を行っているのであり，文法や意味的な面から見る文章の整合性は加味していないと考えられる．

データセット 2 も同様に観察したところ，例文 1 や例文 2 のような，極端な文章の短さや同じ文字の繰り返しといった特徴を持つ文章はデータセット 1 よりも少ないように見受けられた．技術の発展により，一目で LLM 生成文と分かるような異様な特徴が排除されたのだと考えることができる．

新しいモデルはより自然な単語分布に従って文章を生成するため，人間が頻繁に使う単語を優先して使い，あまり使われない単語は抑制する傾向にある．そのため，

古いモデルによる文章に見られた異様な繰り返しが，新しいモデルによる文章では過度に抑制され，人間との乖離が相対的に大きくなるという結果につながったと考えられる．

データセット内の文章の観察から，比較的新しいモデルによる LLM 生成文は一見人間生成文と区別しにくいように見えても，Binoculars スコアにおいては十分な差が出ていることが分かった．実験 4 では，この十分な差が機械学習モデルを用いた二値分類器においてもきちんと機能するかどうか，実際に二値分類器を作成してその判別性能を検証していく．

5.6 実験 4

実験 4 では，Binoculars スコアを一次の特徴量として機械学習モデルに判別境界を学習させ，LLM 生成文と人間生成文を高精度に判別できるかを検証する．そのために，機械学習モデルとして SVM とロジスティック回帰の 2 種類を用い，判別性能を比較することで二値分類器の性能を総合的に評価する．また，評価指標には Accuracy と AUC の二つを用いる．

Accuracy は，2.4.1 に示したように，ある閾値をもってデータをどのくらい正しく判別できたかを見るための指標である．本実験では，機械学習の結果推定された判別境界が実際にデータを正しく判別できるかを評価するために用いる．

AUC は，2.4.2 に示したように，閾値に依存せず評価をする方法である．本実験では，機械学習の結果推定された判別境界を定める決定化関数に，Binoculars スコアを入力した結果の出力スコアが，LLM 生成文と人間生成文を判別しやすいように並べられているかを評価するために用いる．

実験 1 や 2 から，performer モデルよりも observer モデルの方がパラメータ数が大きい場合に LLM 生成文と人間生成文の Binoculars スコアの乖離が大きい傾向があることが分かった．実験 3 から，比較的新しいモデルによる LLM 生成文の方が人間生成文との Binoculars スコアの乖離が大きいことが分かった．以上の結果を踏まえて，実験 4 では performer モデルより observer モデルの方によりパラメータ数が大きいモデルを配置し，かつデータセット 1 ではなくデータセット 2 を用いて実験を行うものとする．

5.6.1 実験 4 の手順

実験の手順は以下の通りである．

1. 表 5.7 に従いモデルを選択する．
2. データセット 2 内の各文章について Binoculars スコアの算出を行う．
3. SVM とロジスティック回帰それぞれに，トレーニングデータとしてデータセット全体の 8 割の Binoculars スコアと正解ラベルを渡し，境界推定のための学習を行わせる．
4. SVM とロジスティック回帰それぞれに，テストデータとしてデータセット全体の 2 割の Binoculars スコアを渡し，判別を行わせる．
5. テストデータの判別結果と正解ラベルを照らし合わせ，Accuracy と AUC をそれぞれ算出して比較する．

なお，データのサンプル数は先述したように LLM 生成文が 2000 件と人間生成文が 4671 件で計 6671 件である．トレーニングデータはその 8 割にあたる 4585 件，テストデータは 2 割にあたる 1137 件で行っている．

表 5.7: 実験 4 モデル組み合わせ表

	performer	observer
1	deepseek-llm-7b-chat	deepseek-llm-67b-chat
2	Qwen3-Embedding-8B	Qwen3-14B
3	gemma-3-12b-it	gemma-3-27b-it

5.6.2 実験 4 の結果と考察

実験 4 の結果からは，performer モデルよりも observer モデルの方がパラメータ数が大きくリリース日が新しいモデルであり，判別対象の LLM 生成文が比較的新しいモデルによって作成されている場合，Binoculars スコアを一次の特徴量として機械学習に判別境界を学習させた二値分類器は最高で Accuracy0.81 の判別性能を持つことが示された．

実験 4 の結果を以下の表 5.8 に示す．表中の Accuracy と AUC の列について，簡略のため (S) を SVM，(ロ) をロジスティック回帰として表記する．

表 5.8: 実験 4 結果表

	Accuracy(S)	Accuracy(ロ)	AUC(S)	AUC(ロ)
1	0.69	0.69	0.59	0.59
2	0.79	0.79	0.80	0.80
3	0.81	0.80	0.81	0.81

1，2，3 のすべてのケースにおいて，判別性能はランダム分類の 0.50 を上回る結果となっており，最高で Accuracy 及び AUC が 0.81 を示している．また，使用しているモデルのパラメータ数が大きくリリース日が新しいほど Accuracy と AUC の値が大きい．さらに，各ケース各分類手法での Accuracy と AUC の値は似通っている．

以上の結果から，パラメータ数が大きくリリース日が新しいモデルを使用したほうが結果的により判別性能の高い二値分類器ができることが示唆される．Accuracy と AUC の値が似通っていることから，Binoculars スコアそのものの限界性能を評価できていると考えられ，また Accuracy の方が若干数値が高いケースもあるためこの分類器は Accuracy での評価と相性が良い可能性がある．

5.7 実験総考察

本実験では，Binoculars スコアを一次の特徴量として機械学習モデルに判別境界を学習させた二値分類器が，LLM 生成文と人間生成文を高精度に判別できるかどうかを段階的に検証した．実験の結果から判明した点を以下に示す．

- 提案手法の有効性
  - performer モデルよりも observer モデルとしてよりパラメータ数が大きくリリース日が新しいモデルを配置した場合，LLM 生成文と人間生成文の Binoculars スコアについて統計的な有意差が見られた．
  - 比較的古いモデルによる LLM 生成文と人間生成文より，むしろ比較的

新しいモデルによる LLM 生成文と人間生成文の方が Binoculars スコアについて統計的な差が大きいことが明らかになった.
  - 上記の二つの条件をそろえた上で，二値分類器の性能として最高で Accuracy0.81，AUC0.81 が示された.
- 提案手法の限界点及び課題
  - performer モデルと observer モデルが同一であった場合，LLM 生成文と人間生成文の Binoculars スコアにおける差が大きい可能性があったが，Binoculars スコアの構造上適さないという判断のもと同一のモデルの配置を検証対象外とした.
  - Binoculars スコアの算出に用いるモデルのパラメータ数やリリース日がスコアに影響することが示唆されたが，どちらがどの程度スコアに影響しているのか特定するのは実験設計上不可能であった.
  - Binoculars スコアについて，LLM 生成文は人間生成文より大きな分散の値を示し，分布が重なるため，一次の特徴量のみでは現状より高い判別性能を示すのは難しいと考えられる.
  - Binoculars スコアは文章の長さや同じ文字の繰り返しといった特徴に応じて上下する傾向があり，文法の間違いや意味的齟齬については検知できないことが示唆されたため，二値分類器においても文法や意味の点で文章を判別することは難しいと考えられる.

実験 1 及び実験 2 では，Binoculars スコアの算出に用いるモデルの組み合わせ方について検討し，LLM 生成文と人間生成文のスコアの乖離が大きくなる条件を明らかにした．実験 3 では判定対象の文章が Binoculars スコアにどのように反映するかを観察し，二値分類器がどのような特徴の文章を判別できてどのような特徴の文章を判別できない可能性があるかを検討した．実験 4 では実験 1〜3 で判明したことを頼りに二値分類器を構成し評価した結果，一定の判別性能が確認できた.

本研究ではニュース記事について取り扱っているため，人間生成文には基本的に文法の明らかな間違いや事実と異なる記述はないと捉えている．そこで，提案手法の二値分類器の拡張性を活かし，文法の間違いや文章の意味的齟齬を検知できる特徴量を二値分類器に追加することで，判別性能の向上が期待できる.

第 6 章　おわりに

6.1　まとめ

本研究では, LLM 生成文と人間生成文の判別に用いられる指標である Binoculars スコアについて，性質の確認及び Binoculars スコアを一次の特徴量として機械学習モデルで判別境界を学習する二値分類器の判別性能の評価を行った.

実験の結果，Binoculars スコアは算出に用いるモデルの組み合わせ方や判別対象の変化によって分布が変動する値であることが示され，固定の閾値による判別には適さない指標であることが示唆された．また，Binoculars スコアを一次の特徴量として機械学習モデルに判別境界を学習させた二値分類器は，最高で Accuracy0.81 という判別精度を示し，提案手法は日本語ニュース記事において LLM 生成文と人間生成文を判別する上で一定の有用性を持つことが明らかになった.

先行研究においては Binoculars スコアを用いることで 90 %以上の LLM 生成文を検出したと報告されている．しかし，先行研究では英語の文章を判定対象としているのに対し，本研究では日本語の文章，特にニュース記事の領域を判定対象としている．さらに，先行研究では使用モデルの変化による分布の変動に考慮していないが，本研究では分布の不安定性に対処するために機械学習モデルを用いて判別境界を推定している．これらの先行研究との取り組み方の違いから，先行研究と本研究の成果において単純な判別性能の比較は成り立たない.

本研究の成果は，Binoculars スコアの性質を調べ，その不安定性に着目したことで，使用したモデルの影響を抑え安定した判別性能を得られる可能性が示された点である．Binoculars スコアの算出に用いるモデルの組み合わせ方に依存した分布の不安定性や，文法や意味的齟齬を捉えられない性質，特徴量が一次のみであることを踏まえると，日本語ニュース記事という領域における文章判別において，Accuracy0.81 という判別精度は現状到達できる限界点であると考えられる．本研究で明らかになった Binoculars スコアの性質や提案手法の有用性を活用し，今後さらに多様な領域において LLM 生成文と人間生成文を判別する手法が現れることが期待される.

6.2 今後の展望

実験での観察結果より，Binoculars スコアには文法や意味的齟齬を捉えられない性質がある．この性質が本研究の提案手法におけるボトルネックになっていると考えられる．そこで，本手法の二値分類器には Binoculars スコア以外の特徴量も追加できるという拡張性を利用して，現状よりも高精度な二値分類器の実現を検討する．

本研究で扱っているニュース記事では，人間生成文については文法が正確であり意味的齟齬も少ないと考えられる．そのため，文法や意味的齟齬といった特徴を捉えられる特徴量を追加することで，現状よりも高精度に LLM 生成文と人間生成文を判別できることが期待される．例えば，単語の意味をベクトル空間で捉え，近い単語同士はベクトル空間内でも近い位置に配置される Word2Vec や，類似の考え方である Doc2Vec を活用することで，文章の意味に焦点を当てた指標を作れるのではないかと考える．また，文中の主語と述語の関係を検知する係り受け解析を使って，文のねじれを検出するための指標を作れるのではないかと考える．さらに，日本の都道府県は 47 個あるなどといった一般常識を多数記録した辞書を作ることで，常識から逸脱している内容の文章を検出するための指標を作れるのではないかと考える．

これらの指標を本手法の二値分類器の特徴量空間に追加することによって，より多角的な視点で LLM 生成文と人間生成文を見分けられるようになると考える．ただし，先述したようにニュース記事という領域では文法や意味的齟齬という観点が LLM 生成文と人間生成文を見分ける鍵になりそうだが，その他の領域あるいは全般的な文章の判別の際にどのような特徴量を用いるのが有効かは不明である．

また，本研究の実験結果から，比較的新しいモデルによる LLM 生成文の方が比較的古いモデルによる LLM 生成文よりも人間生成文との Binoculars スコアにおける乖離が大きいことが判明している．この乖離の理由として，仮説ではあるが，新しいモデルほど人間が頻繁に使う単語のみを選択する傾向があり，極度に平坦な文章を生成することによって，Binoculars スコア上ではかえって人間生成文からかけ離れてしまうことが考えられる．直感的には最近のモデルによる LLM 生成文と人間生成文を見分けることが難しくても，本手法を利用すればスコアの乖離を観測できるため，今後さらに進化したモデルが現れても対応できる可能性がある．

謝辞

卒業研究にあたり，研究室の皆様には非常にお世話になりました．鈴木先生にはどのように研究を進めたら良いかのアドバイスを逐一いただき，とても感謝しています．最後までご指導いただきありがとうございました．

事務補佐員の駒沢さん，研究室で顔を合わせると明るい笑顔で挨拶をしてくださってありがとうございました．駒沢さんからいただいた元気が，自身の研究のモチベーションにつなげられていたと思います．

鈴木研究室の先輩方には，口数の少ない自分にも気さくに接してくださり，本当に感謝しています．いつでも何でも聞いて良いという空気を普段から見せてくださっていたおかげで，相談がしやすかったです．また，先輩方同士の仲の良さも研究室全体の空気の良さを作っていたと思います．そして実際に質問をすると，必ず複数人で集まって一緒に考えて意見をくださるので，いつも大変参考になる答えを貰うことができていました．

鈴木研究室の同輩の皆さんにも，長いことお世話になりました．約 1 年半という期間，同じ立場で学習や勉強に取り組む仲間がいて良かったです．これからの進路はそれぞれ違いますが，各々が選んだ道で頑張っていけることを願っています．

鈴木研究室の後輩の皆さんも，短い期間でしたがお世話になりました．同輩や先輩と共に卒業研究を頑張って欲しいなと思います．

家族には，生活面で 4 年間支えていただいたことを感謝しています．この数年の間で大変なことがたくさん起きましたが，それでも衣食住の心配をせずに勉強や研究に集中することができたのは両親のおかげです．就職してお金や人生経験を貯めて，いつか何かの形で恩返しができれば良いなと考えています．

家や大学以外でも，たくさんの人にお世話になりました．勉強や研究に直接は関わらなくても，新たな経験や価値観，そして精神面の支えをいただけたことは本当にありがたく思っています．結果としてそれらは私自身の成長につながるだけでなく，研究の中での少しの考え方やものの見方にも変化を生んでくれました．

最後になりますが，大学生活および卒業研究を支えてくださった皆様のおかげでこの論文を完成させることができました．本当にありがとうございました．

参考文献

[1] Alberto J. Gutiérrez-Megías, L. Alfonso Ureña-López, and Eugenio Martínez-Cámara. The influence of the perplexity score in the detection of machine-generated texts. *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*, pp. 80–85, 2024.

[2] 佐藤修. Llm の普及と利用リスクの一考察. 日本情報経営学会第 86 回全国大会予稿集【特定自由論題：技術が媒介する越境経済・社会】, pp. 27–30, 2023.

[3] 鈴木雅弘, 坂地泰紀, 和泉潔. 異なる単語分割システムによる日本語事前学習言語モデルの性能評価. 言語処理学会 第 29 回年次大会 発表論文集, pp. 894–898, 2023.

[4] 萩原正人, 小川泰弘, 外山勝彦. perplexity を用いた類義語獲得の自動評価. 言語処理学会第 12 回年次大会 (NLP2006), pp. 767–770, 2006.

[5] Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*, 2024.

[6] 大林準. 統計検定手法の選び方　— 基本編 —. 天理医学紀要, Vol. 25, No. 1, pp. 60–65, 2022.

[7] CORINNA CORTES and VLADIMIR VAPNIK. Support-vector networks. *Machine Learning*, pp. 273–297, 1995.

[8] 中澤秀夫. ロジスティック回帰 (logistic regressions). 日本医科大学医学会雑誌 (Nihon Ika Daigaku Igakkai Zasshi), pp. 186–191, 2014.

[9] Tom M. Mitchell. Machine learning. 1997.

[10] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, Vol. 27, No. 8, pp. 861–874, 2006.

[11] 岩田理央, 綱川隆司, 西田昌史. 大規模言語モデルにより生成された日本語テキストの検出性能の検証. 2025 年度人工知能学会全国大会 (第 39 回), 2025. セッション ID: 4G1-GS-6-04.

発表リスト