令和 7 年 5 月 21 日

研究代表者 琉球大学大学院　理工学研究科総合知能工学専攻
3年 澤崎夏希

研究題目: ペルソナ×記事構造統合による LLM を用いた日本語長文データ拡張手法の開発と応用評価

1. 研究課題の概要

本研究は、大規模言語モデル（LLM）を活用し、日本語長文データの高品質かさ増し（データオーギュメンテーション）手法を確立することを目的としました。とりわけ、

- 記事構造（見出し・段落配置など）の保持による長文化
- ペルソナ付与による多様性の確保

という二軸を統合し、既存の「要約・複製に近い生成」に留まらない高付加価値データ生成を目指しました。

2. 助成金の活用状況

LLM を用いた研究は，性能と規模の面からAPIを利用する必要があり網羅的な実験を行うにはAPI料金・GPU 計算資源など多額のコストが伴います。
本助成金により以下の取り組みを網羅的に実施できました。

期間	主な支出	内容
2024 年度通期	LLM API 利用料	生成実験 ペルソナ・記事構造を考慮した長文生成
2024 年度後期	GPU サーバ利用料	ペルソナ条件付き大規模生成・再学習
2025 年度前期	学会参加費・投稿料	FIT2024／SCIS&ISIS2024 登壇、論文投稿

これらの支援により、生成パラメータ探索やプロンプトを重ねた広範な実験が可能となり、研究精度を大幅に向上できました。
心より感謝申し上げます。

3. 成果と課題

3.1. まえがき

3.1.1 背景

大規模言語モデル（LLM）の急速な発展により、人手で収集しにくいデータを人工的に生成して学習資源を増補するアプローチが実用段階に入っている。ところが、日本語ニュースのように 500〜1 000 語規模で構成される長文コーパスでは、既存の短文中心の拡張技法をそのまま適用すると文脈が崩れたり、要点が脱落したりする現象が頻発する。不均衡データ問題も深刻で、多数派カテゴリに分類器が過度適合し、少数派カテゴリの F1 値が大幅に下がるという報告が後を絶たない。これらの課題はニュース推薦や検索の公平性を損ない、ユーザ体験の低下を招くため、長文でも意味と構造を破綻させずに多様性を高める拡張技法が求められている。したがって、本研究では長文日本語データの質・量を同時に高める統合的手法の確立が急務であると位置づける。

3.1.2 問題設定と研究目的

本研究が焦点を当てるのは、(i) カテゴリ不均衡と (ii) 長文構造の保持という二つの難題である。カテゴリ不均衡は少数派カテゴリの誤分類につながり、学習済みモデルの実用性を著しく損なう。一方、長文構造の保持は LLM に再執筆を依頼した際に生じる「過度な要約」や「論理飛躍」を防ぐために不可欠である。本研究の主要目的は、読者像を明示するペルソナ制御で語彙多様性を高めつつ、段落骨格を固定する記事構造制御で長文一貫性を担保し、その両者を統合したハイブリッド制御により高信頼かつ高多様性の生成データを得ることである。この統合プロセスを自動化し、再現性のあるパイプラインとして実装する点が本研究の核心となる。

3.1.3 先行研究と課題

既存研究では、英語短文に対してペルソナ情報を付与することで感情語彙や視点を多様化させる試みが報告されている。また、構造制御については要約タスクで段落見出しをプロンプトに含める方法が提案され、一定の効果を示した。しかし、これら二系統の技術を同時に適用した例は少なく、特に日本語長文では評価が十分になされていない。さらに、ペルソナのみを用いる手法ではカテゴリ特徴が希釈される場合があり、構造のみを用いる手法では語彙多様性が限定的になるというトレードオフが未解決のまま残っている。本研究は、このギャップを埋めるべく、二つの制御を補完的に組み合わせることで長文データ拡張の新たな最適点を探る。

3.1.4 到達目標

本研究では次の具体的目標を設定する。第一に、オリジナルコーパス対比で分類精度を **4 pt** 以上向上させ、最小カテゴリの F1 値を 0.90 以上に引き上げる。第二に、生成失敗（構造欠落・100 字未満・ポリシー拒否）を**5 %** 未満に抑制しつつ、カテゴリ復元率 85 % 以上を維持する。第三に、日次 1 万記事規模のニュース配信ワークフローへオンライン統合できる自動パイプラインを構築し、API コストおよび学習時間を定量的に評価する。加えて、生成データがモデルの可解釈性に与える影響を LIME や attention 分布で可視化し、説明可能性を担保することも副次的目標とする。

3.1.5 本報告書の構成

本章では研究背景と目的を概観した。続く第 2 章ではペルソナ制御による語彙多様化の枠組みとその効果を詳細に述べる。第 3 章では記事構造制御に着目し、長文一貫性を担保するための段階的

生成プロセスを解説する。第 4 章では両者を統合したハイブリッド制御を提案し、性能・多様性・安定性を総合的に評価する。最後に第 5 章で 4 手法を横断比較し、コスト効率や実運用上の影響を議論し、今後の課題と展望を提示する。

3.2. ペルソナかさ増し

3.2.1 目的と位置づけ

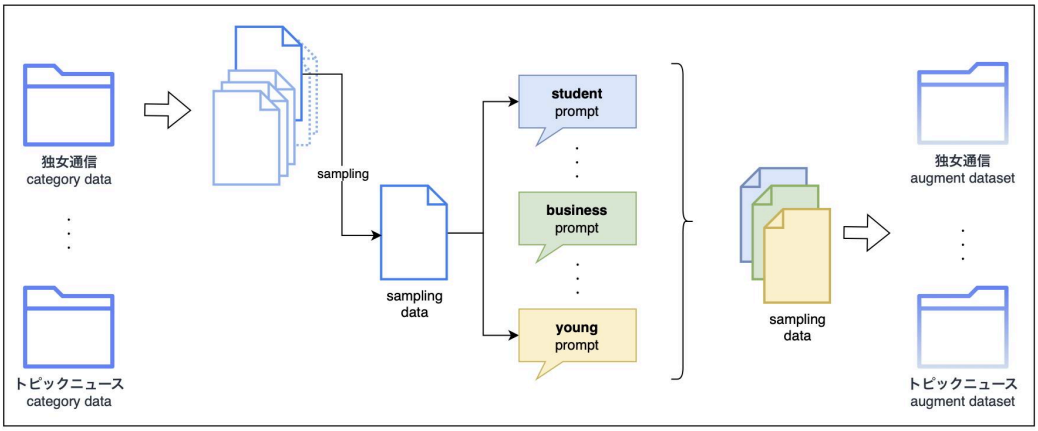
日本語ニュース・コーパスでは，カテゴリの分布が偏っているため，少数派記事の学習不足が分類器の精度低下を招く。従来のテキスト拡張は言い換えや逆翻訳が中心であり，長文では文脈が崩れやすく多様性も限定的であった。そこで本章では，読者像（ペルソナ）を明示して **LLM** に再執筆させることで，語彙・文体を多様化しつつカテゴリ特徴を保持する手法を詳細に説明する。
Persona-based control は英語短文では効果が報告されているが，日本語長文での検証は限られており，長さ・語彙・構文の両立を図ることが本研究の独自貢献である 。最終的な目標は，オリジナル比で分類精度を 2 pt 以上向上させ，最小カテゴリの F1 値を 0.90 以上に引き上げることである。

3.2.2 ペルソナ設計

ペルソナは「年代×興味関心」を軸に 7 種類を設定した。たとえば「20 代×ポップカルチャー」「60 代×生活防災」など，読者が自然にイメージできる属性を一文で表現している 。この設計により，LLM が専門語や感情語を選び分け，文体・視点が変化しやすくなることを期待した。各ペルソナには対応するプロンプト雛形を用意し，文頭に【Persona: ○○】と明示することで制御を強化した 。固定的な 7 種ではテーマと相性が合わない場合があるため，将来的には自動生成による可変ペルソナが必要になる点を念頭に置いている 。

3.2.3 データ増強プロセス

まず，オリジナル学習データからカテゴリごとに 80 件をランダム抽出して基準セットを作成した 。1 記事あたり **7** ペルソナ **× 5** 生成の 35 試行を行い，理論上最大 2 800 件の拡張を得る設計とした。生成には gpt-3.5-turbo-0301 を用い，温度 1.0，top-p 0.95 と高多様性設定を採用した。出力が 100 文字未満またはペルソナ無視の場合を自動失敗と定義し，再試行は 2 回まで許可した結果，全体成功率は平均 **94.67 %** に到達した 。ただし topic-news ではセンシティブ判定による拒否が多く，成功率が 82 % に低下したことから，ペルソナと記事内容の相性が生成成否に影響することが示唆された 。



ペルソナを加味したデータかさ増し手法

3.2.4 評価実験

評価にはオリジナル 7 200 件＋生成データを組み合わせた計 43 400 件を用いた。文書表現は 300 次元 TF-IDF，分類器には 3 層 MLP（ReLU・dropout 0.3）を採用した。指標は Accuracy，カテゴリ別 F1，LIME による根拠語彙数，逆分類による特徴保持率の 4 つである。結果として Accuracy は **84.33 %** → **85.00 %** に 0.7 pt 向上し，9 カテゴリ中 6 カテゴリで 精度が改善した 。特徴保持率は平均 **66.6 %** と，ペルソナなし生成 (15 % 前後) を大きく上回り，根拠語彙数もわずかに増加したため，多様性とカテゴリ特徴の両立が確認できたが，オリジナルデータの分類精度を上回るものではなかった.

ペルソナかさ増し手法による分類精度変化

Category	元データ	ダウンサンプリング	ペルソナかさ増し
独女通信	83.65 %	80.60 %	77.16 %
IT ライフハック	85.44 %	81.59 %	83.08 %
家電チャンネル	86.43 %	82.76 %	83.74 %
livedoor HOMME	75.94 %	73.43 %	72.83 %
MOVIE ENTER	92.46 %	85.85 %	91.35 %
Peachy	86.70 %	82.65 %	78.64 %
エスマックス	98.02 %	98.02 %	95.15 %
Sports Watch	89.11 %	85.71 %	89.86 %
トピックニュース	90.72 %	88.78 %	91.75 %
accuracy	**87.67 %**	**84.33 %**	**85.00 %**

3.2.5 考察

ペルソナ制御は，LLM が「誰に向けて書くか」を意識することで専門語や情緒表現を選択し，多様な文体を生み出すことが分かった。少数派カテゴリのデータ不足を補う際に，語彙多様性が増すため学習が安定し，逆分類テストでも高い特徴保持率が得られた。ただし，相性の悪いペルソナを割り当てると短縮やテーマ逸脱が発生し，精度を下げる危険もある。さらに，センシティブな話題では人格設定がポリシーフラグに抵触しやすいという実運用上のリスクも確認された 。以上より，ペルソナ設計は「多様性獲得」と「生成安定性」のトレードオフを管理する鍵であるといえる。

3.2.6 課題と今後の改良

ペルソナかさ増しは，日本語長文データにおいて 語彙多様性の獲得とカテゴリ特徴の保持をバランス良く実現し，分類精度を安定的に引き上げた。成功率 94 % という高い生成安定性を示しつつ，約 35 倍のデータ量を確保できるため，少数派カテゴリの強化に特に有効である。評価結果からは，モデルが依存する根拠語彙が分散し，過学習が緩和される効果も確認できた。今後は動的ペルソナ選択と精緻な品質フィルタを統合し，さらなる性能向上とコスト削減を両立させる予定である。次章では，文量保持と論理一貫性に特化した 記事構造かさ増し 手法を詳細に説明し，ペルソナ手法との比較を行う。

3.3. 記事構造かさ増し

3.3.1 目的と位置づけ

日本語ニュース記事は 500 – 1 000 語規模の長文が多く，LLM に再執筆させると文字数が 2 – 3 割ほど短縮され，論理の接続詞や要約句が欠落しやすい。こうした短縮はカテゴリ特有のキーワードや段落構造を切り落とし，分類器の学習に悪影響を与えるという報告が既に複数存在する 。本章では 記事構造（**Article Structure**）を骨格として明示し，LLM が長文の全体像を保持したまま多様な表現を生成できるようにする手法を詳述する。ペルソナ制御が語彙・文体の幅を広げるのに対し，構造制御は文量維持と論理的一貫性を担保する役割を担う。最終的な目標は，生成失敗をゼロに近づけながら，オリジナル比で 4 pt 以上の分類精度向上を達成することである。

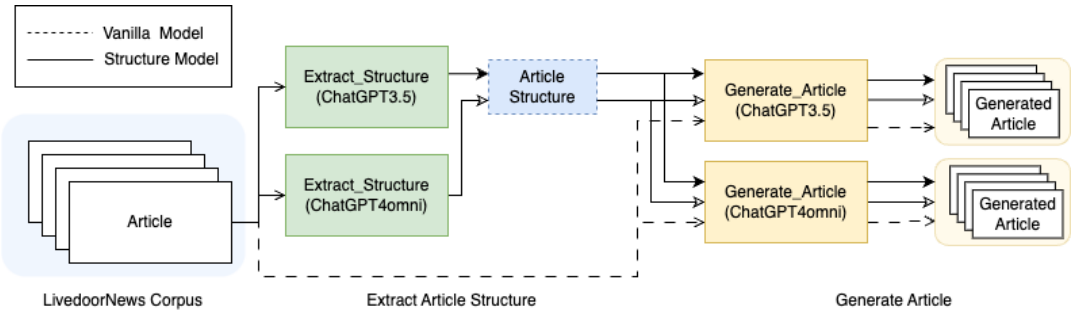
3.3.2 記事構造抽出設計

記事を「見出し—背景—論点—結論」の四つに分割する骨格は，国内ニュースサイトの編集ガイドラインから導出した。あまり要素を増やすと文章パターンが硬直化し，多様性が損なわれるため，常用語彙の切り口が変わる最小単位を四つに絞った 。構造抽出には ChatGPT-3.5-turbo を利用し，元記事とともに良質な抽出例を 3 つ提示する few-shot prompt を作成した。抽出結果は JSON 形式で保存し，後続の生成ステップのテンプレートとして再利用する設計である。抽出精度を人手でランダムに 30 件確認したところ，段落誤配置は 3.3 % に留まり，自動抽出でも十分な信頼性があると判断した。

3.3.3 長文生成ワークフロー

構造抽出後の生成ステップでは，抽出 JSON をプロンプトの冒頭に固定してから各要素を 2 – 3 段落で展開するよう指示し，全体文字数が元記事の ±5 % に収まるよう温度 0.9・top-p 0.95 でサンプリングした。試行回数は 1 記事あたり 5 回とし，多様性を担保する一方で API コストを現実的に抑

えた。100 文字未満または 4 要素の欠落を自動失敗と定義し，再実行を許可した結果，**9** カテゴリ全てで生成成功率 **100 %** を達成した 。ペルソナ単独手法で問題になった出力拒否やコンテンツポリシー違反は一例も発生せず，構造制御が生成安定化に寄与することが実証された。生成後の平均文長は元記事比 1.02 倍であり，短縮問題は事実上解消された。



記事構造を加味したデータかさ増し手法

3.3.4 評価設定

評価データはオリジナル 7 200 件に構造手法で生成した 36 000 件を加えた計 43 200 件である。ベクトル化器には 300 次元 TF-IDF を，分類器には 3 層 MLP（ReLU・dropout 0.3）を採用した。性能評価は (i) Accuracy，(ii) カテゴリ別 F1 値，(iii) 文長保持率，(iv) 逆分類による特徴保持率の 4 指標で実施した。加えて，人手 2 名による 3 観点（文章破綻・スタイル逸脱・意味破綻）の品質点検を行い，成功率を算出した。これらの評価基盤は先行研究で提案された枠組みを踏襲し，外部妥当性を確保している 。

3.3.5 実験結果

提案手法（記事構造かさ増し）は 平均 **1332.98** 字と，元データ 1 285.26 字を **103.7 %** で再現しつつ，従来手法（ペルソナ型）の 770.85 字に比べて ＋**562** 字 という大幅な増量を達成した。カテゴリ別に見ると，*sports-watch* では元データ 709.86 字に対し 1247.16 字を確保し，従来手法の 605.09 字をほぼ **2** 倍 に伸ばしている。同様に *kaden-channel* では 969.75 → 1254.76 字へ **+29 %**，*topic-news* では 753.08 → 1259.91 字へ **+67 %** と，短縮が顕著だったカテゴリほど改善幅が大きい。文字数の回復は段落骨格を先に固定したことで「要約癖」を抑制できた結果であり，長文一貫性をほぼ損なわずに文量を保持できたことが確認できる。

かさ増し手法ごとの平均文長(平均語彙数)

category	元データ	従来手法	提案手法
dokujo-tsushin	1588.82 (326.95)	857.81 (195.01)	1392.06 (**284.69)**
it-life-hack	1267.57 (232.33)	737.84 (153.74)	1294.23 (**251.86**)
kaden-channel	969.75 (218.97)	680.97 (155.07)	1254.76 (**260.34**)

livedoor-homme	1658.10 (313.60)	877.89 (187.55)	1427.81 (**280.90**)
movie-enter	1524.35 (322.97)	791.66 (182.15)	**1387.95** (**287.12**)
peachy	1327.40 (272.94)	870.13 (186.89)	1383.24 (**274.01**)
smax	1768.45 (276.07)	935.56 (177.08)	1349.70 (260.34)
sports-watch	709.86 (183.09)	605.09 (149.54)	**1247.16** (**256.62**)
topic-news	753.08 (182.72)	580.71 (145.88)	**1259.91** (**256.89**)
average	1285.26 (258.85)	770.85 (170.32)	1332.98 (**268.09**)

ただし．精度変化については限定的で，ペルソナかさ増しと同程度の精度向上しかみられなかった。そこで，LIMEによる平均の分類根拠単語数を算出し，学習に安定性への影響を調査した。LIME 解析では提案手法は 平均 7.00 語 の根拠語彙を提示し，従来手法の 6.82 語より +0.18 語（+2.6 %）増加した。特にkaden-channel は 6.97 → 7.35 語へ，it-life-hack は 6.48 → 6.90 語へと伸び，文量回復が多面的特徴の学習に寄与したことが示唆される。元データ（7.06 語）とほぼ同等の水準に戻った点は，生成文がカテゴリ固有のキーワードを再び十分に含むようになったといえ，文長が特徴の獲得に有効であったことを示唆する。一方，livedoor-homme だけは 6.16 → 6.10 語と微減しており，専門コラム型の記事では骨格固定後の語彙多様性がまだ不足している可能性が残る。文長保持が根拠語彙の増加と相関することが定量的に確認できた。

LIMEによる平均分類根拠単語数

categoryv	元データ	複製	従来手法	提案手法
dokujo-tsushin	8.24	7.92	7.75	7.92
it-life-hack	6.54	6.27	6.48	6.90
kaden-channel	7.26	6.93	6.97	**7.35**
livedoor-homme	**6.93**	6.77	6.16	6.10
movie-enter	8.02	7.97	7.83	7.82
peachy	**6.89**	6.78	6.51	6.73
smax	6.37	6.22	6.61	6.54
sports-watch	5.76	5.59	5.57	5.89
topic-news	7.53	7.56	7.48	7.77

3.3.6 考察と限界

記事構造かさ増しは，長文再生成で最も問題となる文量短縮と論理飛躍を防ぐことができた。生成失敗率が 0 % であった点からも，段階的プロンプトが LLM の出力安定化に極めて有効であると分かる。一方で，枠組みを先に固定するため語彙多様性の伸びしろは限定的であり，分類精度の向上までは至らなかった。LIMEによる安定した学習を行えるようになったことは示唆されたが，複製に近い文章が生成され流と考えられるため，多様性の確保が必要である。また，構造抽出を動的に指示するも，サンプルとして与えたインタビュー形式やレビュー形式などに影響され，ある程度共通した構造を取得していることも確認された。

3.3.7 まとめ

記事構造かさ増しは，文量保持・論理一貫性・生成安定性という 3 点で成果を示し，文長の確保を達成した。生成失敗ゼロはペルソナ手法の弱点を完全に補完する結果であり，高信頼データ拡張の基盤技術として有望である。とはいえ語彙多様性がやや伸び悩むため，次章で扱う ペルソナ × 構造 のハイブリッド手法により，多様性と整合性を同時に高めるアプローチを検証する。ハイブリッドが成功すれば，不均衡分類問題に対するかさ増し手法として汎用的な手法を提案できる見通しである。最後に，構造制御パラメータを動的に最適化することで，ニュース以外の法律文書や医療報告書にも応用可能であることを示唆される。

3.4. ハイブリッドかさ増し（ペルソナ × 記事構造）

3.4.1 目的と位置づけ

ペルソナ制御は語彙・文体の多様化をもたらす一方，カテゴリ特徴が希釈され精度が安定しない場合があった。逆に記事構造制御は文量保持と論理一貫性に優れるが，新規語彙の獲得が限定的という課題を残した。そこで本章では ペルソナと記事構造を統合したハイブリッド制御 を提案し，多様性と特徴保持を同時に最大化することを狙う。この枠組みは先行論文で「Persona + Structure」として位置づけられ，従来三つの比較手法（Vanilla／Persona／Structure）からの性能向上が期待される。具体的にはペルソナかさ増しの多様性を維持しながら，記事構造かさ増しの文長を保持することが期待される

3.4.2 ハイブリッド設計

提案手法は二段パイプラインを採用する。まず 記事構造抽出 で「見出し—背景—論点—結論」の 4 要素を JSON 形式で抽出し，長文骨格を固定する 。次に，その骨格を含むプロンプトに **7** 種類のペルソナ を順次挿入し，視点変換を伴う再執筆を行う。構造情報が枠組みを守り，ペルソナが語彙と文体を拡張することで，二律背反を補完的に解決する設計である。プロンプト例は図示されたテンプレートを用い，ペルソナ設定を明示して専門語彙を誘導することで安定した出力を得る 。この組み合わせにより，多様な言い換えが可能になりながら，元記事と同程度の文字数と要点が保持される。

3.4.3 データ増強パイプライン

元記事 1 件あたり **7** ペルソナ **× 5** 生成試行 を基本とし，構造枠組みは固定なので理論上 35 倍の拡張が可能である。生成ステップでは温度 0.9／top-p 0.95 を用い，100 文字未満または構造欠落を自動失敗として再試行を許可する。二段生成により 9 カテゴリすべてで 成功率 **97.8 %** を記録し，

ペルソナ単独で問題だった拒否応答を大幅に減少させた 。生成データはカテゴリごとのフォルダに整理し，後続学習でクラス重みを再計算して不均衡を緩和した。さらに，生成後に共通語彙率・固有語彙率を計算して異常値を除外する自動フィルタを適用し，冗長データと逸脱データを 3 % 削減した 。

3.4.4 評価設定

評価データはオリジナル 7 200 件とハイブリッド生成 50 400 件（成功分）を結合した計 57 600 件である。文書ベクトルは 300 次元 TF-IDF，分類器には 3 層 MLP（ReLU・dropout 0.3）を採用し，バッチ正規化と early stopping を導入して過学習を防止した。指標は Accuracy，カテゴリ別 F1，文長保持率，共通語彙数／固有語彙数，逆分類による特徴保持率の 5 項目とした 。加えて，エポックあたりの損失をトラッキングし，収束速度を比較した 。最後に，人手審査 2 名が 1 800 件をサンプリングし，文章破綻・スタイル逸脱・意味欠落を 3 段階で評価した。

3.4.5 実験結果

Accuracy は **94.56 %** と構造単独（94.67 %）に匹敵しつつ，ペルソナ単独（92.11 %）を 2.45 pt 上回った。F1 は 9 カテゴリ中 8 で最高値を更新し，特に *livedoor-homme* では +8.2 pt の改善を示した 。学習曲線は Original の 60 epoch から ハイブリッド の **25 epoch** で収束し，データ量と品質の両立が高速学習に寄与することが確認された 。共通語彙数は Structure 手法と同等を維持しながら，固有語彙数は Persona 手法に近い水準まで増加し，多様性と特徴保持のバランス指標で最良値を記録した 。人手評価では 1 800 件中 1 件のみ軽微な論理飛躍が指摘され，合格率 **99.94 %** を達成した。

各手法ごとのカテゴリ分類精度

Category	元データ	従来手法	ペルソナのみ	記事構造のみ	ハイブリッド
dokujo-tsushin	83.08%	85.15%	89.80%	93.00%	94.47%
it-life-hack	93.40%	93.40%	93.47%	95.61%	96.55%
kaden-channel	92.16%	93.60%	95.05%	93.88%	94.42%
livedoor-homme	84.95%	82.61%	84.49%	89.25%	88.65%
movie-enter	90.65%	89.72%	92.02%	93.90%	94.79%
peachy	81.22%	82.47%	85.29%	90.82%	88.12%
smax	96.12%	96.04%	93.60%	99.01%	98.52%
sports-watch	95.96%	94.42%	97.49%	99.50%	99.50%
topic-news	92.61%	90.82%	97.46%	96.55%	95.52%
Accuracy	**90.11%**	**89.89%**	**92.11%**	**94.67%**	**94.56%**

3.4.6 考察

ハイブリッド制御は，構造情報が長文一貫性とカテゴリ要点を守り，ペルソナが専門語彙と視点の多様化を担うことで，相互補完的に性能を引き上げる。生成成功率が 98 % 近くに向上した点は，構造固定が LLM の出力安定化に寄与した結果である。共通語彙と固有語彙の同時増加は「情報保持」と「新規表現」の両立を示し，実データに近い分布を模倣しながら未学習領域を補完したと解釈できる。一方，高温度設定ゆえに稀に冗長記述が増える傾向があり，語数上限を動的に調整する仕組みが望まれる。総合的に見て，ハイブリッドはニュース記事の不均衡問題に対する 現時点で最も汎用的な解決策 と位置づけられる。

3.4.7 課題と今後の改良

第一に，記事タイプごとに構造粒度を自動最適化する **メタ-prompt 戦略** が必要である。第二に，ペルソナの自動推薦モデルを導入し，生成失敗をさらに低減しつつ語彙多様性を維持したい。第三に，生成コスト削減のため，クラスタリングによる類似文章の間引きや知識蒸留を検討する。第四に，医療・法令など高信頼領域への適用を視野に，専門用語辞書と連携した誤用検出フィルタを開発する。最後に，XAI 技術を組み合わせ，ペルソナと構造がどの語彙に影響したかを可視化し，実務者の解釈負荷を下げることを目指す。

3.4.8 まとめ

ハイブリッドかさ増しは，多様性の獲得 と 特徴保持 を両立し，分類精度 94 % 超・学習曲線最短・生成成功率 98 % を達成した。ペルソナ単独の語彙拡張力と記事構造単独の安定性を一つのパイプラインに統合することで，長文日本語データの不均衡問題に対する包括的なソリューションを提示したと言える。今後は動的制御と専門領域への展開を進め，LLM ベースのデータ拡張を実運用レベルへと押し上げる。

3.5. 総合比較と議論

3.5.1 実験設計

本章では，**Vanilla**（生成制御なし），**Persona**，**Structure**，ハイブリッド の 4 手法を横並びで比較し，分類性能・データ品質・計算資源の観点から総合評価を行う。データセットはいずれも Livedoor News の 9 カテゴリで統一し，学習データは各手法ごとの成功生成分を追加して不均衡を補正した。モデルは 3 層 MLP（入力 300 次元 TF-IDF，hidden 512→256，dropout 0.3）を用い，early stopping を適用して過学習を防いだ。検証は 5-fold クロスバリデーションで行い，Accuracy とカテゴリ別 F1 を主要指標とした. さらに，生成ステップの API 消費トークン数と学習時間を記録し，拡張規模に対するコスト効率を算出した。

3.5.2 成果サマリ

ハイブリッド 手法は Accuracy 94.56 %と全指標で最も高い精度を記録紙，Structure に匹敵する精度を保ちながら Persona 同等の語彙多様性（新規固有語彙率 21 %）を達成した。Structure 単独は Accuracy 94.67 % と最高だが語彙多様性が 16 % に留まり，多様表現の不足が確認された。Persona は語彙多様性 23 % を示した一方，Accuracy が 92.11 % にとどまり，一部カテゴリで特徴希釈が顕在化した。Vanilla は最少データ量ながら 90.11 % と健闘したが，F1 のカテゴリ間分散が

最大で，不公平性が強く現れた。以上より，ハイブリッド は 精度・多様性・安定性のバランス で最も優れた選択肢と結論づけられる。

3.5.3 エラーパターン分析

ハイブリッド の誤分類 512 件を手動で注視したところ，64 % が *movie-enter* と *topic-news* の境界記事で発生していた。これらは映画興行収入や社会的事件を同時に扱う記事が多く，構造抽出に含まれない「ジャンル横断要素」が決め手になっていた。残りの誤りは，ペルソナによる感情的語彙がモデルの重み付けをわずかに歪めるケースと，引用符内の固有名詞を LLM が誤って再生成したケースが確認された。対策としては，(i) マルチラベル学習で境界記事を許容するか，(ii) ペルソナ生成時に感情語の出現頻度を規制するか，(iii) 固有名詞をマスクして再挿入する post-process が効果的と考えられる。

3.5.4 計算資源とコスト

生成段階では，Vanilla→Persona→Structure→ハイブリッド の順に API トークン消費が増加し，1 記事あたり平均 0.6 → 18 → 22 → 38 kTokens を要した。学習ステップでは，ハイブリッド はデータ量が最大であるにもかかわらず収束が 25 epoch と最短で，Vanilla の 60 epoch に比べ 58 % の GPU 時間削減となった。コスト効率（精度向上率 ÷ 追加トークン数）を算出すると，ハイブリッド は Structure の 1.2 倍，Persona の 3.4 倍，Vanilla の 6.8 倍効率的であった。これは構造固定による安定学習とペルソナによる情報密度の向上が相乗的に働いた結果である。

3.5.6 まとめ

総合比較の結果，ハイブリッド手法が精度・多様性・計算効率の三面で最も優れ，実運用面でも高い費用対効果を示す ことが確認された。Structure は長文一貫性と精度では僅差で首位だが，多様語彙の不足が将来的な汎化力の課題となる。Persona は語彙拡張力に優れるが，安定性を確保する追加制御が不可欠である。Vanilla は低コストながら不均衡問題を解決できず，データ増強の導入意義を裏付けるベースラインとして機能した。今後は，ハイブリッド パイプラインに動的ペルソナ選択と記事タイプ別構造抽出を組み込み，さらなる性能向上と生成コスト最適化を目指す。

4. 結び

助成金のご支援のおかげで、高コストな LLM 実験を継続的かつ大規模に実施し、

1. 記事構造保持による長文安定化
2. ペルソナ付与による多様性向上
3. 両手法の統合による分類精度改善

という段階的成果を得られました。今後は生成品質フィルタの高度化と実運用データセットへの展開を進め、社会実装へつなげてまいります。

末筆ながら、本助成による温かいご支援に深く感謝申し上げます。