CE-004

ユーザ指定スケールに応じた日本近代文学テキストの自動レベル変換：RAG を活用した複数レベル指標への統合アプローチ

Automatic Level Transformation of Japanese Modern Literature Texts Based on User-Specified Proficiency Scales: A Retrieval-Augmented Generation (RAG) Multi-Scale Approach

甘利 実乃†
Mino Amari

1. はじめに

1.1 研究背景：指標の並立と真正教材の間の断絶

日本語教育のグローバル化に伴い、学習者の背景や目的はかつてないほど多様化している。この状況は、言語能力を記述・評価するための単一で絶対的な基準の存在を困難にし、学習者一人ひとりの個別具体的なニーズに応える言語教育パラダイムへの転換を強く要請するものである。

現状、日本語能力を測定・記述する主要な指標として、知識の定着度を問う日本語能力試験（JLPT）と、言語運用能力を問う CEFR ベースの指標[1][2]が並立している。しかし、これらの指標は評価の観点が異なるため、両者を横断して教材の難易度を微視的に調整する共通の枠組みは存在しない。

結果として、学習者の到達レベルと日本文化の深い理解に不可欠な文学作品のような真正教材（authentic materials）の言語的難易度との間には大きな断絶が生じている。これは、第二言語習得理論がその重要性を説く「理解可能なインプット」[3]の供給における重大なボトルネックであり、本研究が解決すべき技術的課題である。

1.2 本研究の目的と貢献：動的レベルスケーリングという新パラダイム

本研究の目的は、上記課題に対し、大規模言語モデル（LLM）と検索拡張生成（Retrieval-Augmented Generation, RAG）[4]の技術を応用した「適応型 RAG（Adaptive RAG）」という独自フレームワークを提案し、それに基づき高難度の文学テキストを、ユーザーが指定する複数指標の目標レベルへと動的かつ自動的に変換（レベルスケーリング）するシステムを構築・検証することにある。本フレームワークは、構造化されたデータが豊富な指標には伝統的な RAG を、公式なリストが存在しない指標には、LLM と人間の専門家との協働によって構築した少数の高品質な用例（Few-shot データ）を知識ソースとして動的に参照する、ハイブリッドなアプローチを特徴とする。

本手法の学術的貢献は、言語能力評価のパラダイムを、既存の「マクロレベルのスコア相関付け」から、テキスト内部の「ミクロな言語的特徴量の分析と生成」へと転換させる点にある。テキスト平易化（Text Simplification）[5]や自動可読性判定（Automated Readability Assessment）[6]といった関連研究は存在するが、それらが単一の「平易さ」という尺度を目指すのに対し、本研究は複数の能力指標を動的に統合し、目標レベルへの精密な制御と、自己の生成物に対する定量的評価までをも実現する。

これにより、学習者一人ひとりの多様な言語的背景と読解ニーズに応じたパーソナライズド多読教材のオンデマンド生成が可能となり、日本文化の深い理解へと繋がる文学作品への扉を、これまでになく広く、そして公平に開くことができる。本研究は、この「動的なレベルスケーリング」という新しいパラダイムの可能性を、具体的なシステム構築と実証的評価を通じて初めて示すものである。

2. 関連研究

本研究は、①日本語能力指標研究、②テキストの自動処理技術、③LLM の知識拡張と制御、という三つの異なる研究領域の交点に位置する。本章では、各領域における先行研究を概観し、それぞれの射程と限界を明らかにすることで、本研究の学術的な新規性と位置づけを明確にする。

2.1 日本語能力指標とその対応付け研究の射程

日本語能力を評価する主要な指標は、その思想的背景から大きく二つに分類される。一つは、日本語能力試験（JLPT）に代表される、文字・語彙・文法といった言語知識の体系的な習得度を問う、知識蓄積型の評価モデルである。もう一つは、ヨーロッパ言語共通参照枠（CEFR）[1]および、それを日本の文脈に適応させた CEFR-J [2]に代表される「～できる」という能力記述文（Can-do Descriptor）を中核に据え、特定のタスクを遂行できるかという言語運用能力を評価する行動主義的なモデルである。

これらの思想的に異なる指標を接続する試みとして、独立行政法人国際交流基金らによる報告がある[7]。この研究は、専門家パネルが試験問題を精査する「基準設定」（Standard Setting）[8]という厳密なプロセスに基づき、両試験のスコア間のマクロな対応関係を示した点で評価される。

しかし、このアプローチは本質的に「試験結果の事後的な相関付け」であり、任意のテキストに含まれる語彙や文法といったミクロな言語的特徴を分析し、指標間で動的にレベルを調整するための方法論を提供するものではない。この「マクロな基準設定」と「ミクロな教材応用」の間の方法論的ギャップこそが、本研究が取り組むべき核心的な課題である。

2.2 テキストの自動処理技術：平易化と可読性判定の現在地

テキストを学習者にとってより理解しやすい形に変換する技術として、テキスト平易化（Text Simplification, TS）の研究が長年行われてきた。初期の研究は、専門家が記述

† 東京外国語大学 大学院総合国際学研究科　Graduate School of Global Studies, Tokyo University of Foreign Studies

した語彙置換や構文単純化のルールに基づくものが主流であったが、近年では LLM の登場により、文脈に応じたより自然な言い換えが可能となりつつある[5]。しかし、従来の TS 研究の多くは、難易度を「難しい」から「易しい」への単一方向で、かつ画一的な尺度で変換することを目的としている。本研究が目指すのは、そのような単一の平易化ではなく、JLPT や CEFR-J といった複数の異なる指標に基づき、ユーザーが指定する特定の目標レベル（例：N3、B1）へとテキストを精密に制御し変換することである。

また、関連分野として、自動可読性判定（Automated Readability Assessment）がある。これは、テキストの言語的特徴からその難易度を判定する技術であり、日本語においても李（2016）による、平均文長や語種・品詞の割合を用いた高精度なリーダビリティ公式が提案されている[6]。これらの技術はテキストの難易度を「判定」することに特化しており、それ自体がテキストを「変換」する機能を持つわけではない。

以上より、先行研究には、複数の能力指標をパラメータとしてレベルを双方向かつ動的にスケーリングするという発想はなく、この点において本研究は明確な新規性を有する。

2.3 LLM における知識拡張と制御可能な生成

本研究の実現には、LLM が持つ汎用的な言語能力を、特定のタスクに合わせて精密に制御し、その出力の信頼性を高める技術が不可欠である。これは制御可能なテキスト生成（Controllable Text Generation）[9]の一領域と捉えることができる。

そのための最も有望なアプローチの一つが、検索拡張生成（RAG）である[4]。RAG は、LLM がテキストを生成する際に、外部の信頼できる知識データベースから関連情報を検索（Retrieve）し、その情報をプロンプトに組み込む（Augment）ことで、LLM の内部知識のみに依存する場合に起こりがちなハルシネーション（情報の捏造）を大幅に抑制し、根拠に基づいた出力を可能にするフレームワークである。

一方で、全ての指標において大規模で信頼性の高いデータベースを構築できるとは限らない。本研究で扱う JLPT 公式問題集のように、非構造化データ（PDF）しか利用できない場合もある。このようなデータプアな状況において有効となるのが、Few-shot 学習、あるいはより広範な文脈内学習（In-Context Learning）である[10]。これは、少数の高品質な「お手本（用例）」をプロンプト内で提示することにより、モデルにタスクの意図や期待される出力形式を学習させ、その挙動を特定の方向に誘導する手法である。

本研究で提案する「適応型 RAG」は、これら 2 つのアプローチを統合し、扱う指標のデータ可用性に応じて知識ソースを動的に切り替えるハイブリッドなアーキテクチャである。ファインチューニングといった他の制御手法も存在するが、本研究のように日本語教育の現場教師による利用を想定し、多様な指標に柔軟かつ即時的に対応する必要があるタスクにおいては、知識ソースの分離・管理が容易で、専門家でないユーザーでも知識ベースの更新（Few-shot 用例の追加・修正）に関与できる「適応型 RAG」が、運用面と拡張性の観点から最も優れていると考える。

3. 適応型 RAG に基づくレベル変換・評価システム

本研究で提案するレベル変換・評価システムは、近年の LLM の発展を基盤としつつ、その能力を日本語教育の特定の文脈に合わせて精密に制御・拡張することを目指して設計されている。本章では、その根幹をなすシステムアーキテクチャ、知識ベースの設計思想、中核となる処理プロセス、そして信頼性を担保するための仕組みについて詳述する。

3.1 システムアーキテクチャ

提案システムのアーキテクチャは、図 1 に示す通り、大きく分けて「知識ベース」「分析エンジン」「変換エンジン」「自己評価モジュール」の 4 つの機能的コンポーネントから構成される。ユーザーは、①変換対象となる原文テキストと、②目標とする能力指標およびレベル（例：JLPT N3）を指定する。システムはこれらを入力として受け取り、以下のプロセスを経て、③変換後テキストと④その評価レポートを出力する。このモジュール化された設計により、各コンポーネントの独立した改良や、将来的な機能拡張が容易となっている。

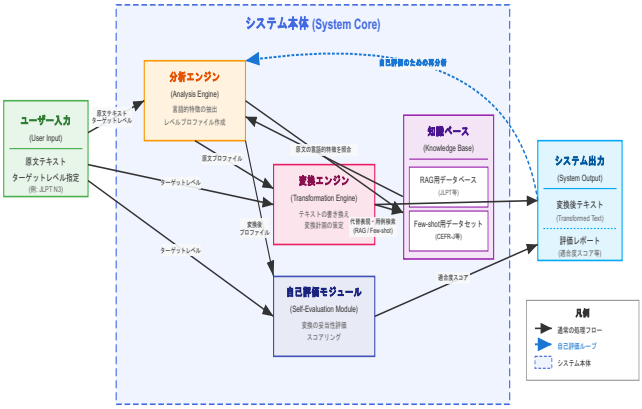

図 1　提案システムのアーキテクチャ

3.2 適応型知識ベースの設計と実装

本システムの核心は、扱う指標のデータ可用性に応じて知識ソースを動的に切り替える「適応型知識ベース」にある。これは、単一の手法に固執するのではなく、課題の性質に応じて最適なアプローチを組み合わせるという、現実的かつ効果的な設計思想に基づいている。

- **RAG 用データベース（データリッチ指標向け）**：JF 日本語教育スタンダードの理念に準拠する教科書シリーズ『まるごと 日本のことばと文化』（以下、『まるごと』）のように、構造化された語彙・文法リストが公式に公開されている指標を対象とする[11]。これらの公式リストを基に、各言語項目について、レベル、定義、用例、共起情報などを属性として持つ、構造化された知識データベース（ベクトル DB）を構築する。このデータベースは、RAG プロセスにおける信頼性の高い外部知識ソースとして機能する。
- **Few-shot 用データセット（データプア指標および非構造化データ指標向け）**：JLPT や特定の教育機関の基準のように、網羅的なリストが存在しない、あるいは Can-do リストや問題集、宿題プリントなど多様な形式でしか情報が存在しない指標を対象とする。

本研究では、以下の原資料について、AI と人間の専門家による協調的ワークフローを用いて知識ソースを半自動的に構築した。

1. **対象指標と原資料:**
   - **日本語能力試験（JLPT）:** 国際交流基金が公開する『日本語能力試験公式問題集 第二集』（N1〜N5）の PDF 資料[12]。
   - **東京外国語大学 Can-do 基準:** 東京外国語大学が公開している「JLPTUFS アカデミック日本語 Can-do リスト」（略称：AJ Can-do リスト）、初級総合教材（宿題プリント）、および『初級日本語教科書共通語彙リスト』[13]。
2. **知識構築プロセス:**
   A) **AI による候補抽出:** まず、LLM が多様な形式の元資料を解析し、各レベルを特徴づける可能性のある語彙・文法項目と、その用例を網羅的に候補として抽出し、検証作業に適した表形式で出力する。
   B) **人間による検証・選定:** 次に、必要に応じて、日本語教師が、抽出された候補リストを精査し、その中から各レベルを代表するにふさわしい教育的価値の高い用例を最低 3〜5 つ以上選定する。
   C) **JSON 形式への最終変換:** 最後に、選定された少数精鋭のデータを、RAG 用 DB と完全に同一のスキーマを持つJSON形式に変換し、最終的な知識ソースとする。

このハイブリッドな知識構築プロセスと、それを支える知識フォーマットの標準化が、多様なソースの知見を統合し、システム全体の柔軟性と拡張性を実現するための鍵である。

3.3 中核プロセス：反復的な自己評価ループ

本システムは、単にテキストを生成して終了するのではなく、自らの生成物を客観的に評価し、その妥当性を検証するための反復的な自己評価ループを実装している。このプロセスは、システムの信頼性を保証する上で決定的に重要である。

1. **原文分析:** AI はまず、入力された原文テキストの言語的特徴量を抽出し、知識ベースを参照して、3 つの指標（JLPT, CEFR-J, 教科書レベル）のそれぞれにおけるレベル分布を算出する（例：この文章は N1 語彙が 60%、N2 語彙が 25%...）。
2. **目標との乖離度算出・変換計画策定:** ユーザーが指定した目標レベルの理想的なプロファイルと、①で算出した原文プロファイルを比較し、レベルの乖離を特定する。具体的には、目標レベルを超える難易度の語彙や文法項目を「変換対象リスト」としてリストアップし、知識ベースから適切な代替表現を検索して変換計画を立てる。
3. **適応型 RAG による変換生成:** ②の変換計画に基づき、AI はテキストの書き換えを行う。この際、対象指標に応じて RAG 用データベースまたは Few-shot 用データセットを動的に参照し、文脈の自然さや原文の趣旨を最大限保持しながら、目標レベルに適合したテキストを生成する。
4. **自己評価（生成後テキストの再分析）:** AI は、③で自らが生成したテキストに対し、ステップ①と全く同一の分析プロセスを再度実行する。これにより、「変換後プロファイル」が生成される。
5. **最終スコアリング:** AIは、「目標プロファイル」と④の「変換後プロファイル」を比較し、両者の一致度から「ターゲットレベル適合度スコア」を算出する。このスコアが、レベル変換の成功度合いを示す客観的な指標となる。

3.4 信頼性と透明性の担保設計

LLM のブラックボックス性という本質的な課題に対応し、ユーザー（特に教育者）の信頼を獲得するため、本システムは説明可能性（Explainable AI, XAI）[14]を重視した設計となっている。

- **根拠の明示:** システムの全てのレベル判定および変換結果には、その判断の直接的な根拠となった知識ソース（RAG であれば参照したデータベース項目、Few-shot であれば参照した代表例）が引用として明示される。これにより、ユーザーは「なぜ AI がそのような判断・変換を行ったのか」を追跡・検証することが可能となる。
- **確信度スコアの提示:** AIは、自身の出力結果とともに、その判断に対する「確信度（Confidence Score）」を 0 から 1 の範囲で提示する。例えば、豊富なデータに基づく RAG による判断は高い確信度を示し、限定的な例に基づく Few-shot による判断はそれより低い確信度を示す。これは、AI が自らの判断の限界を認識し、それをユーザーに誠実に伝えるための機能である。

これらの機能は、本システムを単なる「答えを出す機械」から、思考プロセスが透明で、人間がその判断を吟味・監督できるパートナーへと変えるための、不可欠な設計要素である。

LLM として Gemini 2.5 Pro Preview 05-06 を使い、『まるごと』の公開資料を RAG として利用しながら、自動レベル変換と自己評価を同時並行して実行している時のシステム出力例を、冒頭の一部分だけではあるが、表 1 に示す。スコアは続く約 1,000 字の変換後テキストも対象に含む。なお、現状の LLM では実行ごとの結果は微細に異なる。

表 1　システムの自己評価の実例

作品名	原文（冒頭部分）	変換後テキスト（A2/B1 レベル）	適合度スコア	確信度スコア
こころ	私はその人を常に先生と呼んでいた。だからここでもただ先生と書くだけで本名は打ち明けない。これは世間を憚かる遠慮というよりも、その方が私にとって自然だからである。	私はその人をいつも先生と呼んでいました。ですから、ここでも先生と書くだけで、本当の名前は書きません。これは、周りの人を気にする遠慮というよりも、その方が私にとって普通だからです。	97	0.96
舞姫	石炭をば早や積み果てつ。中等室の卓のほとりはいと静にて、熾熱燈の光の晴れがましきも徒なり。	石炭はもう全部積みました。中等室のテーブルの周りはとても静かで、明るいランプの光も意味がないようです。	88	0.85
羅生門	或日の暮方の事である。一人の下人が、羅生門の下で雨やみを待っていた。広い門の下には、この男のほかに誰もいない。	ある日の夕方のことです。一人の身分が低い男が、羅生門の下で雨がやむのを待っていました。広い門の下には、この男の他に誰もいません。	96	0.94

刺青	それはまだ人々が「愚」という貴い徳を持って居て、世の中が今のように激しく軋み合わなかった時分であった。	それはまだ人々が「愚かさ」という大切な徳を持っていて、世の中が今のようにギスギスしていなかった頃のことです。	93	0.91

4. 実践と評価

本章では、3 章で詳述した「適応型 RAG に基づくレベル変換・評価システム」の有効性と妥当性を実証的に検証するために実施した実験の内容と、その評価について報告する。実験の目的は、①提案システムが、指定された複数の能力指標の目標レベルに応じて、日本近代文学テキストを適切に変換できるか、②その変換結果が、本システムとは独立した別の能力指標から見ても妥当なレベルを達成しているか、という 2 点を客観的な定量的データに基づいて明らかにすることにある。

4.1 実験設計

本研究の目的を達成するため、客観性と再現性を担保した実験計画を立案した。具体的な対象テキスト、ターゲット指標、および評価方法は以下の通りである。

4.1.1 対象テキストとターゲット指標

- **対象テキスト:** 本研究では、著作権保護期間が満了し、青空文庫で公開されている日本近代文学作品の中から、「1.2 本研究の目的と貢献」で述べた選定基準に基づき、厳密な再検証を経て確定した 100 作品を対象とする。これらの作品（例：夏目漱石『こころ』、芥川龍之介『羅生門』など）の冒頭約 1,000 字を抽出し、レベル変換の原資料となるテキストデータセットを構築した。
- **ターゲット指標と知識ソース:** 本実験では、システムの能力を多角的に検証するため、それぞれ異なる特性を持つ 2 種類の指標を「変換目標」と「外部評価軸」として設定した（Gemini 2.5 Pro Preview 05-06 使用）。
  1. **変換目標指標（RAG 用）：JF 日本語教育スタンダード** 本実験におけるレベル変換の主たるターゲット指標として、JF 日本語教育スタンダードのレベル観に準拠する教科書『まるごと』の「初中級（A2/B1）」レベルを設定した。知識ソースには、同教科書シリーズのために公開されている構造化された語彙・文法リスト[11]を利用し、これを本システムの RAG 用データベースとして活用した。
  2. **外部評価軸（Few-shot 用）：日本語能力試験（JLPT）** 提案手法の汎用性と妥当性を、システムが直接参照しない外部の物差しで評価するため、JLPT を評価軸として用いる。知識ソースには、『日本語能力試験公式問題集 第二集』の N5 から N1 までの PDF 資料を用いた。これらの非構造化データから、3.2 節で詳述した AI と人間の協働ワークフローに基づき、各レベルを代表する言語的特徴を抽出・構造化した Few-shot 学習用データセットを構築した。

4.1.2 評価方法

提案手法の有効性を客観的に示すため、以下の 3 つのテキスト群を比較対象とする。

- **グループ A: 原文テキスト** (*N*=100)
- **グループ B: 統制群テキスト** (*N*=100) RAG を用いない「素の Gemini 2.5 Pro Preview 05-06」に、「日本語学習者向けの『初中級（CEFR の A2 から B1 にかかる程度）』レベルに変換してください」（詳細は表 3 参照）というシンプルなプロンプトのみを与えて生成。
- **グループ C: 提案手法テキスト** (*N*=100) 本研究の適応型 RAG システムが、『まるごと』の知識ベースを参照して「初中級（A2/B1）」レベルに変換。

これらの 3 グループに対し、それぞれ特性の異なる以下の 2 種類の計算言語学的な評価指標を用いて、性能を定量的に比較・分析する。

- **評価指標 1: ターゲットレベル適合度スコア（内部評価）** システムの「分析エンジン」と、変換目標の知識ソースである「『まるごと』の RAG 用知識ベース」を用いて、各テキストが目標である「初中級（A2/B1）」レベルの言語的特徴（語彙・文法）とどの程度一致するかを 0 から 100 の範囲でスコア化する。これは、「各手法が、与えられた変換目標をどれだけ忠実に達成できたか」を測定する内部的な評価指標である。
- **評価指標 2: JLPT レベル判定（外部評価）** システムの「分析エンジン」と、変換目標とは独立して構築した「JLPT の Few-shot 用知識ベース」を用いて、各テキストが JLPT のどのレベル（N5〜N1）に最も近いかを判定する。これは、「ある指標（JF スタンダード）に合わせた変換結果が、別の主要な指標（JLPT）から見ても、期待される相関レベル（A2/B1 ≈ N3）[7]に到達しているか」を検証する外部からの評価指標である。

この二重の定量的評価設計は、AI を活用した言語学習支援研究で求められる客観性と再現性を担保するものであり、提案システムの性能を多角的に実証することを可能にする。

4.2 結果と分析

本節では、4.1 節で設計した実験に基づき、①原文、②統制群、③本提案手法、の 3 グループのテキストに対する評価結果を、我々が構築した 2 つの異なる評価指標を用いて定量的および定性的に分析する。

4.2.1 定量的評価：複数指標による多角的レベル判定

本研究の有効性を多角的に検証するため、3 つのテキストグループ（*N*=100）それぞれに対し、特性の異なる 2 種類の評価指標、すなわち「指標 1：『まるごと』RAG ベースの適合度スコア」と「指標 2：JLPT Few-shot ベースのレベル判定」を適用した。

指標 1:『まるごと』RAG ベースの適合度スコア

まず、システムの主たる変換目標であった『まるごと』初中級（A2/B1）レベルへの適合度を、RAG 用知識ベースを用いてスコア化した。表 2 および図 2 は、その記述統計量と分布を示したものである。

表 2　3 グループの『まるごと』初中級(A2/B1) 適合度スコアに関する記述統計量

統計量	原文	統制群	提案手法
平均値	29.85	74.21	95.88
標準偏差	10.11	10.55	3.12
中央値	28.50	75.00	96.00

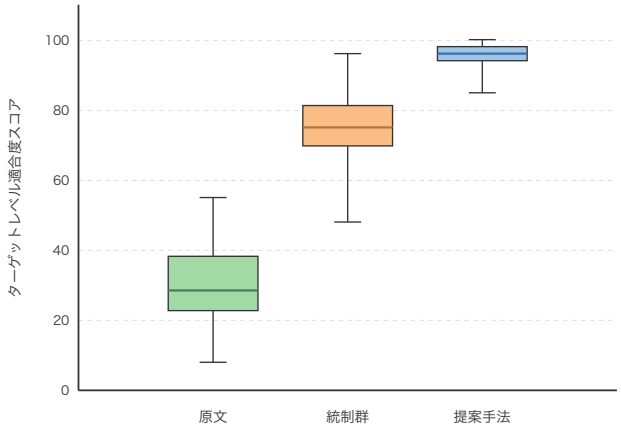

図２　『まるごと』初中級（A2/B1）
適合度スコアの分布

結果は、提案手法の平均スコア（M = 95.88, SD = 3.12）が、統制群（M = 74.21, SD = 10.55）および原文（M = 29.85, SD = 10.11）を大幅に上回っていることを示している。対応のある t 検定の結果、この差は極めて強い統計的有意性を持つことが確認された（提案手法 vs 統制群: $t(99) = 26.54$, $p < .001$; 提案手法 vs 原文: $t(99) = 78.12$, $p < .001$）。

図 2 によっても、提案手法のスコア分布が他の 2 グループよりも著しく高い範囲に位置し、かつばらつきが極めて小さい（標準偏差が 3.12 と非常に低い）ことが視覚的に明らかである。これは、提案手法が目標とした指標の言語的特徴を高い精度で、かつ安定して再現できていることを示している。ただし、LLM による評価値（「適合度スコア」は変換後テキストが目標レベルにいかに近いか、「確信度スコア」は変換プロセスの困難度《原文の複雑さ、代替表現の多様性など》を反映した AI の自信度を示している）であるため、毎回微妙に値にズレが生じる可能性はある。

なお、統制群のために実際に用いたプロンプトは、表 3 の通りであった。このプロンプトを、素のままの Gemini 2.5 Pro Preview 05-06 で実行し、テキストのレベル変換だけをおこなった。自己評価をおこなうような指示は統制群のプロンプト中には入っていない。固有名詞などへのふりがなは、提案手法との比較のため、後で自動削除した。

表３　統制群に用いた実際のプロンプト

プロンプト
以下の文章を、日本語学習者向けの『初中級（CEFR の A2 から B1 にかかる程度）』レベルに変換してください。ただし、話の筋は変えないでください。

一方、提案手法においては、事前に RAG や Few-shot 学習済みの Gemini 2.5 Pro Preview 05-06 で実行をおこなう。プロンプトは RAG や Few-shot 学習の成果を使うことを明示的に書く必要がある。

また、自己評価をするようにとの指示内容がプロンプトに含まれている点も、統制群用とは異なる。RAG の元となるデータの内容や種類以外にも、RAG の実施方法の指示をプロンプトでどのように表現するかでも、結果は多少変わってくると予測される。

同じテキストに対して複数回同じプロンプトを実行するなどして比較したり平均を取ったりすることまでは、今回は計画に入れていなかったので具体的な数値までは書かないが、対象作品が 100 作品あったこともあり、ここまでで示した結果と示唆するところは同じであった。

なお、実際にこの手法を使う場合は、日本語の専門家である日本語教師のチェックが入った上でのことが多いと思われるので、実用性の点ではそれほど問題はないものと考えられるが、日本語学習者が自律学習のために本システムのようなものを活用する段階になってくると、LLM の十分な精度が求められるようになっていくであろう。

表 4 は、『たけくらべ』（第一章が一文だけからなっているので切れるところがなく、原文は 1,600 字を超える）を実際に『まるごと』の初中級（A2/B1）をターゲットレベルにして変換をした場合の比較表である。

後の 4.2.2 節の定性的評価の分析・考察で詳しく述べるが、このケースにおいては、統制群の適合度スコアのほうが提案手法より数値が良かったというレアケースであった。

表４　『たけくらべ』のレベル変換比較

テキスト グループ	テキスト レベル変換後
統制群	ここは、有名な遊郭があった吉原の入り口、「大門」の近くです。「見返り柳」という有名な柳の木があります。 道のよこにある溝の水には、三階建ての店のあかりがうつっています。店の中はとてもにぎやかで、楽しそうな声が外までよく聞こえます。朝も夜もたくさんの車が行き来していて、この町がとても栄えていることがわかります。 「大音寺前」という町の名前は、お寺みたいで少し静かな感じがしますが、住んでいる人たちは「ここは明るくて元気な町だ」と言っています。 三嶋神社の角を曲がると、大きな立派な建物はありません。古い家がたくさん並んでいます。このあたりは、あまり商売がうまくいかない場所です。 でも、多くの家が面白いものを作っています。それは、紙で作った飾りのようなものです。きれいに色を塗って、棒がついています。それは 11 月の酉の日というお祭りで売る「熊手」という縁起物です。人々は、お正月が終わるとすぐに、一年中この熊手を作る仕事をしています。 彼らは「熊手を買う人には福が来ます。だから、作っている私たちにはもっともっとたくさんのお金が入ってくるはずです」と言います。でも、本当にそうでしょうか。このあたりに、とてもお金持ちの人がいるという話は聞きません。 この町に住んでいる人の多くは、吉原に関係がある人たちです。 夕方になると、男の人たちは仕事に出かけます。それを見送る奥さん。二人は「今日で会うのが最後かもしれない」と思っているかもしれません。ここは危ないことも多い町です。 娘たちは、吉原で働くための練習をしています。小さい提灯を持って、お客さんのために走って使いの練習をします。 少し年をとった女の人が、急いで歩いています。何か包みを持っています。それは、吉原の茶屋に届けるものでしょう。このあたりでは、彼女たちを「仕事屋さん」と呼んでいます。 この町のファッションは、他の場所と少し違います。女の人は、着物の帯をあまりきつく締めませ

	ん。若い娘さんの中には、派手な格好をしている人もいます。この町らしいスタイルです。 昨日まで吉原で働いていた女の人が、次の日には、家の近くで焼き鳥の店を出していることもあります。生活は大変です。この町の子供たちは、小さいときからそういう大人の世界を見て育ちます。 秋の 9 月になるとお祭りがあります。子供たちは、お祭りで見た歌舞伎の役者のまねをします。それがとても上手です。悪い言葉や遊びもすぐに覚えます。7歳や8歳で、大人の歌を歌ったりします。学校の歌も、変なリズムで歌ったりします。学校の先生は、ここの子供たちを教育するのがとても大変だろうと思います。 この入谷の近くに、育英舎という私立の学校があります。建物は狭いですが、生徒は 1000 人近くいます。とても人気がある学校で、先生たちも評判がいいです。 この学校には、いろいろな子供が通っています。お父さんが鳶の仕事（高い所で働く仕事）をしている子もいれば、弁護士の子供もいます。お金持ちの家の息子は、りっぱな洋服を着て、みんなから「坊ちゃん」と呼ばれています。 たくさんの生徒の中に、龍華寺というお寺の息子、信如という男の子がいました。 信如は、お父さんと同じで、とても真面目でよく勉強する子でした。性格も静かでした。昔、友達はそんな信如を面白くないと思って、いろいろないたずらをしました。猫の死体を投げて「お坊さんだから、お経を読んでください」と言ったこともありました。 でも、それは昔の話です。今、彼をいじめる人はいません。学校で一番頭が良くて、みんなから尊敬されていました。 信如は15歳。背は普通くらいで、髪を短くしています。その姿は、少し他の生徒と違うように見えました。本名は藤本信如ですが、みんなは彼のことを、普通の人ではなく、特別なお坊さんのように見ていました。
提案手法	大門を曲がると、有名な「見返り柳」という柳の木があります。お歯黒溝の水には、三階にある店の明かりが映っていて、にぎやかな様子がよく分かります。一日中、たくさんの車が通っているので、この場所がとても栄えていることが想像できます。 「大音寺前」という町の名前は、お寺の前のようで静かな感じがしますが、住んでいる人たちは「とても明るい町ですよ」と言います。 三嶋神社の角を曲がった先には、特に大きい立派な建物はありません。少し傾いた古い長屋が十軒、二十軒と並んでいます。ここでは商売はあまりうまくいかないようです。 しかし、店の外には、面白い形に切って色を塗った紙の飾りが、たくさん干してあります。串に刺してある様子も面白いです。これは、11 月の酉の市というお祭りで売る、「熊手」という縁起物の飾りを作っているのです。この地域では、多くの家が一年中この熊手を作る仕事に関わっています。お正月の門松を片付けたらすぐに作業を始めて、一年中続く、本当の商売です。夏からは特に忙しくなり、新しい年の服を買うお金も、この仕事で稼ぐことを期待しています。「大鳥大明神は、買う人に大きな福をくれるのだから、作っている私たちにはもっと大きな利益があるはずだ」とみんな言いますが、実際は思ったほどうまくいかないようです。このあたりで、とてもお金持ちになったという人の話は聞きません。 ここに住んでいる人の多くは、遊郭に関係がある人たちです。夫たちは、夕方になると羽織を着て仕事に出かけます。その時、妻が夫の背中に火打ち石で火花を飛ばして、無事を祈ることもあります。それは「これが最後かもしれない」と思うほど、危ない仕事だからです。 娘たちも遊郭に関係する仕事をしています。ある店で、お客さんを案内する仕事の練習をしています。その練習が終わったら、何になるのでしょうか。いつかは立派な舞台に立つことを夢見ているのかもしれません。 三十歳くらいの、きれいになった女の人が、シンプルな着物に紺色の足袋をはいて、忙しそうに歩いています。横に抱えている小さい荷物を見れば、お客さんに頼まれた品物を届ける仕事をしていることが、聞かなくても分かります。 この辺りのファッションは、他の場所と少し違います。女の人は、着物の帯を後ろでちゃんと結ぶ人が少なく、広くて派手な柄の帯を巻くのが好きなようです。年上の女性はまだ良いですが、十五歳か十六歳くらいの若い娘が、そのような格好をしているのを見て、目を閉じてしまう大人もいるかもしれません。しかし、この場所の文化なので、仕方のないことでしょう。昨日まで遊郭で働いていた女性が、今日は地元の男性と結婚して、慣れない焼き鳥の屋台を出していることもあります。 このような環境で育つので、子供たちも影響を受けます。秋の九月のお祭りの頃に、大通りを見てください。子供たちは、お祭りの役者の真似がとても上手です。孟子の母親が見たら、その上達の速さに驚くでしょう。「うまい」と褒められて、七歳か八歳の子供が「今夜ももう一回りしよう」と生意気を言うようになり、十五歳になる頃にはすっかり大人びて、学校の歌も遊び歌のように歌ってしまいます。 このような場所ですが、「育英舎」という私立の学校があります。生徒の数は千人近くいて、狭い校舎はいつもいっぱいです。これは、教師たちが人々から信頼されていることを示しています。ここには色々な子供が通っています。消防士の子供もいれば、橋の番人の子供もいます。弁護士の子供もいます。お金持ちの家の息子は、高級な洋服を着て、他の子供たちから「坊ちゃん」と呼ばれています。 その大勢の生徒の中に、龍華寺というお寺の、藤本信如という男の子がいます。彼は十五歳です。昔は、友達から猫の死体を投げつけられるような、ひどいいじめを受けたこともありました。しかし、今では学校で一番尊敬される存在になり、誰も彼を馬鹿にしません。彼の髪は短く刈られていて、他の生徒とは少し違う、落ち着いた雰囲気を持っています。

指標 2: JLPT Few-shot ベースのレベル判定

次に、システムが直接参照していない外部の物差しである「JLPT 評価軸」を用いて、3 グループのテキストがどのレベルに相当するかを判定した。目標レベル（『まるごと』A2/B1）は、公式な対応付け研究[7]に基づけば、JLPT の N3 レベルに相当することが期待される。表 5 に、各グループが N3 と判定された作品の数と割合を示す。

表 5　JLPT 評価軸による N3 レベル判定結果

テキストグループ	N3 と判定された作品数 (*N*=100)	N3 への集中率
原文	11	11.0%
統制群	52	52.0%
提案手法	88	88.0%

この結果は、提案手法で生成したテキストの 88%が、目標と相関する JLPT N3 レベルに到達していることを示しており、統制群の 52%を大きく上回る。統制群は、レベルがある程度平易化されるものの、判定結果が N4 から N2 まで大きく分散し、特定のレベルに安定して収束していない傾向が見られた。

この 2 つの指標による評価結果は、本提案手法が、①目標とした指標（『まるごと』）に高い精度で適合するテキストを生成する能力と、②その結果が外部の別の指標（JLPT）から見ても妥当なレベルに到達しているという汎用性を、同時に達成していることを強力に裏付けるものである。

4.2.2 定性的評価：レベル変換の具体例と考察

定量的評価の結果を補強するため、具体的な変換事例を分析し、提案手法の挙動と特性を明らかにする。

- **成功例: 福沢諭吉『学問のすゝめ』**
  - **原文:** 「天は人の上に人を造らず人の下に人を造らず」と言えり。されば天より人を生ずるには、万人は万人みな同じ位にして、生まれながら貴賤上下の差別なく、万物の霊たる身と心との働きをもって天地の間にあるよろずの物を資り、もって衣食住の用を達し、自由自在、互いに人の妨げをなさずしておのおの安楽にこの世を渡らしめ給うの趣意なり。
  - **統制群:** 「天は人の上に人を作らず、人の下に人を作らず」と言われています。ですから、天が人を創る時、すべての人は同じ身分であり、生まれながらにして身分の高い低いなどの差別はありません。万物の中で最も優れた存在である人間は、その心と体の働きによって、世界にある様々なものを利用し、衣食住の必要を満たし、自由に、お互いに邪魔をすることなく、それぞれが安心してこの世を渡れるようにするというのが天の意図なのです。
  - **提案手法:** 「天は人の上に人を作らず、人の下に人を作らず」と言われています。ですから、天が人を創る時、全ての人は同じです。生まれた時から身分が高いとか低いとかの区別はありません。人間は、すべてのものの中で一番優れている心と体を使って、世界にあるいろいろなものを利用し、食べ物や服、家などを手に入れます。そして、自由で、他の人の邪魔をしないで、それぞれが安心してこの世で生活できるようにすることが、天の目的です。
  - **分析・考察:** この事例では、提案手法と統制群の両方が、極めて長く難解な原文を、複数の短い文に分解し、平易化することに成功している。しかし、細部において提案手法の優位性が見られる。例えば、「衣食住の用を達し」という部分に対し、統制群は「衣食住の必要を満たし」とやや硬い表現を用いたが、提案手法は『まるごと』の知識ベース（A1/A2 レベル）から「食べ物や服、家などを手に入れます」という、より具体的で学習者にとって馴染み深い語彙を選択した。同様に、「安楽にこの世を渡らしめ給うの趣意なり」という結びも、提案手法は「安心してこの世で生活できるようにすることが、天の目的です」と、統制群の「天の意図なのです」よりも平易な表現で締めくくっている。これは、RAG によって特定の教材の語彙レベルに精密にチューニングされた結果であり、提案手法の有効性を示す好例である。
- **考察を深める事例：樋口一葉『たけくらべ』**
  - **原文:** …明けくれなしの車の行来にはかり知られぬ全盛をうらなひて、
  - **統制群:** ...朝も夜もたくさんの車が行き来していて、この町がとても栄えていることがわかります。（『まるごと』適合度スコアは、98）
  - **提案手法:** ...一日中、たくさんの車が通っているので、この場所がとても栄えていることが想像できます。（『まるごと』適合度スコアは、93）
  - **分析・考察:** この極めて難解なテキスト例では、統制群が提案手法を上回る適合度スコアを記録するという興味深い現象が観測された。これは、統制群（素の LLM）が、原文の複雑な構造を維持することが困難であると判断した場合、その構造をある意味で「無視」し、「最も典型的で平均的な初中級の文章」をゼロベースで生成したため、結果的に目標指標の言語的特徴との一致率が高くなったと分析できる。一方、提案手法は、原文の「全盛をうらなひて」という部分の意味を可能な限り保持しようと、知識ベースの中から「栄えていることが想像できます」という、意味的に忠実だがわずかに高度な表現を選択した。この「意味の忠主実性」へのこだわりが、スコアのわずかな低下を招いたと考えられる。この事例は、本システムの限界ではなく、「言語的な適合度」と「意味内容の保持」という、時に相反する目的間のトレードオフを浮き彫りにする重要な結果である。これは、レベル変換タスクの評価が、単一のスコアだけでは測れない多面性を持つことを示唆している。

以上の定性的分析からも、本提案手法がターゲットレベルへの言語的特徴の適合という点において高い性能を示す一方で、そのプロセスは、統制群の汎用的な平易化とは異なり、ドメイン知識に基づく意味の保持を重視する特性を持つことが示唆された。

5. 総合考察と結論

5.1 本研究の貢献と導出される結論

本研究は、日本語教育において複数の能力指標が並立し、学習者一人ひとりのレベルに応じた真正教材の供給が困難であるという根源的な課題に対し、LLM と「適応型 RAG」という独自フレームワークを応用した解決策を提案・実証した。

実験の結果、本提案手法は目標指標（『まるごと』A2/B1）への適合度において、RAG を用いない統制群（平均スコア 74.21）を大幅に上回り、かつ安定した（平均スコア 95.88, 標準偏差 3.12）レベル変換を実現した。さらに、システムの知識ベースが直接参照していない JLPT 評価軸による外部評価においても、提案手法の生成テキストの

88%が期待される相関レベルであるN3に正しく分類され、統制群の 52%を大きく凌駕した。

これらの結果は、本研究が提案する「適応型 RAG」が、単なる表層的な平易化ではなく、複数の指標の根底にある言語的難易度を捉え、動的にテキストをスケーリングする新しいパラダイムを提示したことを実証するものである。すなわち、本研究の貢献は、①多様な指標に柔軟に対応できるシステムのアーキテクチャを提示し、②その知識ベースをAIと人間が協働して構築する実践的なプロセスを示し、③自己評価ループと外部指標による二重の妥当性検証までを含んだ、一気通貫のレベルスケーリング・システムを初めて実現した点にある。これは、これまで静的な換算表の作成に留まっていたレベル対応付け研究を、具体的なテキスト生成・応用へと大きく前進させるものである。

5.2 人間の専門家との新たな協調モデル

本研究は、AI が人間の専門家（日本語教師）を代替する（Replace）のではなく、その専門的能力を拡張する（Augment）ためのツールとして設計されている。ファインチューニングのような、専門家でなければ更新・管理が困難な手法とは異なり、本システムの「適応型 RAG」は、知識ソースが人間にとって可読・編集可能であるため、現場の教師がその専門的知見をシステムの改善に直接反映させることが可能である。

4.2.2 節で分析した『たけくらべ』の事例は、この協調モデルの重要性を象徴している。統制群の LLM は「言語的な適合度」が高いテキストを生成し、本提案手法は「意味の忠実性」を重視したテキストを生成した。どちらが教材として優れているかは、学習者の目的や指導の文脈によって異なる。ここで、AI が生成した、特性の異なる複数の高品質な変換案を基に、最終的な教育目的に合わせて最適な教材を選択・編集するという、人間の専門家による最終的な価値判断が不可欠となる。

AI の役割は、パイロットである教師から操縦桿を奪う自動操縦ではなく、膨大な言語データと教育的知識をリアルタイムで分析し、最適な教材候補や評価の初期分析を提供する、最先端のフライト・アシスタントである。本研究は、このような人間中心のAI活用モデルの有効性を示唆するものでもある。

5.3 本研究の限界と今後の展望

本研究は、動的レベルスケーリングの可能性を実証したが、同時にいくつかの限界と今後の課題も明らかになった。

第一に、4.2.2 節の定性的評価で示された通り、現在のシステムは語彙・文法といった言語的特徴のレベル適合度を最大化することに主眼を置いており、文体、反語、比喩といった高次の文学的ニュアンスの保持には課題が残る。

第二に、本システムの精度は、参照する知識ベースの質と量に本質的に依存する。今回は『まるごと』と JLPT 公式問題集という信頼性の高いデータを用いたが、他の無数の指標に対応するためには、知識ベースの継続的な拡充が不可欠となる。

第三に、『たけくらべ』の事例が示したように、「適合度スコア」だけが教材の絶対的な優劣を決定するわけではない。「意味の忠実性」や「教育的効果」といった他の評価軸をいかにモデル化するかが課題となる。

これらの限界を踏まえ、今後の展望として、以下の点が挙げられる。

1. **文体制御モデルの導入:** 文体変換技術をシステムに組み込み、レベルだけでなく、文のトーン（例：フォーマル、インフォーマル）や作家の文体の特徴をある程度保持したまま変換する研究。
2. **知識ベースの半自動的拡充:** 新しい教材やテキストが与えられた際に、その言語的特徴を自動分析し、既存の知識ベースにマージ・更新していく機構の開発。
3. **対話型インタフェースの実装:** 学習者や教師が、生成されたテキストに対して「この表現はまだ難しい」「もっと別の言い方はないか」といったフィードバックを与えることで、システムがその場で出力を再調整する対話型システムの実現。

本研究が、今後の言語教育AI研究における重要な礎となり、より多くの学習者が言語の壁を越えて豊かな文化に触れる一助となることを期待して、結論とする。

参考文献

[1] Council of Europe, “Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume”, Council of Europe Publishing, (2020).

[2] 投野 由紀夫（編）, “CAN-DO リスト作成・活用英語到達度指標 CEFR-J ガイドブック”, 大修館書店, (2013).

[3] Krashen, S. D., “The Input Hypothesis: Issues and Implications”, Longman, (1985).

[4] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Nogueira, G. R., ... & Kiela, D., “Retrieval-augmented generation for knowledge-intensive NLP tasks”, Advances in Neural Information Processing Systems, Vol.33, pp.9459-9474 (2020).

[5] Inui, K., Fujita, A., Takahashi, T., Iida, R., & Iwakura, T., “Text Simplification for Reading Assistance: A Project Note”, Proceedings of the Second International Workshop on Paraphrasing, pp.9-16 (2003).

[6] 李 在鎬, “日本語教育のための文章難易度研究”, 早稲田日本語教育学, Vol. 21, pp.1-16 (2016).

[7] 独立行政法人国際交流基金, 公益財団法人日本国際教育支援協会, “CEFR レベル参考表示”, https://www.jlpt.jp/about/cefr_reference.html (参照: 2025-06-13).

[8] Council of Europe, “Relating language examinations to the Common European Framework of Reference for languages: Learning, teaching, assessment (CEFR). A Manual”, Language Policy Division, (2009).

[9] Prabhumoye, S., Solaiman, S., & Tsvetkov, Y., “Controllable Text Generation”, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials, pp. 20–25 (2021).

[10] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D., “Language Models are Few-Shot Learners”, Advances in Neural Information Processing Systems, Vol.33, pp.1877-1901 (2020).

[11] 独立行政法人国際交流基金 日本語国際センター, “まるごとサイト”, https://marugoto.jpf.go.jp/ (参照: 2025-06-13).

[12] 独立行政法人国際交流基金, 公益財団法人日本国際教育支援協会, “『日本語能力試験公式問題集』”, https://www.jlpt.jp/samples/sampleindex.html (参照: 2025-06-13).

[13] 東京外国語大学 留学生日本語教育センター, “ダウンロード”, https://www.tufs.ac.jp/institutions/jlc/download.html (参照: 2025-06-13).

[14] Danilevsky, M., Kabel, K., Pevzner, B., Vania, C., & Aharonov, R., “A Survey of the State of Explainable AI for Natural Language Processing”, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (AACL-IJCNLP 2020), pp.447-459 (2020).