

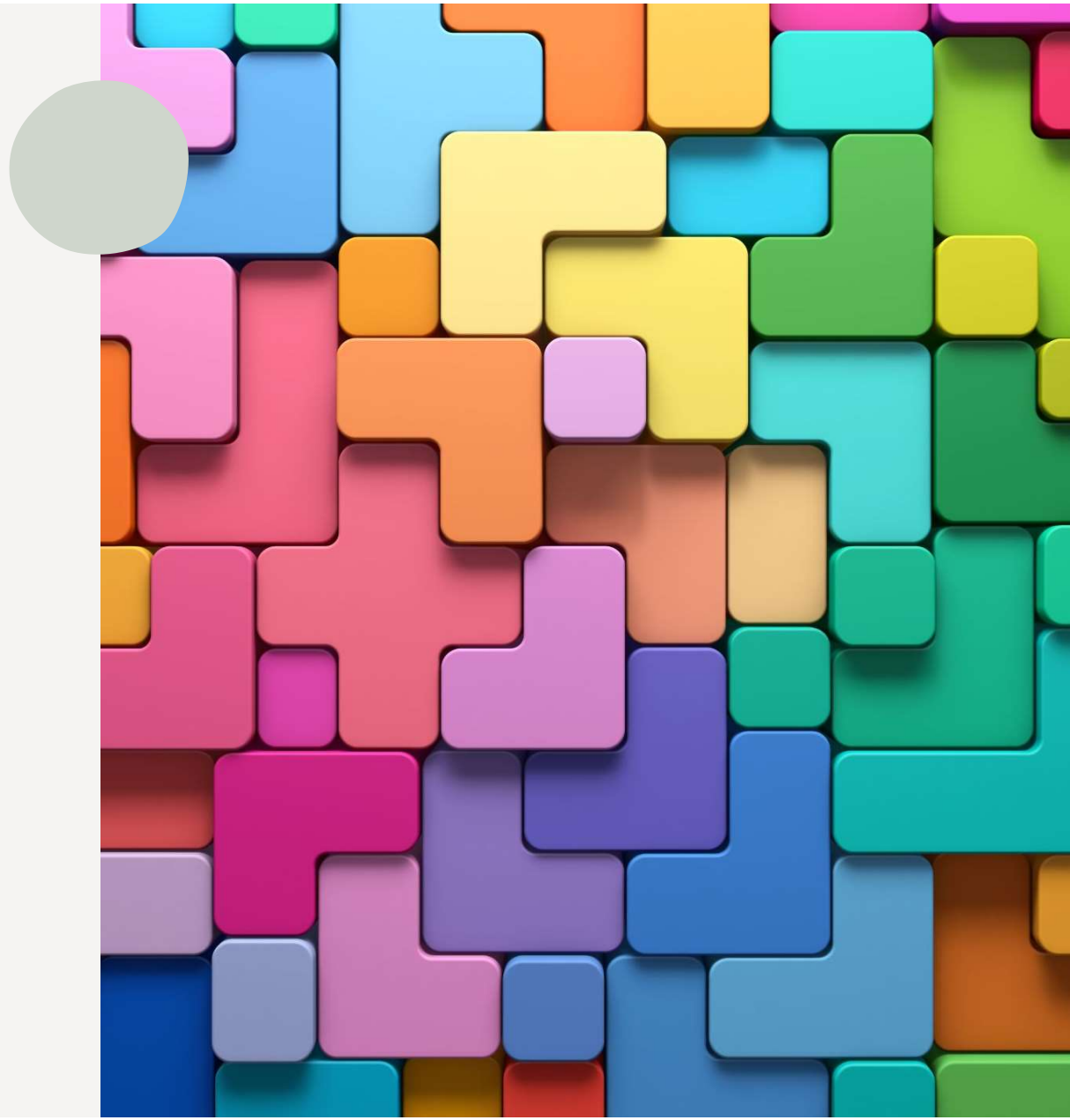
NII DXシンポ (2025.1.14)

日本語教育と生成AI：

パフォーマンス評価の支援ツールとしての ChatGPTの活用例

李在鎬 (りじえほ)

早稲田大学大学院日本語教育研究科



ねらい（伝えたいこと）

1. パフォーマンス評価の重要性と広まっていない現状。
2. 生成AIはパフォーマンス評価を広めるための動力源。
3. 人間教師のフィードバックと**ChatGPT**のフィードバックは類似している。
4. **ChatGPT**のフィードバックは日本語教師がみても納得できるものが多い。



教えること・評価すること

- 大前提：教えることと評価することはコインの表と裏→両者の好循環で良い教育を形作る。

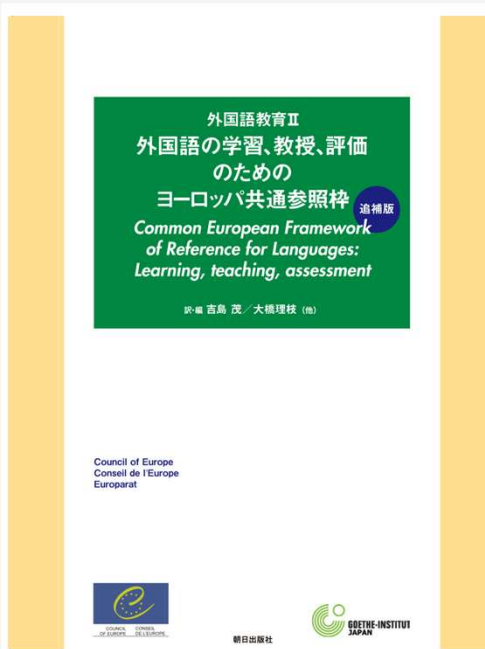
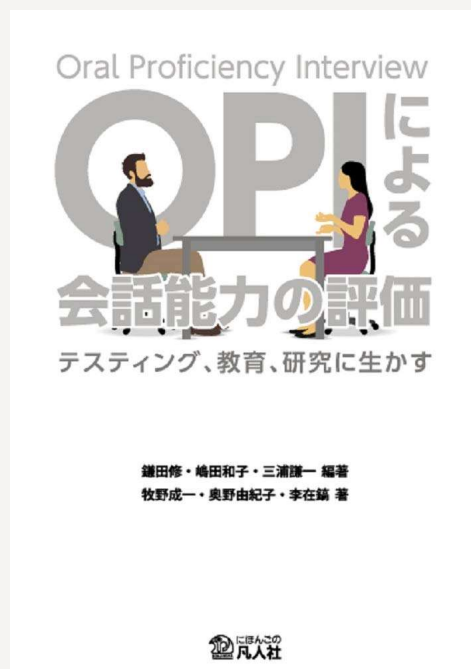


表7 評価の種類

1	達成度評価 Achievement assessment	熟達度評価 Proficiency assessment
2	標準準拠型評価 (NR) Norm-referencing	基準準拠型評価 (CR) Criterion-referencing
3	合否型基準準拠型評価 Mastery leaning CR	連続型基準準拠型評価 Continuum CR
4	継続的評価 Continuous assessment	定評評価 Fixed point assessment
5	形成的評価 Formative assessment	総括的評価 Summative assessment
6	直接評価 Direct assessment	間接評価 Indirect assessment
7	運用評価 Performance assessment	知識評価 Knowledge assessment
8	主観的評価 Subjective assessment	客観的評価 Objective assessment
9	チェックリスト査定 Checklist rating	尺度査定 Performance rating
10	印象評価 Impression judgment	指針に基づいた判断 Guided judgment
11	全体的評価 Holistic assessment	分析的評価 Analytic Assessment
12	シリーズ評価 Series assessment	分野別評価 Category assessment
13	他人による評価 Assessment by others	自己評価 Self-assessment

<https://www.goethe.de/resources/files/pdf191/cefr31.pdf>

日本語教育の文脈で言えば



産出能力の評価

日本語能力試験 JLPT Japanese Language Proficiency Test

言語知識 (文字・語彙) 1 [漢字読み]

——の言葉の読み方として最もよいものを、1・2・3・4から一つ選びなさい。

彼は今、新薬の研究開発に挑んでいる。

1 はげんで 2 のそんで

3 からんで 4 いどんで

<https://www.jlpt.jp/>

つぎへ ▶

理解能力の評価

パフォーマンス評価の教育的価値

- 多肢選択型テストの限界 (Hart 1994)
 1. 理解や反省を犠牲にして、かわりに再生や暗記に過度の価値を置きすぎている
 2. ほとんどの問題には、唯一の正しい答えがあるという誤った印象を強めている
 3. 生徒を、答えや解決法を構築するのではなくただ受容するだけの受け身の学習者にしてしまっている
 4. 生徒にとって学ぶに値する内容にではなく、テストに出されやすい内容に教師の目を向けさせてしまっている
 5. 教える内容を全てマーク形式にしてしまうことによって内容やスキルの発展を軽視してしまう

Diane Hart, *Authentic Assessment : A Handbook for Educators*, Dale Seymour Publications, 1994.

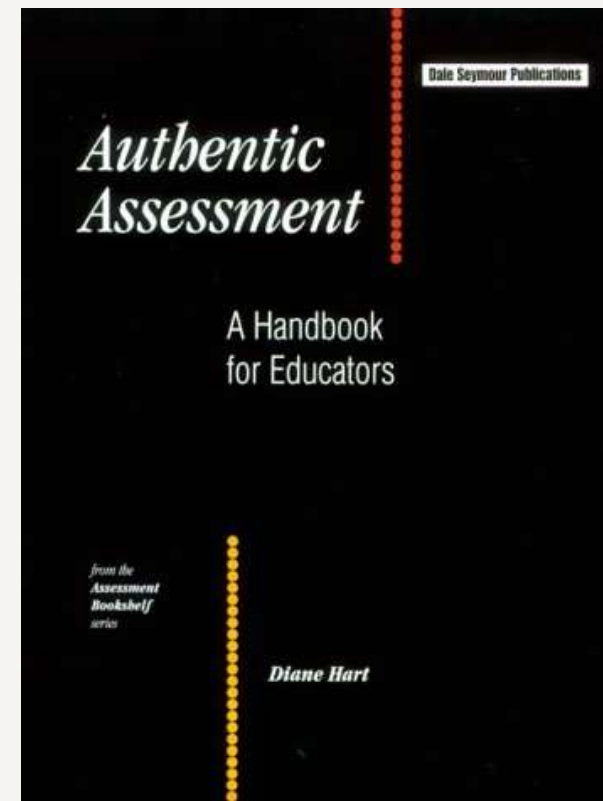


表1 評価の2つのパラダイム

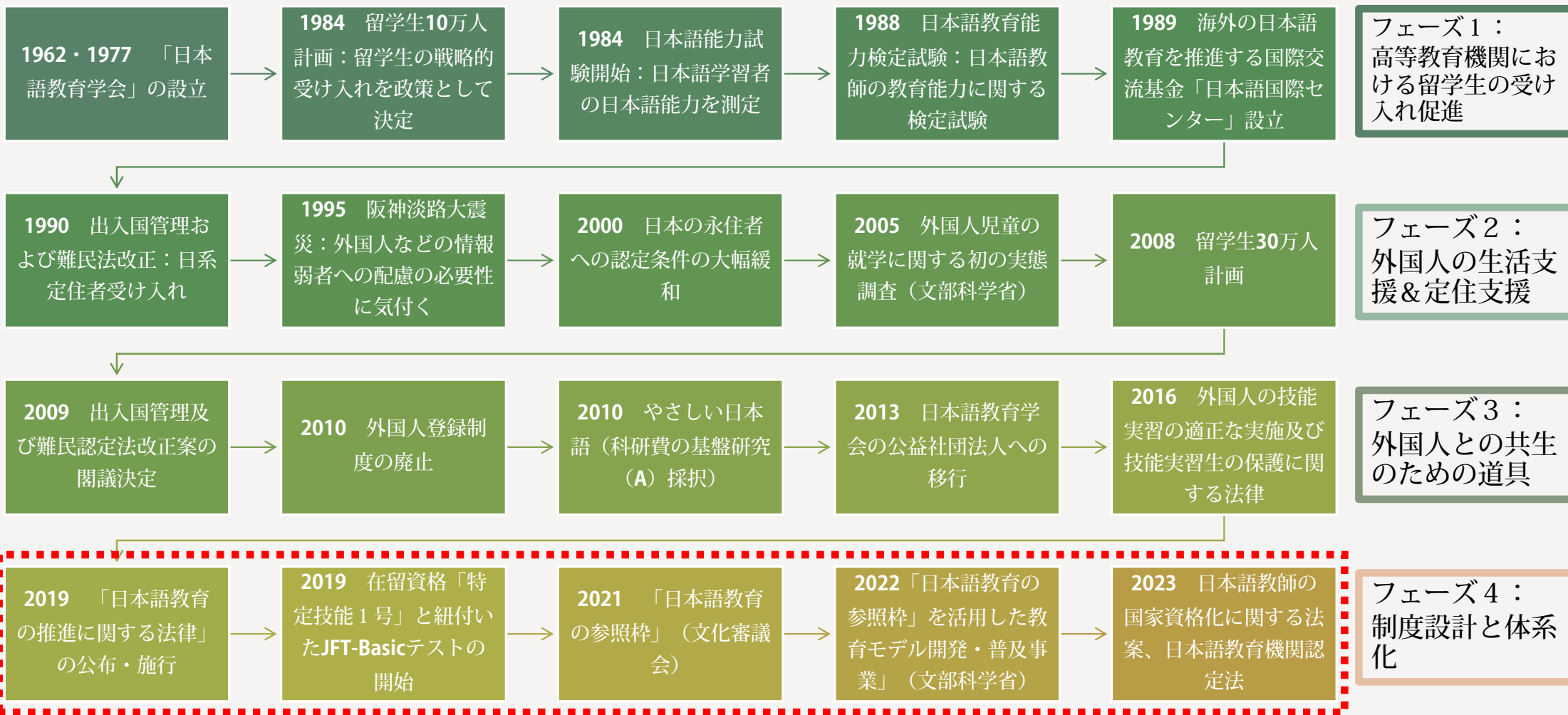
	心理測定学的パラダイム	オルタナティブ・アセスメントのパラダイム
学問的基盤	心理測定学	構成主義、状況論、解釈学など
評価目的	アカウンタビリティ、質保証	教育改善・指導、学生の成長
評価対象	集団	個人
評価機能	総括的評価	形成的評価
評価項目	分割可能性	複合性
評価場面	脱文脈性 統制された条件	文脈性 シミュレーション、真正の文脈
評価基準	客観性	間主観性
評価データ	量的データ	質的データ
評価主体	評価専門家、政策担当者	実践者自身
評価方法	標準テスト、学生調査 など	真正の評価、ポートフォリオ評価、パフォーマンス評価 など

松下佳代 (2012) 「パフォーマンス評価による学習の質の評価—学習評価の構図の分析にもとづいて」『京都大学高等教育研究』第18号, pp.75-114.

パフォーマンス評価が重要なもう1つの理由

- 日本国内の日本語教育の特殊性

1. 日本国内で日本語を学ぶ学習者は「生活者」でもある。
2. 日本語教育は、言語の教育としての特性に加え、共生のための道具としての特性を有する。
3. 社会の変化と連動してきた日本語教育の存在。
4. 日本語運用力をつける教育と、その能力の評価が重要になる。



日本語教育参照枠（日本語能力の熟達度について6レベルで示したもの）

日本語の学習・教授・評価を考える際に必要になる、日本語のレベルを示した全体的な尺度と、「聞く」「読む」「話す（やりとり・発表）」「書く」の言語活動別に実生活において日本語を使ってどんなことができるかを表した言語能力記述文（Can do）を示すもの。

- **日本語教育の推進に関する施策を総合的かつ効果的に推進するための基本的な方針**（令和2年6月23日 閣議決定）
「地域に在住する外国人が**自立した言語使用者**として生活していく上で必要となる日本語能力を身に付け、日本語で意思疎通を図り、生活できるよう支援する必要がある（p.9）」

全体的な尺度（抜粋）

言語使用者 熟達した	C2	聞いたり、読んだりしたほぼ全てのものを容易に理解することができる。自然に、流ちょうかつ正確に自己表現ができ、非常に複雑な状況でも細かい意味の違い、区別を表現できる。
	C1	いろいろな種類の高度な内容のかなり長いテキストを理解することができ、含意を把握できる。言葉を探しているという印象を与えずに、流ちょうに、また自然に自己表現ができる。社会的、学問的、職業上の目的に応じた、柔軟な、しかも効果的な言葉遣いができる。
言語使用者 自立した	B2	自分の専門分野の技術的な議論も含めて、具体的な話題でも抽象的な話題でも複雑なテキストの主要な内容を理解できる。お互いに緊張しないで熟達した日本語話者とやり取りができるくらい流ちょうかつ自然である。
	B1	仕事、学校、娯楽でふだん出会うような身近な話題について、共通語による話し方であれば、主要点を理解できる。身近で個人的にも関心のある話題について、単純な方法で結び付けられた、脈絡のあるテキストを作ることができる。
言語使用者 基礎段階の	A2	ごく基本的な個人情報や家族情報、買い物、近所、仕事など、直接的関係がある領域に関する、よく使われる文や表現が理解できる。簡単で日常的な範囲なら、身近で日常の事柄についての情報交換に応じることができる。
	A1	具体的な欲求を満足させるための、よく使われる日常的表現と基本的な言い回しは理解し、用いることもできる。もし、相手がゆっくり、はっきりと話して、助け船を出してくれるなら簡単なやり取りをすることができる。

・各レベルについての説明は、CEFR日本語版（追補版）の訳文を基にし、CEFR補遺版を参考に一部修正を加えた。

6

パフォーマンス評価の課題

1. 多くの時間と労力を要求する：課題の開発や実施が大変。採点基準の作成と採点は循環的であるため、多くの労力を要する。
2. 課題数が制限される：回答にも採点にも労力を要するため、問題数が限られ、カバー範囲も限られる。結果を一般化しにくい。
3. 信頼性が低い：採点が主観的になる。この課題を解決するため、
 - ルーブリックを使用し、採点基準の共有化
 - 複数の採点者間で事前調整
 - 採点者のトレーニング
 - 採点事例の蓄積

日本語教育の現状

- 解決すべき課題
 1. 大規模テストにおいて、パフォーマンス評価は取り入れられていない。
 2. 教育現場においても、評価の大変さから敬遠されがち。
- 改善のため

生成AIは、パフォーマンス評価を促進する原動力になる

日本語教育のための 言語テスト ガイドブック

李在鎬 編

[執筆]

安高紀子	Noriko Ataka
今井新悟	Shingo Imai
大隅敦子	Atsuko Osumi
小野塚若菜	Wakana Onozuka
加納千恵子	Chieko Kano
鎌田 修	Osamu Kamada
小林典子	Noriko Kobayashi
酒井たか子	Takako Sakai
島田めぐみ	Megumi Shimada
孫 媛	Yuan Sun
野口裕之	Hiroyuki Noguchi
村上京子	Kyoko Murakami
谷内美智子	Michiko Yachi
谷部弘子	Hiroko Yabe
李 在鎬	Jae-ho Lee



生成AI×日本語教育

調査課題

1. ChatGPTの言語理解力は？
2. 日本語教師の観点と生成AIの観点は合致する？
3. 日本語学習者の文章に見られる誤用を訂正できる？

生成AIは日本語読解問題にどのくらい解答できるか

—日本留学試験を対象として—

- 小野塚若菜(ベネッセ教育総合研究所)
- 岩崎拓也(筑波大学)
- 村田裕美子(ミュンヘン大学)
- 李在鎬(早稲田大学)
- 若井誠二(カーロリ・ガーシュパール・カルヴィン派大学)

人間の正答率の平均は70点前後

- 問題項目:2018年6月~2023年11月 **10回分計250問**
- 使用した生成AI:ChatGPT (GPT-4),Gemini (Gemini-1.0 Advanced) 2024年4月時点で最新モデル



結果概要

- 生成AIの正解率:
ChatGPT 96.8%, Gemini 91.2%
いずれかが正解した場合の正解率は98.0%
=人間の最高点(99.0%)に匹敵する成績(日本学生支援機構b)
- 生成AIの比較
全体的にChatGPTのほうがGeminiよりも正解率が高い(図1)。
ChatGPT【正解】—Gemini【不正解】 17問
Gemini【正解】—ChatGPT【不正解】 3問
いずれも【不正解】 5問
内容理解にかかる日本語の判別には違いがある？
- 不正解の問題項目の傾向
 - ①単純な情報取り出しでの誤り
 - ②接続詞の選択問題での誤り(特にGeminiの不正解は5/7問)
 - ③選択肢の設定によって正答に至れていない

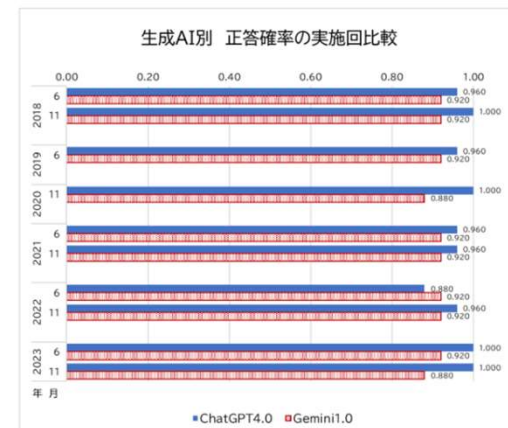


図1 生成AI別 正答率の実施回比較

日本語学習者が書いた意見文の例

マクロレベルから見れば、ある国の人口全体にとって住みやすい国の条件は最低限として人権、自由、民主主義、平和などが含めていると思える。住みやすい国には台風や地震のような自然災害が低く、あっても、保険や国家の助けのおかげで致命的ではない。安全、男女平等や高等教育などもあるし、生活水準が高い。汚職がなく景気がいいし、商品の価格や失業率が低く、仕事で米が食べられるし、自分の子供も将来があって安心できる。経済的な要素などの上に、国の文化生活の提供が多く、人々はその提供に参加する余裕があると住みやすいと思う。

このような条件が満たされていると否とにかかわらず、ミクロレベルで、つまり個人的な経験で、住みやすい国の定義が人によって非常に異なっていると思う。たとえば私が「法治国家」や「先進国」であるドイツに住んでいても、別に住みやすいと思わない。自分の生き方にあっている国はいくつかがあるかもしれないが、将来には日本に住むことしか想像できない。なんとなく、日本にいるうちにビーガンだから何回も外食で困ったことがあっても、仕事で死ぬほど忙しくても、窓がない四畳半の部屋で暮らしていても、他の辛いことがたくさんあっても、飽きていない。日本人の建前や満員電車で絶望するという人に対して、私にとって(これは「外人ボーナス」のおかげであることはさておき)日本人との交流はドイツ人とのよりずっと簡単に感じて、日本の社会や生活でホッとした。自分のやりたいことができる、結局、マイナス面があっても構わず

特別なことに対して熱心があれば、どんな条件でも満足して、
することができて、自分の生活に意味が見つけられる国、私にとって一番大住みやすい国だと思

李在鎬, 加藤恵梨, 堀恵子, 村田裕美子, 毛利貴美 (2023.12)
「ChatGPTの評価観点と人間の評価観点の比較 - 計量テキスト分析の手法を用いた分析 -」 第34回第二言語習得研究会 (JASLA) 全国大会 (予稿集 37-42)

成長

良い

A

- ・（構成）マクロとミクロに分けて、論じているところはとても良い。
- ・（内容）具体例を出して説明しているのでわかりやすい。
- ・（内容）個人の経験などを入れながら、意見を述べているところはオリジナリティがあって面白い。
- ・（日本語）文法や語彙の間違いが少ない。

悪い

- ・（結束性）接続表現などを使って、文と文をつなげるともっと良い。
- ・（構成）序論と結論がもうすこし明確にあると良い。
- ・（構成）最後のまとめは、ミクロレベルのまとめだけだと思うので、マクロレベルも含めた全体のまとめがあるともっと良い。

良い

B

明確な構造: 作文はマクロレベルとミクロレベルの視点から、「住みやすい国の条件と理由」を分析しています。この2つの視点を使うことで、議論のバランスが取れていて、理解しやすいです。

具体性と詳細性: 作文は具体的な例と詳細を提供しています。これにより、あなたの主張が実際の経験に基づいていることが明確になり、読者にとって説得力があります。

悪い

文法と表現: 一部の文は日本語の自然な表現から逸脱しています。例えば、「仕事で米が食べられるし」は「仕事により生計が立つし」、「ビーガンだから何回も外食で困ったことがあっても」は「ビーガンのため外食に困ることが度々あっても」などと修正するとより自然な日本語表現になります。

結論部の明確化: 結論部は全体をうまくまとめることが期待されます。最後のパラグラフは読者にとって異なる視点を提供する重要な要素ですが、そこでの主張が明確でなく、作文全体の結論としての役割を果たしていないように思います。

アカデミックライティングのスタイル: 作文は全体的に話し言葉のような表現が含まれています。アカデミックライティングでは、よりフォーマルな言い方と表現が求められます。例えば、「なんとなく、日本にいるうちにビーガンだから何回も外食で困ったことがあっても」は「私がビーガンであるため外食に困ることがしばしばあったにも関わらず、私は日本にいる間に」などと修正すると良いでしょう。

生成AIの誤用訂正に 日本語教師は どの程度納得するのか

李 在鎬（早稲田大学）

岩崎 拓也（筑波大学）

村田 裕美子（ミュンヘン大学）

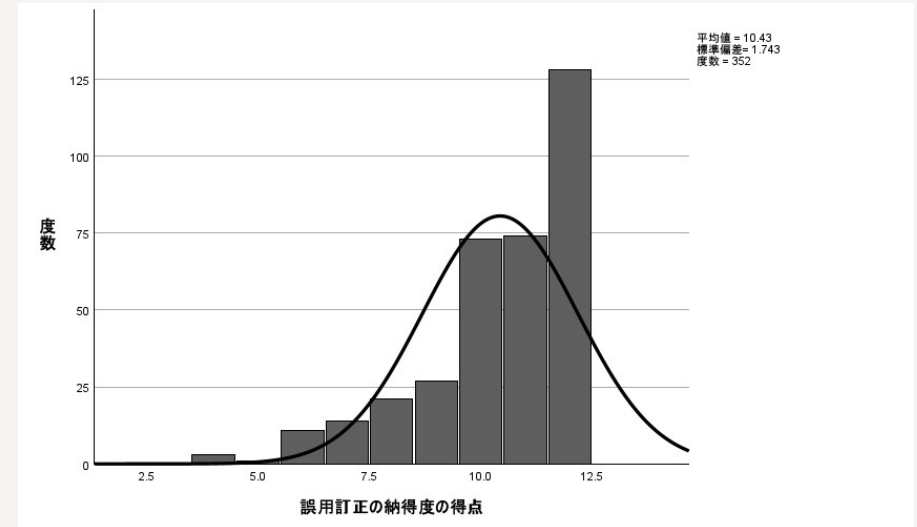
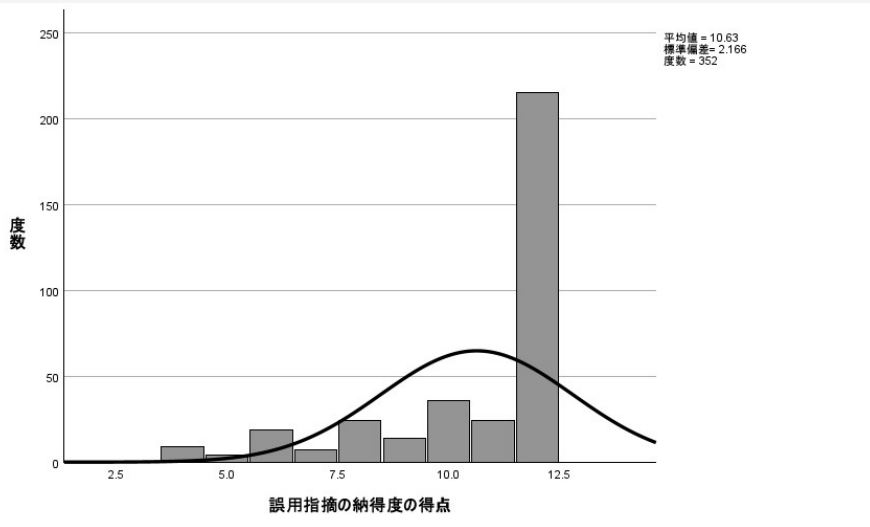
SRDANOVIC Irena（ブーラ大学）

共同研究者 Kristina Hmeljak Sangawa（リュブリャナ大学）

科研費（24K00078）の成果発表です

研究概要

- **調査**：学習者作文コーパス 「住みやすい国コーパス」（*村田2021）をChatGPT-4で分析。ChatGPTに誤用例を指摘してもらうタスクを実行。ChatGPTの指摘を日本語教師4人がチェック
- **結果と示唆**：352件のサンプルを調査。日本語教師は78%の事例において納得。



納得できない事例：AIの直しすぎ問題

- 4名の日本語教師全員が納得できないと回答した「9つ」の事例

生成AIが誤用として指摘した事例	生成AIが訂正案として提案した表現
水を飲むことができます	水が飲めます
6年前にクロアチアに来ました	6年前にクロアチアに来ました
ほかの国と比べると	他の国と比べて
楽しくて、温かくて、気さくな国です	楽しく、温かく、気さくな国です
大切だと思えます	大切であると思えます
言論の自由もとても大事なことです	言論の自由もとても大事です
外国人を受け入れます	外国人を受け入れる
戦争がなく、貧困も割と少なくて	戦争がなく、貧困も比較的少ない
便利なインフラ設備も条件の一つです。	便利なインフラも重要な条件です。

アカデミックリーディングにおける教育実践へ

Step0

文章を読みます。

Step1

読んだ文章に関連するテーマで意見文を書いてもらいます。

Step2

生成AIがフィードバックします。

Step3

学習者がフィードバックをもとに意見文を書き直します。

Step4

人間教師が確認し、フィードバックします。

アカデミックリーディング3～4レベル（初級後半から中級前半）

作文のテーマ：イルカたちは、「超音波」を使って、コミュニケーションをします。一方、私たち人間は、「ことば」を使って、コミュニケーションをします。この2つ（「超音波」と「ことば」）には、どのような共通点（同じところ）と相違点（違うところ）がありますか。また、「超音波」は「ことば」に比べ、どんなところが優れていますか。あなたの考えを書いてください（300字～400字）。

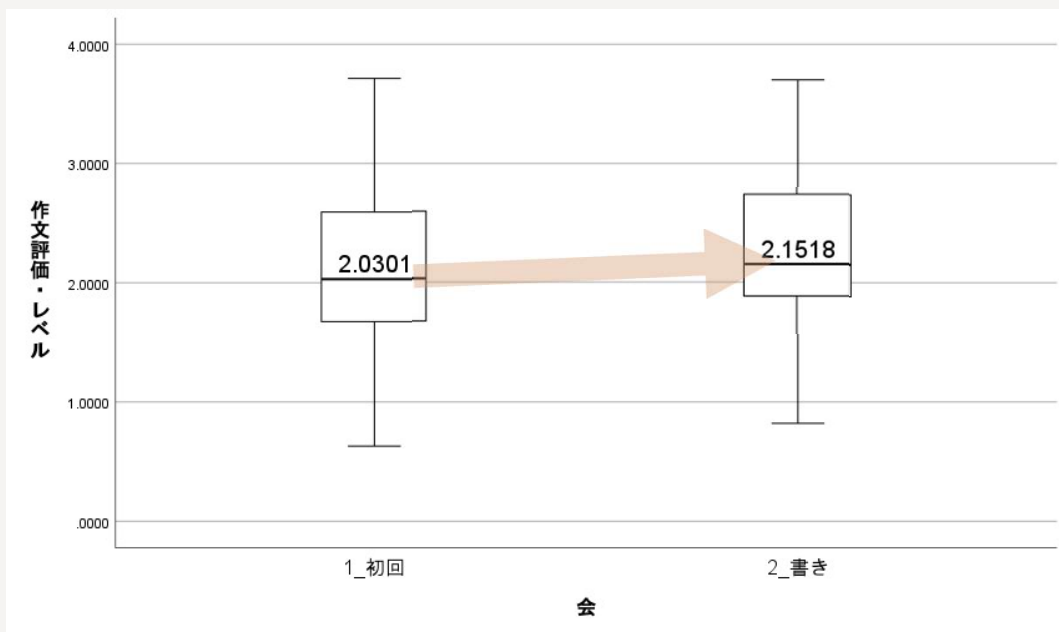
ポイント1：2～3つの段落で書いてください。

ポイント2：段落の中心文を意識しましょう。中心文を段落の最初に書いてください。

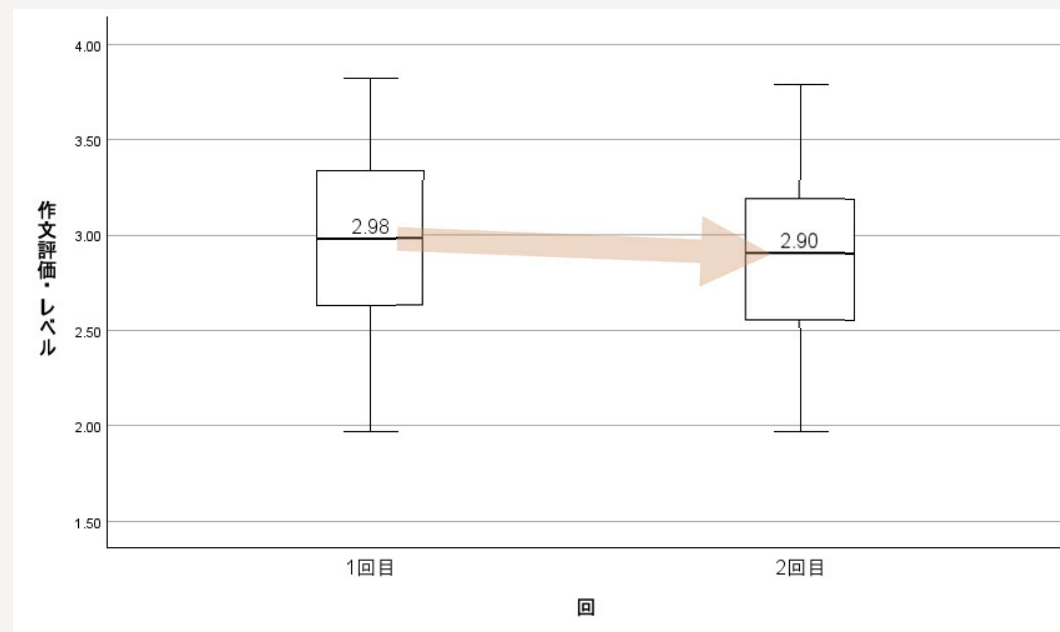
ポイント3：中心文のあとに支持文を書きます。

AIのフィードバックで文章はよくなるか

2024春学期



2024秋学期



まとめ

1. パフォーマンス評価の教育的価値は高く、とりわけ国内で日本語を学ぶ人々（学習者 & 生活者）にとっては不可欠。
2. 運用向上を目指す教育と評価において、生成A Iは教師の強いミカタ。生成A Iは十分な日本語力を持ち、学習者の文章に見られる誤用の指摘精度は教師からみても納得できるもの。「教育のツール」として有望
3. ただし、「学習のツール」としての可否については未知。継続的な調査が必要。

ありがとうございました。

