生成 AI を有能な赤ペン先生にするプロジェクト

李在鎬（早稲田大学）
jhlee.n@gmail.com

【要約】
筆者の研究グループでは、2024 年 4 月から科研費の補助を得て生成 AI を組み込んだ作文診断システムを開発している。以下では、2024 年度に行った３つのケーススタディ（①生成 AI の文章理解力、②生成 AI の誤用抽出の単位、③生成 AI の誤用の指摘に対する日本語教師の納得度）に関する調査結果を報告する。その調査結果から、本研究グループが目指す生成 AI を組み込んだシステム開発の意義は非常に大きいことを示す。

１．背景

筆者はこれまで、大規模データの定量分析を通じて、学習に関わる現象をモデル化し、日本語教育および学習を支援するシステムを開発・公開してきた。具体的には、2013 年に日本語文章難易度判定システム「jReadability」（https://jreadability.net/）（李（編）2017）を開発・公開し、2017 年には日本語学習者作文評価システム「jWriter」（https://jreadability.net/jwriter/）（李（編）2019）を公開した。このような取り組みの延長として、2024 年からは科研費の助成を受け、「jWriter」に生成 AI を組み込み、日本語学習者が入力した作文に対して診断的評価を提示するシステムの開発を目指して研究を進めている。本稿では、この研究プロジェクトの目的を概説した後、初年度に行った生成 AI の定量分析に関する 3 つのケーススタディ（李 2024b、李ほか 2024、小野塚ほか 2024）について報告する。

２．大規模言語モデルのインパクトと科研費の研究が目指すもの

大規模言語モデルを実装した生成 AI は、標準的な言語使用者を超える正確さと流暢さを獲得している。例えば ChatGPT は高品質な日本語文の生成能力があることが定量研究（樽本ほか 2024）によって報告されており、こうした特徴を生かし、言語教育に活用しようとする動きが活発になっている（Song & Song 2023、Allen & Mizumoto 2024）。こうした動きは英語教育において特に顕著であり、生成 AI を利用した言語学習支援システムの開発が盛んに行われている。特にライティング教育支援においては、注目すべき研究成果が上がっており、生成 AI に文法や語彙の誤りを指摘させることで学習者のライティングスキルが向上したことが報告されている（Song & Song 2023）。また、ChatGPT は大規模言語モデルを実装しているため、多様な言語タスクに対応できることが知られており、学習者は様々な種類の文章作成の練習に活用することができる（Kohnke, Moorhouse & Zou 2023）。さらに、自動採点（Mizumoto & Eguchi 2023）や文法エラーの修正（Mizumoto et al. 2024）に関しても高い性能を示していることが報告されている。

日本語教育においては、教材開発のための研究（久野ほか 2023）をはじめ、自動採点に関する研究（李 2023、長谷川 2024）、ChatGPT を語彙資源構築のツールとして用いることを提案する研究（李 2024c）、文章平易化ツールとして活用する研究（李・長谷部 2025）などがなされており、生成 AI を活用するこ

とは教師にとっても学習者にとっても大きなメリットがあるとされている。特に李ほか (2023) では、ChatGPT のフィードバックと教師のフィードバックを比較し、両者に高い類似性が見られることを指摘している。これらの研究成果を受け、李 (2024a) では採点ツールとしての生成 AI の積極的な活用を提案しており、生成 AI の導入を機に外国語教育の方向転換についても検討すべきと主張している。具体的には、次の 3 点を提案している。1)「目的」としての外国語学習から「手段」としての外国語学習への転換、2) 知識中心の教育・評価からパフォーマンス中心の教育・評価への転換、3) 単言語中心の外国語学習から複言語能力の価値を受容する外国語学習への転換の必要性を指摘している。

以上の流れを加速化すべく、筆者の研究グループでは 2024 年の 4 月から生成 AI を組み込んだ日本語作文支援システムの開発に関する研究を行っている。具体的には、科研費基盤研究 (B)「生成 AI を組み込んだ日本語作文診断システムの開発と普及に関する研究」というプロジェクトである。その詳細は、https://kaken.nii.ac.jp/grant/KAKENHI-PROJECT-24K00078 で確認できるが、2024 年度は研究の 1 年目に相当する。1 年目の研究では、生成 AI の定量評価を目的に以下の 3 つの調査研究を行っている。

調査研究１：生成 AI の言語理解力を調べる
調査研究２：生成 AI の誤用訂正力を調べる
調査研究３：生成 AI の文章の内容理解力を調べる

調査研究１では、言語テストの問題バンクを作成し、生成 AI に回答させることで言語理解力を調査する。調査研究２では、生成 AI が学習者の誤用事例に対して、どのような指摘をするか、そして、生成 AI の誤用訂正に対して日本語教師はどう評価するかを調べる。調査研究３では、生成 AI が学習者の文章に含まれる内容にかかわる要素（主張、根拠、論理構成など）に対して、どの程度、適切な指摘ができるかを調査する。調査１に関しては、小野塚ほか (2024) で結果を公開しているし、調査２は李 (2024b) および李ほか (2024) で報告しているが、以下でその概要を報告する。

３．生成 AI の言語理解力

小野塚ほか (2024) の研究では、2018 年 6 月から 2023 年 11 月にかけて実施された日本留学試験の日本語科目、読解領域における 250 の問題をデータベース化し、これを ChatGPT (GPT-4) および Gemini (Gemini-1.0 Advanced) が解答する実験を行った。両システムの平均正解率は、ChatGPT が 96.8%、Gemini が 91.2%であった。どちらか一方が正解する確率は 98.0%に達し、これは人間の最高得点(99.0%)と比較しても遜色のない結果を示している。

小野塚ほか(2024)の研究において ChatGPT が Gemini より正解率が高いことを明らかにしている (図 1)。

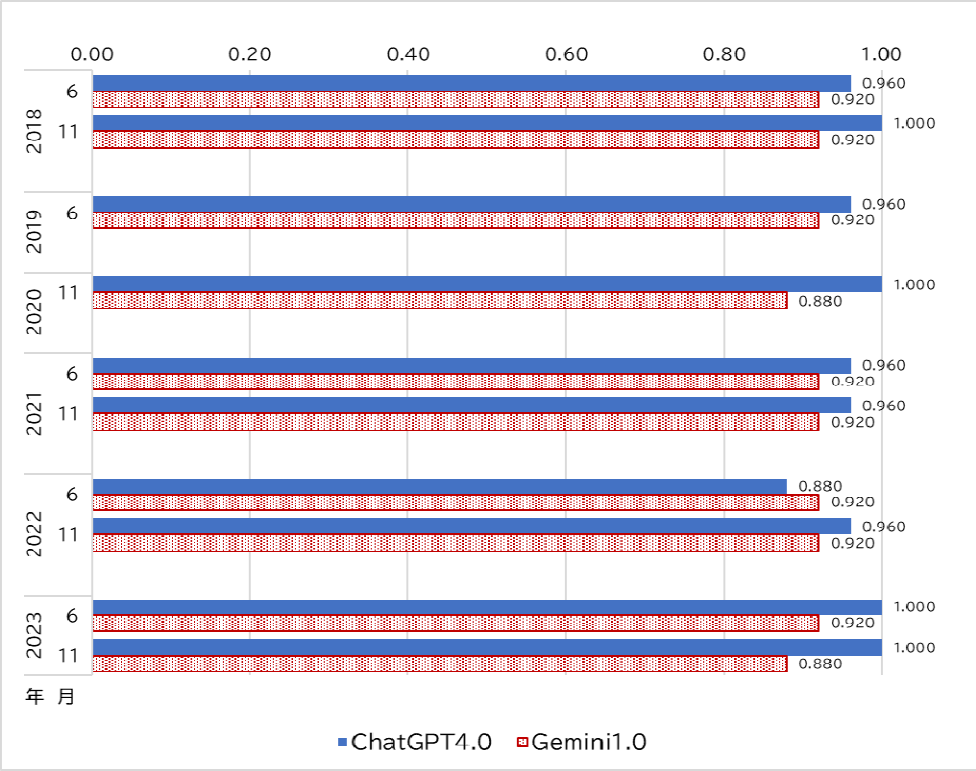

図 1　生成 AI 別　正解率の実施回比較

さらに、性能の詳細について分析すると、Gemini では不正解だったが、ChatGPT は正解できた問題が 17 問、ChatGPT が不正解だったが、Gemini は正解できた問題が 3 問であった。また、両システムが共に不正解だった問題は 5 問である。そして、生成 AI が誤答する問題の傾向に関して、以下の 3 つの点を報告している。

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
3. 選択肢の設定によって正答に至れていない

1 は単純なミスのようであるが、原因は不明であり、今後、改善されることが期待される。一方、2 の接続詞の選択問題と 3 の選択肢の選択問題についてはいくつか興味深い事実が観察された。まず、2 の接続詞を選択する問題に関しては、ChatGPT よりは Gemini において誤答が目立った。具体的には、すべての問題において接続詞に関わる問題は 7 問あったが、Gemini は 5 問（ChatGPT は 1 問）において不正解となった。

次に、3 の「選択肢の設定によって正答に至れていない」ケースの具体例として 3 つのタイプが観察された。

① 上位概念と下位概念の関係が判別できていないタイプ
② 選択肢の構造が異なることが原因のタイプ
③ 選択肢の時制の不一致が原因のタイプ

著作権の関係で実際の試験問題を掲載することはできないが、①～③はそのいずれも高度な推論能力が求められるタスクであり、今後、提案されうる進化版生成 AI を評価する一つの指標になると考えている。なお、小野塚ほか（2024）の調査結果は、2024 年春の時点で最新モデルあった GPT-4 に基づくものであるが、その後の GPT-o1 モデルでは改善されていることも確認できている。とりわけ上記の①に関しては顕著な改善が確認できている。

４．生成 AI の誤用訂正力

生成 AI の誤用訂正に関する定量分析として、李（2024b）および李ほか（2024）がある。まず、李（2024b）では、ChatGPT の GPT-3.5 と GPT-4 のモデルを使い、日本語学習者の作文に含まれる誤用を抽出するタスクを行った。調査データとして「住みやすい国コーパス」（ver.1）（村田 2021）に含まれている 100 名分の作文データを使用した。各作文に対して不自然な日本語を「」で括るように指示し、誤用を抽出した結果、2167 例（初級 496、中級 1268、上級 403）の誤用例が収集できた．なお、GPT-3.5 が指摘できた事例は 761 件、GPT-4 は 1406 件であった．実際の事例を示す．「」で括られている部分が誤用である。

> まず、政治的な状態の基本的なレベルです。「私たちはその質問を主民主義の文脈から見るため」、生活しやすい国も民主的人権がある国だろう。その上に、表現の自由や出版の自由などが守られていることもその条件の二つではないでしょう。
> 自分がどんな国で社会化を受けたことによって、住みやすい国の理想が異なるかもしれません。それにしても、文化や社会の寛容が非常に大切だと言えます。「自分自身が外国からその国に入ってきて、どうやって向かわれているか」、あるいは自分自身がその国で育って、異文化の人々に対してどんな思想があるかによって、生活の質が変わります。「詳しく説明ができるように、例を挙げたいと思います。例えば」、長年日本に住んでいる外国人によると、生活や仕事の条件が順調ですが、外国人として、本当に社会の一員はなれないことだそうです。「人間の幸福のためには帰属意識が重大なので」、それは結局問題になる可能性がある。「その一方」、ドイツでは、特にトルコ人や難民に対して偏見が多く、共存が複雑になってしまいます。「特に異文化の経験があまりない人々は早めに排外思想を持つようになって」、「理由なく差別します」。両方は国民と外国人のために生活しやすい国を作りません。

李（2024b）の成果として、図 2 が挙げられる。

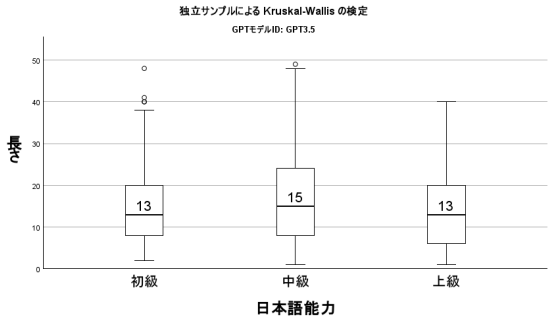

図 2a：GPT-3.5 の誤用指摘の長さ（p=.035）

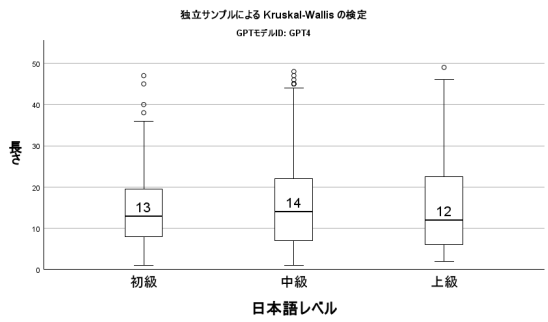

図 2b：GPT-4 の誤用指摘の長さ（p=.254）

図 2 は ChatGPT の 2 つのモデルで誤用として指摘した部分の文字数を数え、日本語レベル別に箱

ひげ図を書いている。ここで注目すべきは、GPT のモデルや日本語能力の違いによる顕著な差はなく、12 文字から 15 文字の長さに対して誤用を指摘していることである。「寺村誤用例集データベース」（https://www2.ninjal.ac.jp/teramuradb/）をはじめとする従来の誤用分析の場合、語彙論、品詞論、統語論（助詞類）のレベルで誤用を捉えることが主流であり、いわゆる「語の単位」で整理するアプローチが主流であると言える．しかし、生成 AI の指摘は、その長さから語や形態素とは言えない単位で指摘をしている。12～15 文字の長さは文として捉えるには短く、語として捉えるには長いため、「表現の単位」とも言える独自の単位で誤用を捉えている可能性が示唆される。

李（2024b）が明らかにしたような生成 AI の独自性は、日本語教育の当事者にとって、どのように捉えられるのであろうか。こうした問題意識のもとで、李ほか（2024）では生成 AI の誤用訂正に対して人間の日本語教師はどの程度、合意できるかを調べた。具体的には、「住みやすい国コーパス」からサンプリングした 20 編（下位群 6 編、中位群 7 編、上位群 7 編）の作文を Open AI の GPT-4 モデルで誤用抽出と訂正タスクを行った後、そのサンプルに対して、日本語教師はどの程度納得できるかをアンケート調査した。分析では 20 編の作文から抽出した 352 例のサンプル（上位群 112 例、中位群 140 例、下位群 100 例）に対して、4 名の日本語教師が 2 つの観点から評定を行った。

観点 1）生成 AI が指摘した誤用に対してあなたも誤用だと思うか。
観点 2）生成 AI が訂正した表現に対してあなたは納得するか。

観点 1）は誤用指摘の納得度を調べるためのもので、観点 2）は誤用に対する訂正表現の納得度を調べるためのものである。評定では「納得する、納得しない、どちらとも言えない」のいずれかを選んでもらったあと、筆者がスコア化（納得しない→[1]、どちらとも言えない→[2]、納得する→[3]）を行った。調査の結果、観点 1）に関しては図 3、観点 2）に関しては図 4 の結果が得られた。

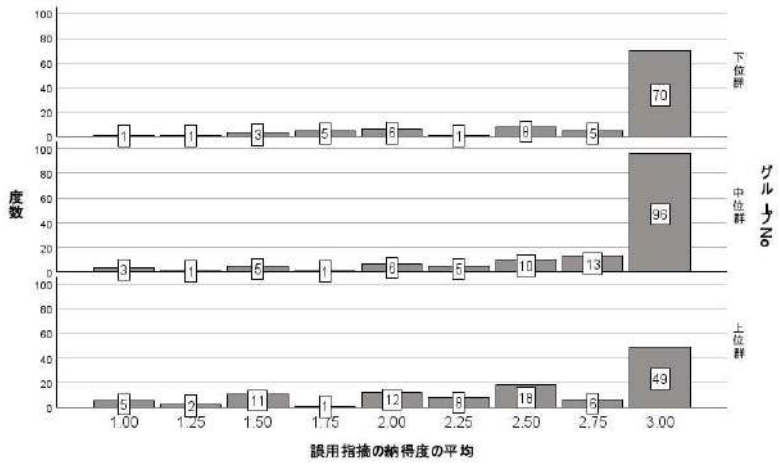

図 3. 誤用指摘の納得度の平均値のヒストグラム

図 3 において注目すべきは次の点である。4 名が 3 点をつけた事例がいずれのレベルにおいてももっとも多く、全体の 6 割を占めている。実質的な納得と捉えられる 2.5 点以上を含めた場合、3 つの群の平均として 78.1%（下位群 83%、中位群 85%、上位群 65.3%）の指摘に対して納得していると解釈できる。

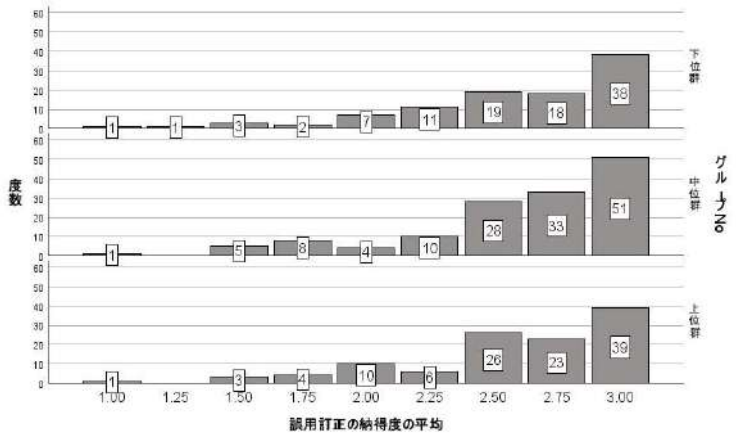

図 4. 誤用訂正の納得度の平均値のヒストグラム

図 4 では、誤用を修正した表現（訂正表現）の納得度を示している。図 3 と同様に 4 名が 3 点をつけた事例がいずれのレベルにおいてももっとも多く、全体の 4 割を占めている。実質的な納得と捉えられる 2.5 点以上を含めた場合、78.1%（下位群 75%、中位群 80%、上位群 78.6%）の指摘に対して納得していることが分かる。

李ほか（2024）によって、生成 AI が指摘する誤用および訂正した表現の多くは、日本語教師にとっても納得できるものであることが明らかになった。しかし、課題として生成 AI が過度に誤用を指摘していることも明らかになった。

表 1 ChatGPT による過度な誤用判定の事例

No	ChatGPT による誤用判定	ChatGPT による訂正
1	水を飲むことができます	水が飲めます
2	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
3	ほかの国と比べると	他の国と比べて
4	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
5	大切だと思います	大切であると思います
6	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
7	外国人を受け入れます	外国人を受け入れる
8	戦争がなくて、貧困も割と少なくて	戦争がなく、貧困も比較的少ない
9	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。

表 1 の左側は、ChatGPT が誤用として指摘した学習者の文例であり、右側が誤用に対する訂正表現の例である。しかし、4 名の日本語教師全員がこれらを誤用とすることに対しては納得できないと回答している。これらは、ChatGPT による「直しすぎ問題」として捉えることができる。表 1 には、表記に関わることが 2 件（No.2、3）、文法に関わることが 4 件（No.1、4、6、8）、文体に関わることが 2 件（No.5、7）、語彙に関わることが 1 件（No.9）となり、必ずしも一貫した傾向は確認できなかった。従って、現時点においては生成 AI の「直しすぎ問題」については、ユーザーレベルで注意すること以外に有効な対策がないと言えよう。

５．まとめ

本稿では、生成 AI の定量分析に関する 3 つのケーススタディを報告した。1 つ目のケーススタディは、生成 AI の言語理解力を確認するため、日本留学試験の試験問題を解答させる調査であったが、結果としては人間の最高点（99.0%）に匹敵する得点（98.0%）であったことが明らかになった。2 つ目のケーススタディとして、生成 AI が指摘する誤用の単位を調べた結果、生成 AI はこれまで人間の分析者が考えている形態素や語や文といった単位とは異なる単位（表現の単位）で誤用を捉えていることが明らかになった。3 つ目のケーススタディとして、生成 AI の誤用の指摘に対して人間の日本語教師はどの程度、納得できるかを調べた結果、78.1%の事例に対して納得できることが明らかになった。これらのケーススタディから、生成 AI の高い言語能力が明らかになったため、本研究グループが目指す生成 AI を組み込んだシステム開発の意義も大きいことが示唆される。

[謝辞] 本研究は、科研費（24K00078）の研究成果である。また、共同研究者の小野塚若菜(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)、若井誠二(カーロリ・ガーシュパール・カルヴィン派大学)、Srdanović Irena（プーラ大学） 、Kristina Hmeljak Sangawa（リュブリャナ大学）に感謝する。

参考文献

小野塚若菜・岩崎拓也・村田裕美子・李在鎬・若井誠二（2024）「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」（日本語教育学会 2024 年度秋大会口頭発表）.

久野かおる・波村慎太郎・津坂朋宏（2023）「文章教材の作成に生成 AI を活用する試み―語彙学習の効果を高める文章を読む練習―」『日本語教育方法研究会誌』30、pp.8-9. https://doi.org/10.19022/jlem.30.1_8

樽本空宙・畠垣光希・宮田莉奈・梶原智之・二宮 崇（2024）「ChatGPT の日本語生成能力の評価」『自然言語処理』 31(2)、pp.349-373. https://doi.org/10.5715/jnlp.31.349.

長谷川由香（2024）「ChatGPT による日本語会話評価の可能性」（第 63 回日本語教育方法研究会ポスター発表）.

村田裕美子（2021）「小規模コーパスの構築方法」、李在鎬（編）『データ科学×日本語教育』ひつじ書房、pp.34-53.

李在鎬（編）（2017）『文章を科学する』ひつじ書房.

李在鎬（編）（2019）『ICT×日本語教育』ひつじ書房.

李在鎬（2023）「ChatGPT による日本語作文の自動採点」（日本語教育学会 2024 年度秋大会口頭発表）.

李在鎬・加藤恵梨・堀恵子・村田裕美子・毛利貴美（2023）「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 口頭発表）.

李在鎬（2024a）「ChatGPT のインパクトと今後の外国語教育の方向性について」『大学英語教育学会中国・四国支部研究紀要』21、pp.31-44. https://researchmap.jp/jhlee/published_papers/45984219

李在鎬（2024b）「生成 AI の誤用訂正に対する定量分析：生成 AI は日本語学習者作文をどう捉えているのか」（計量国語学会第 68 回大会口頭発表）.

李在鎬（2024c）「ChatGPT による意味分類の予測について―生成 AI は語の抽象的な属性が理解できるか―」『計量国語学』34(6)、pp.432-442.

李在鎬・岩崎拓也・村田裕美子・SRDANOVIC Irena・Kristina Hmeljak Sangawa（2024）「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：生成 AI を有能な添削先生に育てるための取り組み」（日本語教育学会 2024 年度秋大会口頭発表）.

李在鎬・長谷部陽一郎（2025）「ChatGPT による日本語ニュースの平易化：生成 AI と「やさしい日本語」『計 量国語学』34(8)、pp.563-573

Allen, T. J., & Mizumoto, A. (2024). ChatGPT over my friends: Japanese English-as-a-foreign-language learners' preferences for editing and proofreading strategies. *RELC Journal*. https://doi.org/10.1177/00336882241262533

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 537-550. https://doi.org/10.1177/00336882231162868

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Mizumoto, A., Shintani, N., Sasaki, M., & Teng, M. F. (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. *Research Methods in Applied Linguistics, 3*(2), 100116. https://doi.org/10.1016/j.rmal.2024.100116

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology, 14*, 1260843. https://doi.org/10.3389/fpsyg.2023.1260843