Journal of Interdisciplinary Cultural Studies
JICS 2026, 1: 53-65
DOI: 10.29329/almamater.2026.1417.4

Evaluating the Error Analysis Performance of the Deepseek AI Chatbot in Foreign Language Teaching

Muhammed Mustafa Uçar[1]*, Muhammet Koçak*[2]

Abstract

This study evaluates the error analysis performance of the DeepSeek artificial intelligence chatbot in foreign language teaching, focusing on its ability to detect, correct, and provide feedback on errors in German texts at the B1 proficiency level. Utilizing a structured methodology, 11 texts from the Menschen B1 textbook were modified to include 88 intentional errors, categorized into four types (omission, addition, selection, ordering) across grammatical and lexico-semantic levels. The chatbot's performance was assessed through three stages: error detection, correction accuracy, and feedback quality. Results revealed that DeepSeek detected 85% of errors overall, with higher success rates for addition (91%) and ordering (91%) errors compared to omission (77%). Correction accuracy stood at 91%, with all selection errors corrected flawlessly, though lexico-semantic ordering errors showed lower correction accuracy (70%). Feedback for correctly corrected errors was 86% accurate, with lexico-semantic errors receiving near-perfect feedback (96%) compared to grammatical errors (75%). While DeepSeek demonstrated robust capabilities in error detection and correction, inconsistencies were observed, particularly in addressing omission errors and providing feedback for grammatical ordering errors. The findings highlight DeepSeek's potential as a supplementary tool in foreign language education but underscore the need for refinement in handling specific error types and feedback clarity. This study contributes to the limited literature on DeepSeek's educational applications and offers insights for optimizing artificial intelligence driven error analysis in language learning contexts.

Keywords: *DeepSeek, Error analysis, Foreign language teaching, Chatbot, Artificial intelligence*

Introduction

The methods used in foreign language education naturally evolve with technological advancements. It would not be incorrect to say that artificial intelligence has now become part of the process that began with the Grammar-Translation Method. Today, artificial intelligence is increasingly expanding its presence in almost all fields. Consequently, it has already begun to establish itself in foreign language teaching. Numerous scientific studies have been conducted in this field, and research continues to grow (Haristiani, 2019; Hong, 2023; Mohamed, 2024; Pokrivcakova, 2019). The integration of artificial intelligence into foreign

[1] Dr. Karamanoglu Mehmetbey Üniversitesi, Karaman, Türkiye, mucar@kmu.edu.tr, ORCID: 0000-0002-8841-0367.
[2] Prof. Dr. Gazi Üniversitesi, Ankara, Türkiye, muhammetkocak@gazi.edu.tr, ORCID: 0000-0001-6387-0765.

How to cite this article: Uçar, M. M., Koçak, M. (2026). Evaluating the Error Analysis Performance of the Deepseek AI Chatbot in Foreign Language Teaching. In: *Alma Mater Journal of Interdisciplinary Cultural Studies 3:1*, 53-65.

Submitted: 25.09.2025 & **Accepted:** 07.11.2025

language education is an inevitable development. In the not-so-distant future, an education model, method, or technique without artificial intelligence will likely be unimaginable. The development of different methods and techniques in foreign language education due to the widespread use of the internet and technology serves as an example of this progress.

Although various methods and techniques have been developed in foreign language education, their sole purpose is to enable students to learn the target language more effectively. One common aspect of foreign language classes is teacher feedback. Correcting students' errors is one of the primary responsibilities of a teacher; however, providing feedback is even more crucial than correction itself. This process, referred to as therapy, holds a significant place in foreign language education (Harden, 2006; Nickel, 1972).

Before defining the concept of Error Analysis, it is necessary to distinguish between error and mistake. This distinction has frequently been made in the literature; however, Brown explains this issue quite clearly:

Mistakes are performance lapses (e.g., slips of the tongue, hesitations) that occur randomly in both native and second language speakers. These are not due to a lack of competence but temporary breakdowns and can be self-corrected. In contrast, errors reflect a learner's systematic deviation from native-speaker norms, revealing gaps in their current competence. For example, a learner asking "Does John can sing?" demonstrates an error rooted in their internalized rule system (e.g., overgeneralizing do-auxiliary rules). Unlike mistakes, errors persist until the learner's linguistic system evolves (2007).

In foreign language teaching, the focus is naturally on errors. The concept of error has been defined similarly. A general definition would be as follows: an error is a deviation from the norms of the target language (Bohnensteffen, 2010; Dulay, Burt and Krashen, 1982; Ellis, 1994).

Erdoğan describes the emergence of Error Analysis as follows: Prior to the late 1960s, behavioristic theory dominated second language acquisition, framing learning as habit formation and attributing errors to native language interference. This led to contrastive analysis—comparing languages to predict errors—while overlooking non-transfer-related mistakes. Error analysis emerged as a corrective, positing that errors stem not only from native language influence but also universal learning strategies and cognitive processes used to process target language input. Unlike contrastive analysis, it treats errors as evidence of learners' evolving competence, offering insights into second language acquisition mechanisms and informing pedagogical approaches (2005).

According to Subekti, Error Analysis provides numerous benefits for teachers, researchers, and students (2018). Dulay, Burt, and Krashen explain this as follows:

> Studying learners' errors serves two major purposes: (1) it provides data from which inferences about the nature of the language learning process can be made; and (2) it indicates to teachers and curriculum developers which part of the target language students have most difficulty producing correctly and which error types detract most from a learner's ability to communicate effectively (1982).

According to Abbot, "[t]he aim of any EA [Error Analysis] is to provide a 'psychological explanation' — a reliable account of the causes of the errors" (1980). This is because an error can only be systematically corrected once its cause is understood (Corder, 1982). A review of the literature in this field reveals that the causes of errors have been categorized in detail.

However, the errors included in the experimental section of this study were deliberately made by the authors. Detailed information is provided in the methodology section.

As explained by Arslan, the human-like abilities of a computer—such as logical reasoning, problem-solving, inference, and the generalization of behaviors, in other words, the use of high-level cognitive skills—can be defined as artificial intelligence (2020). Artificial intelligence is the ability of a machine to imitate the complex processes of the human mind in order to reach a judgment and exhibit human-like behavior. It would be appropriate to provide a different definition for a chatbot: A chatbot (also known as smart bot, interactive agent, or digital assistant) is a computer program designed to simulate conversation with human users, typically over the Internet. It interacts through text or voice and uses Natural Language Processing to understand and respond to human language (Adamopoulou and Moussiades, 2020).

By the end of 2024, many artificial intelligence chatbots have emerged and are in use. ChatGPT, Gemini, and Copilot are chatbots that are widely known and used by almost everyone today. However, the chatbot released by DeepSeek on January 10, 2025, with the same name, has managed to shake things up not only in the field of artificial intelligence but also in many other areas. The reason for DeepSeek's significant impact is its lower cost and faster development compared to other chatbots. Despite this, it is said to be capable of competing with other chatbots and even surpassing them in some areas.

It has been previously stated that many scientific studies have been conducted on artificial intelligence chatbots, especially ChatGPT. However, there appears to be very little research on DeepSeek in the literature. Considering that it has been less than 2 months since its release, this is quite understandable.

This study lies at the intersection of Error Analysis and Artificial Intelligence Chatbots in Foreign Language Teaching. The study aims to evaluate the ability of the artificial intelligence chatbot DeepSeek to detect errors, correct them, and provide feedback in written products. Therefore, the research questions are as follows:

- How many of the errors in written products can the artificial intelligence chatbot DeepSeek detect?
- How many of the errors in written products can the artificial intelligence chatbot DeepSeek correctly correct?
- How much correct feedback (i.e., therapy) can the artificial intelligence chatbot DeepSeek provide regarding the errors in written products?

The hypothesis of the study is as follows: The artificial intelligence chatbot DeepSeek is successful in error analysis in foreign language teaching. It detects all errors, corrects them, and provides appropriate feedback. The limited number of studies on the use of DeepSeek in foreign language education increases the importance of this study.

Method

In this study, Corder's suggestion regarding Error Analysis will be utilized. Corder suggests the following steps in Error Analysis research:

1. Collecting a sample from the student language
2. Identifying the errors
3. Describing the errors
4. Explaining the errors
5. Evaluating the errors (as cited in Ellis, 1994).

In the first stage, Corder suggests collecting examples from students; however, in this study, examples will not be collected from students. The data used in this study will be taken from the Menschen B1 textbook, which is designed for teaching German as a foreign language at the B1 level, and will be modified to contain errors. The reason for not collecting data from students is to ensure equality in the error classification, which will be addressed in later sections of the study (Table 1). Although this may seem artificial, it is important for the validity and reliability of the study. Had data been collected from students, there would have been an imbalance in the types of errors, and it would not have been possible to obtain equal amounts of data for each error level and type. Therefore, texts containing errors will be used in such a way that they are equal in each level and type.

In the second stage, Corder suggests identifying the errors. This task will be carried out by DeepSeek.

In the third and fourth stages, the errors must be described and explained. Errors are generally categorized as errors of omission, which refer to the absence of a necessary element in a sentence; errors of addition, which refer to the presence of an unnecessary element in a sentence; errors of selection, which mean the preference of an incorrect element instead of the correct one; and errors of ordering, which occur when elements are correct but presented in the wrong order. This classification is only a starting point for a more systematic analysis and primarily provides data or evidence. These four types of errors can occur at the Graphological/Phonological, Grammatical, and Lexico-semantic levels (Corder, 1982). Since there is no data at the Graphological/Phonological level in our study, the focus will be on the errors at the Grammatical and Lexico-semantic levels. Based on this, the classification of the errors in this study is as follows:

Error type (ET) Error level (EL)	Omission (Om.)	Addition (Ad.)	Selection (Se.)	Ordering (Or.)
Grammatical (Gr.)				
Lexico-semantic (Ls)				

Table 1: *Error types and error levels in this research*

In the final stage, the errors will again be evaluated by DeepSeek, and feedback will be provided.

The sample of the study has been selected as personal conversations conducted with the DeepSeek artificial intelligence chatbot. The population of the study consists of all the conversations of the DeepSeek artificial intelligence chatbot in this field.

For data collection, the structured interview model, one of the quantitative research methods, has been selected. An interview is an interaction created by two or more individuals meeting face-to-face for a specific purpose and using verbal or non-verbal communication tools and techniques (Özgüven, 1992, cited in Cemaloğlu, 2014). In the Structured Interview technique, all stages of the interview are planned in advance, and during the interview process, this plan is not deviated from (Cemaloğlu, 2014). Since the author will have direct conversations with DeepSeek in this study and the erroneous texts have already been prepared, it is believed that the mentioned method and technique are the most suitable for the study.

A total of 11 texts were randomly selected from the previously mentioned Menschen B1 textbook for use in the study. A word count limit of 200 to 400 words was set for the texts included in the study. Each text was then modified to include four types of errors (omission, addition, selection, ordering) at both the Grammatical and Lexico-semantic levels. Each text contains 4 grammatical errors and 4 lexico-semantic errors, totaling 8 errors per text. The errors found in each text are shown in the table below.

ET EL	Om.	Ad.	Se.	Or.
Gr.	1	1	1	1
Ls.	1	1	1	1

Table 2: *The distribution of errors in each text*

Therefore, in the 11 texts, there is a total of 88 errors, equally distributed.

Findings

The error detection capability of the artificial intelligence chatbot DeepSeek in identifying errors within texts is specified in the table below.

Text Number	Gr.				Ls.			
	Om.	Ad.	Se.	Or.	Om.	Ad.	Se.	Or.
1	+	+	+	+	+	+	+	+
2	+	+	+	+	+	+	+	+
3	-	+	-	+	+	+	+	+
4	+	+	+	+	+	+	-	+
5	+	-	+	+	-	+	+	+
6	+	+	+	+	-	+	+	+
7	+	+	+	+	+	+	+	+
8	-	+	-	-	-	+	+	-
9	+	+	-	+	+	+	+	+

10	+	+	+	+	+	-	+	+
11	+	+	+	+	+	+	+	+
Total detected	9	10	8	10	8	10	10	10
Total undetected	2	1	3	1	3	1	1	1

+ : error detected

- : error not detected

Table 3: *Detection of errors according to error levels and error types*

It has been observed that, while detecting errors, some errors fail to be identified. As a result, these undetected errors remain uncorrected in subsequent steps, and no feedback has been given concerning them. The data regarding whether the detected errors have been corrected and whether the corrections made are accurate or inaccurate is similarly presented below.

Text Number	Gr.				Ls.			
	Om.	Ad.	Se.	Or.	Om.	Ad.	Se.	Or.
1	+	+	+	+	+	+	+	+
2	+	+	+	+	+	+	+	+
3	X	+	X	+	+	+	+	+
4	+	+	+	+	-	+	X	-
5	+	X	+	+	X	+	+	+
6	+	+	+	+	X	-	+	+
7	+	+	+	+	+	+	+	+
8	X	+	X	X	X	+	+	X
9	-	+	X	+	+	+	+	+
10	+	+	+	+	+	X	+	-
11	+	+	+	+	+	-	+	-
Total Correctly Corrected	8	10	8	10	7	8	10	7
Total Incorrectly Corrected	1	0	0	0	1	2	0	3
Total Uncorrected	0	0	0	0	0	0	0	0

+ : the detected error has been correctly corrected

- : the detected error has been incorrectly corrected

X : the error could not be corrected because it was not detected at the previous stage.

Table 4: *Correction status of detected errors according to error levels and types*

It has been observed that some errors are incorrectly corrected during the error rectification process. Examining the feedback related to these incorrectly corrected errors is not considered meaningful. Consequently, only the errors that were detected in the first stage and correctly rectified in the second stage have progressed to the final stage. The data regarding DeepSeek's feedback capability for these errors is presented below.

Text Number	Gr.				Ls.			
	Om.	Ad.	Se.	Or.	Om.	Ad.	Se.	Or.
1	+	-	+	+	+	+	+	+
2	-	+	+	+	+	+	+	+
3	X	+		-	+	+	+	+
4	+	+	+	-	X	+	X	X
5	+	X	+	+	X	+	+	+
6	+	+	+	+	X	X	+	+
7	+	+	+	-	+	+	+	-
8	X	+	X	X	X	+	+	X
9	X	+	X	-	+	+	+	+
10	-	+	+	+	+	X	+	X
11	-	+	-	+	+	X	+	X
Total Correct Feedback	5	9	7	6	7	8	10	6
Total Incorrect Feedback	3	1	1	4	0	0	0	1
Total No Feedback	0	0	0	0	0	0	0	0
+ : the detected error has been correctly corrected - : the detected error has been incorrectly corrected X : feedback could not be given because the error was not detected in the first stage or was not correctly corrected in the previous stage.								

Table 5: *Feedback status of correctly corrected errors according to error levels and types*

Every error level and type found in the texts has been recorded in earlier tables. In order to make these data more interpretable and to simplify the overall analysis, it is beneficial to examine them statistically. First, the proportional data concerning the initial phase, that is, error detection, is provided in the table below.

ET		EL		
		Gr.	Ls.	ET’ Average
Om.	Detected	82%	73%	77%
	Undetected	18%	27%	23%
Ad.	Detected	91%	91%	91%
	Undetected	9%	9%	9%
Se.	Detected	73%	91%	82%
	Undetected	27%	9%	18%
Or.	Detected	91%	91%	91%
	Undetected	9%	9%	9%
EL’ Average	Detected	84%	86%	85%

	Undetected	16%	14%	15%

Table 6: *Error detection rates by level and type*

The proportional data pertaining to the second stage, namely the correction of detected errors, is also provided below.

ET		EL		
		Gr.	Ls.	ET' Average
Om.	Correctly Corrected	89%	88%	88%
	Incorrectly Corrected	11%	13%	12%
	Uncorrected	0%	0%	0%
Ad.	Correctly Corrected	100%	80%	90%
	Incorrectly Corrected	0%	20%	10%
	Uncorrected	0%	0%	0%
Se.	Correctly Corrected	100%	100%	100%
	Incorrectly Corrected	0%	0%	0%
	Uncorrected	0%	0%	0%
Or.	Correctly Corrected	100%	70%	85%
	Incorrectly Corrected	0%	30%	15%
	Uncorrected	0%	0%	0%
EL' Average	Correctly Corrected	97%	84%	91%
	Incorrectly Corrected	3%	16%	9%
	Uncorrected	0%	0%	0%

Table 7: *Correction rates of detected errors by level and type*

At the final stage, the data concerning the feedback on errors that have been correctly rectified is presented in the table below.

ET		EL		
		Gr.	Ls.	ET' Average
Om.	Correct Feedback	63%	100%	81%
	Incorrect Feedback	38%	0%	19%
	No Feedback	0%	0%	0%
Ad.	Correct Feedback	90%	100%	95%
	Incorrect Feedback	10%	0%	5%
	No Feedback	0%	0%	0%
Se.	Correct Feedback	88%	100%	94%
	Incorrect Feedback	13%	0%	6%
	No Feedback	0%	0%	0%

Or.	Correct Feedback	60%	86%	73%
	Incorrect Feedback	40%	14%	27%
	No Feedback	0%	0%	0%
EL’ Average	Correct Feedback	75%	96%	86%
	Incorrect Feedback	25%	4%	14%
	No Feedback	0%	0%	0%

Table 8: *Feedback rates of correctly corrected errors by level and type*

It is beneficial to share graphs derived from the averages of the error type data at both levels. Initially, a graph concerning the average error detection is provided.

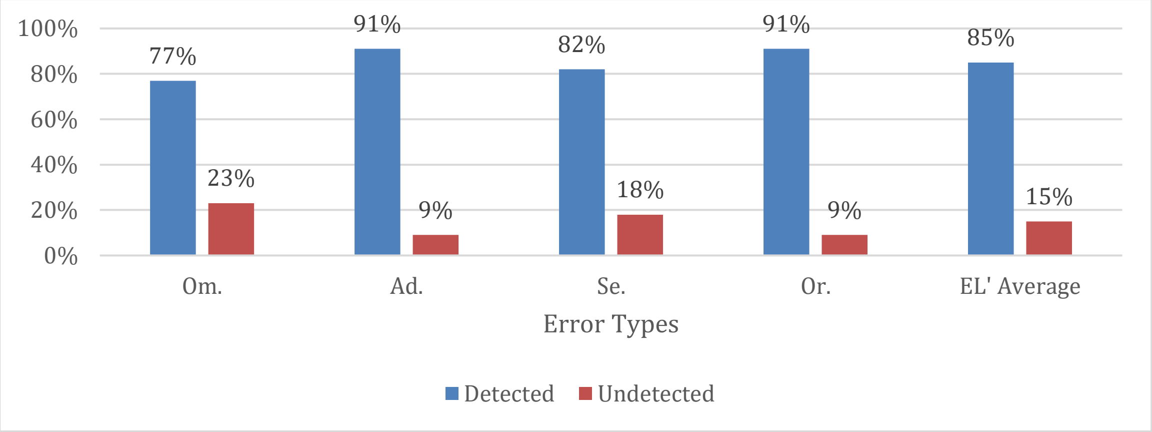

Figure 1: *Average error detection by error type*

The graph obtained from the average of the data concerning the correction of detected errors at both levels is presented below.

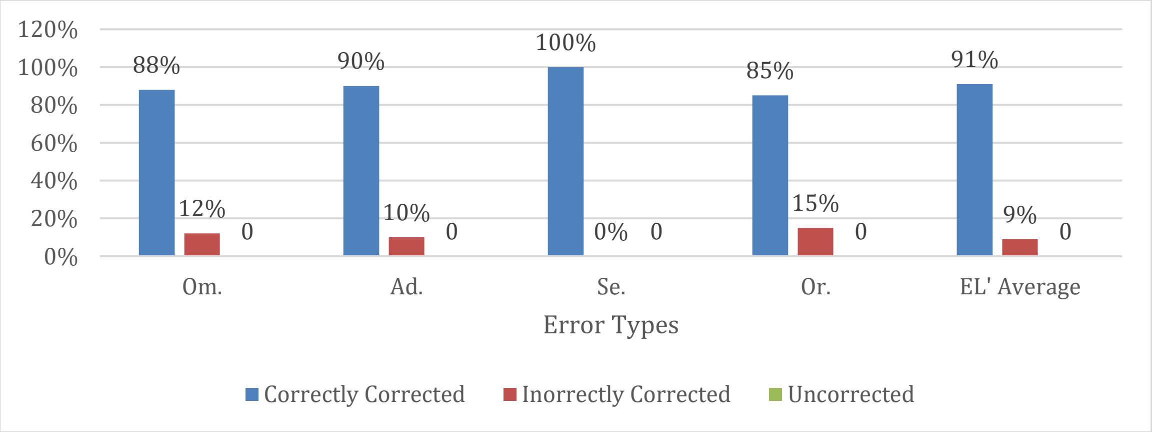

Figure 2: *Average correction of detected errors by error type*

In the final phase, the graph obtained from the average of the data pertaining to the feedback on accurately corrected errors at both levels is provided below.

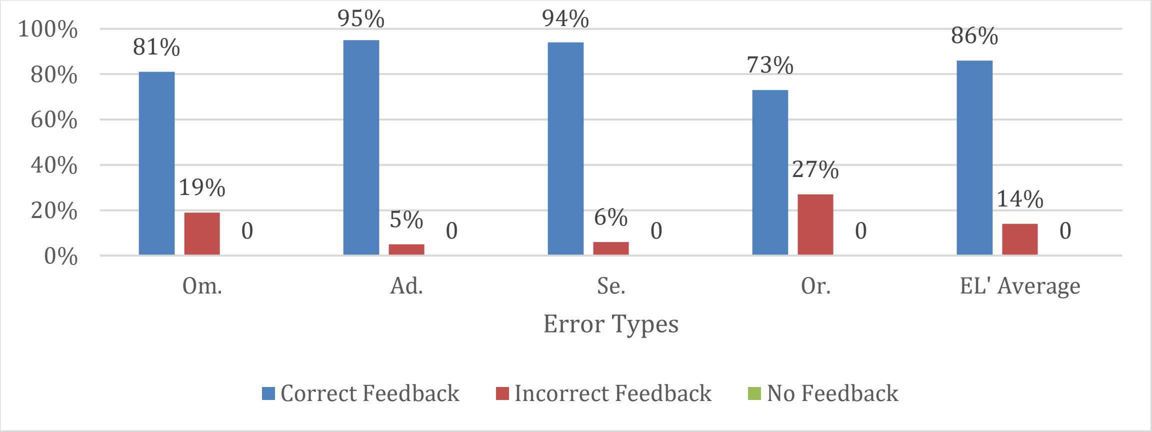

Figure 3: *Average feedback for correctly rectified errors by error type*

Results and Discussion

Based on the obtained findings, the following observations can be made regarding the ability of the artificial intelligence chatbot DeepSeek to detect errors, correct them, and provide feedback on them.

It has been observed that DeepSeek can detect, on average, 85% of errors regardless of error level (Gr. or Ls.) or error type (Om., Ad., Se., Or.), while failing to identify 15% of them. Although this rate cannot be considered poor, it is also not perfect. In fact, the inability to detect 15% of errors poses significant problems in subsequent stages, preventing any corrections or feedback from being provided. This limitation highlights the importance of accuracy in the very first stage of the error treatment process—namely, error detection. If an error goes unnoticed at this initial stage, then the potential for successful correction and meaningful feedback becomes automatically impossible.

A closer examination of error types reveals noteworthy variation. For example, DeepSeek detects 77% of Om. errors while failing to identify 23%, making it the error type with the lowest detection rate and the highest undetected rate. The low detection rate of Om. errors at either the Gr. or Ls. level is concerning, as identifying a missing structure in a sentence should be relatively easier due to the disruption it causes in sentence meaning. Missing components typically leave a visible gap that affects comprehensibility, and thus one might expect a higher detection rate. However, DeepSeek appears to be insufficient in this regard. On the other hand, the detection rate for Ad. and Or. errors is 91% (see Figure 1). This indicates that DeepSeek is more efficient at recognizing errors related to additions and ordering, which suggests that structural and positional cues may be easier for the model to identify.

When it comes to error correction, DeepSeek has performed quite well in addressing the errors it was able to detect. It successfully corrected 91% of the identified errors accurately, while 9% were incorrectly corrected. Additionally, it can be stated that DeepSeek attempts to correct all the errors it detects, regardless of type or level. Notably, it was able to correct all Se. errors accurately, which indicates strong performance in handling substitution errors. However, the lowest correction accuracy rate, at 85%, was observed for Or. errors (see Figure 2). This suggests that while DeepSeek may notice ordering problems, resolving them effectively still poses a challenge.

In terms of feedback provision, DeepSeek has demonstrated a performance that can be considered successful, particularly when providing feedback on correctly corrected errors. It provided accurate feedback for 86% of the correctly corrected errors, while 14% received incorrect feedback. Furthermore, there were no cases in which errors were ignored at the feedback stage. The highest accurate feedback rate was observed for Se. errors at 94%, while the lowest accurate feedback rate was found in Or. errors at 73% (see Figure 3). This disparity implies that DeepSeek's understanding of certain error types extends beyond mere correction, allowing it to explain some corrections more effectively than others.

On average, across all analyses, 84% of errors were detected, while 16% remained undetected. The highest detection rate was observed in Ad. and Or. error types, both at 91%. The lowest detection rate was found in Se. errors, at 73% (see Table 6). Such results are interesting because substitution errors are usually expected to be relatively easy for automated systems to recognize, as they often result in semantic or syntactic anomalies. Yet, the relatively poor detection rate suggests that DeepSeek sometimes interprets substituted words or structures as acceptable variants, reflecting a limitation in its linguistic sensitivity.

With respect to feedback accuracy, it has been observed that DeepSeek provided accurate feedback for 75% of the correctly corrected errors at the Gr. level. The accuracy rates for providing correct feedback were 60% for Or. errors, 63% for Om. errors, 88% for Se. errors, and 90% for Ad. errors. As these rates indicate, DeepSeek exhibited a highly inconsistent performance in this regard. In particular, the accuracy of feedback for Or. and Om. errors was relatively low (see Table 8). These inconsistencies raise concerns about the model's reliability as a tool for language learning and teaching, since effective feedback is crucial for learner development.

When the data are analyzed again, a slightly different but complementary picture emerges. On average, 86% of errors were detected, while 14% remained undetected. The detection rate was 91% for Ad., Se. and Or. errors, whereas Om. errors were detected at a lower rate of 73%. If the performance on Om. errors is disregarded, it can be considered a successful performance overall (see Table 6). The accuracy rate for correctly correcting detected errors in this analysis is 84%. Interestingly, DeepSeek was able to accurately correct all Se. errors in this category, reinforcing the earlier observation that substitution errors are handled well once detected. However, only 70% of Or. errors were corrected, which is the lowest accuracy rate in this area (see Table 7).

Feedback accuracy provides additional insights. DeepSeek was able to provide correct feedback for 96% of the correctly corrected errors. While Om., Ad. and Se. errors received entirely accurate feedback, the accuracy rate for Or. errors was 86%. This indicates that DeepSeek was able to provide correct feedback for almost all (96%) of the correctly corrected errors. Additionally, there were no errors for which feedback was omitted (see Table 8).

A more detailed breakdown across levels also shows consistency in some respects but variation in others. Similar results were obtained for error detection at both levels. The Gr. Se. and Ls. Om. errors had the lowest detection rates at 73%. On the other hand, the highest detection rates were observed for Gr. Ad., Gr. Or., Ls. Ad., Ls. Se. and Ls. Or. errors, all at 91%. Gr.

Ad. and Ls. Ad. errors, as well as Gr. Or. and Ls. Or. errors, share the same detection rates (91%). This indicates that DeepSeek is equally capable of detecting Ad. and Or. errors, regardless of error level. Additionally, the difference in the detection rates for Gr. Se. and Ls. Se. errors is noteworthy (73% - 91%). This suggests that DeepSeek is much better at detecting Se. errors at the lexical-semantic level (see Table 6).

A notable difference is observed in the accuracy of correcting detected errors. DeepSeek performs much more successfully in correcting errors at the Gr. level compared to the Ls. level (97% - 84%). The lowest accuracy rate for correct corrections was found in Ls. Or. errors (70%). Additionally, the correction of Gr. Ad., Gr. Se., Gr. Or. and Ls. Se. errors was flawless (100%). This indicates that DeepSeek accurately corrects Se. errors at a 100% rate, regardless of error level. A similar performance is observed in Om. errors (89% - 88%), while a significant difference exists in the correction of Or. errors. DeepSeek is able to correctly correct 100% of Gr. Or. errors, whereas it can only correctly correct 70% of Ls. Or. errors (see Table 7).

In terms of feedback, another noticeable difference is observed. DeepSeek is able to provide accurate feedback for 75% of errors at the Gr. level, while it can provide correct feedback for 96% of errors at the Ls. level. The lowest accuracy rate for correct feedback is 60%, observed in Gr. Or. errors. The highest accuracy rate for correct feedback, at 100%, is found in Ls. Om., Ls. Ad. and Ls. Se. errors (see Table 8). This clearly shows that DeepSeek has a stronger ability to provide accurate and pedagogically useful feedback at the lexical-semantic level compared to the grammatical level.

When considering the raw numbers, the total number of detected errors at the Gr. and Ls. levels is 37 and 38, respectively (see Table 3). These values are very close to each other, reflecting a balance in error detection capability across levels. The data on correctly correcting the detected errors is as follows: Almost all errors detected at the Gr. level were corrected correctly, with only one being incorrectly corrected. At the Ls. level, however, six errors were incorrectly corrected (see Table 4). During the feedback stage, the number of errors is also similar for both levels: Gr. 36, Ls. 32. DeepSeek was able to provide correct feedback for 27 of the 36 errors at the Gr. level, and for 31 of the 32 errors at the Ls. level (see Table 5). Therefore, it can be concluded that DeepSeek demonstrated consistent and high success in detecting errors at both the Gr. and Ls. levels; while it correctly corrected almost all errors at the Gr. level, it was relatively less successful in correcting errors at the Ls. level. However, it showed near-perfect performance in providing correct feedback for Ls. errors, while being less successful at the Gr. level.

Taken together, these findings point to a nuanced picture. In conclusion, DeepSeek is equally capable of detecting errors at both the grammatical and lexical-semantic levels (84% - 86%). It can correct errors at the grammatical level more accurately than those at the lexical-semantic level (97% - 84%). However, it provides less accurate feedback for correctly corrected errors at the grammatical level compared to the lexical-semantic level (75% - 96%).

Overall, DeepSeek is able to detect any error at an 85% rate, regardless of error level or type. It can correct 91% of the detected errors accurately and provide correct feedback for 86% of the errors that were correctly corrected. These findings illustrate both the potential and the limitations of using AI chatbots such as DeepSeek in educational contexts, particularly in error analysis, correction, and feedback provision. While the results are promising, areas such as handling omissions and ensuring consistent feedback remain important directions for future improvement.

References

Abbott, G. (1980). Towards a more rigorous analysis of foreign language errors. *International Review of Applied Linguistics in Language Teaching, 18*(2), 121–134.

Adamopoulou, E., & Moussiades, L. (2020). An overview of chatbot technology. In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Eds.), *IFIP International Conference on Artificial Intelligence Applications and Innovations* (pp. 373–383). Cham: Springer.

Arslan, K. (2020). Eğitimde yapay zeka ve uygulamaları. *Batı Anadolu Eğitim Bilimleri Dergisi, 11*(1), 71–88.

Bohnensteffen, M. (2010). *Fehler-Korrektur: Lehrer- und lernerbezogene Untersuchungen zur Fehlerdidaktik im Englischunterricht der Sekundarstufe II.* Frankfurt am Main: Peter Lang.

Brown, D. H. (2007). *Principles of language learning and teaching* (5th ed.). White Plains, NY: Pearson.

Cemaloğlu, N. (2014). Veri toplama teknikleri: Nicel-nitel. In A. Tanrıöğen (Ed.), *Bilimsel araştırma yöntemleri* (pp. 135–163). Ankara: Anı.

Corder, S. P. (1982). *Error analysis and interlanguage.* Oxford: Oxford University Press.

Dulay, H., Burt, M., & Krashen, S. (1982). *Language two.* New York: Oxford University Press.

Ellis, R. (1994). *The study of second language acquisition.* Oxford: Oxford University Press.

Erdoğan, V. (2005). Contribution of error analysis to foreign language teaching. *Mersin University Journal of the Faculty of Education, 1*(2), 261–270.

Harden, T. (2006). *Angewandte Linguistik und Fremdsprachendidaktik.* Tübingen: Gunter Narr Verlag.

Haristiani, N. (2019). Artificial intelligence (AI) chatbot as language learning medium: An inquiry. *Journal of Physics: Conference Series, 1387,* 012020. doi:10.1088/1742-6596/1387/1/012020

Hong, W. C. H. (2023). The impact of ChatGPT on foreign language teaching and learning: Opportunities in education and research. *Journal of Educational Technology and Innovation, 5*(1). https://doi.org/10.61414/jeti.v5i1.103

Mohamed, A. M. (2024). Exploring the potential of an AI-based chatbot (ChatGPT) in enhancing English as a foreign language (EFL) teaching: Perceptions of EFL faculty members. *Education and Information Technologies, 29,* 3195–3217. https://doi.org/10.1007/s10639-023-11917-z

Nickel, G. (1972). *Fehlerkunde: Beiträge zur Fehleranalyse, Fehlerbewertung und Fehlertherapie.* Berlin: Cornelsen-Velhagen und Klasing.

Pokrivcakova, S. (2019). Preparing teachers for the application of AI-powered technologies in foreign language education. *Journal of Language and Cultural Education, 7*(3), 108–122. doi:10.2478/jolace-2019-0025