International Journal of Teaching, Learning and Education (IJTLE)
ISSN: 2583-4371
Vol-5, Issue-1, Jan-Feb 2026
Journal Home Page: https://ijtle.com/
Journal DOI: 10.22161/ijtle

A Study on the Effectiveness of Large Language Models in Foreign Language Teaching: A Case Study of French Grammatical Error Correction

Jiaqi Hou[1*], Jinxian Ji[2], Liuyi Yang[3], Xinyi Peng[4], Raphaël El Haddad[5]

[1]Language Centre, Tsinghua University, Beijing 100084, China
[2]Foreign Languages College, China University of Geosciences (Beijing), Beijing,100083, China
[3]Rixin College, Tsinghua University, Beijing 100084, China
[4]Department of Foreign Languages and Literatures, Tsinghua University, Beijing 100084, China
[5]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
* Corresponding Author, houjiaqi@mail.tsinghua.edu.cn

Received: 14 Jan 2026, Received in revised form: 16 Feb 2026, Accepted: 21 Feb 2026, Available online: 26 Feb 2026

Abstract

With the rapid development of artificial intelligence technology, Large Language Models (LLMs) have shown tremendous potential in enhancing teaching effectiveness and promoting personalized learning for students. In foreign language learning, LLMs have been widely used by learners for grammar error correction and writing improvement, making it crucial to understand their advantages and limitations in grammar correction. This study systematically evaluates the effectiveness of three mainstream LLMs—ChatGPT, DeepSeek, and Le Chat—in French Grammar Error Correction. The study found that all three models are capable of identifying and correcting errors in French texts to some extent: DeepSeek performs best in overall capability, ChatGPT excels at comprehensively identifying potential errors, Le Chat performs relatively weaker. All three models perform well in spelling correction, but limitations remain in correcting vocabulary errors and specific grammatical issues, such as agreement in gender and number. This study provides valuable insights into the practical application of LLMs in French education. Future research could further enhance the application of LLMs in foreign language teaching by expanding corpus data and optimizing prompt design.

***Keywords*— Large Language Models (LLMs), Grammar Error Correction (GEC), French as a second foreign language**

I. INTRODUCTION

The development of modern educational technology has greatly accelerated innovation in education. From computer-assisted learning to Web-based learning, and further to AI based learning, technological evolution has continuously driven transformations in teaching practice. In foreign language teaching, learners are expected to comprehensively develop integrated skills in listening, speaking, reading, writing, and translation, as well as intercultural communicative competence. Benefiting from massive corpora and continual optimization of deep-learning algorithms, large language models (LLMs) have demonstrated distinctive advantages across multiple dimensions, such as grammar error correction (GEC), translation, AI-assisted writing and speaking.

In grammar error correction (GEC), LLMs can identify syntactic, lexical, and orthographical errors in various languages, offering relatively accurate and fluent revision suggestions (Bryant et al., 2023). As for translation, although LLM-generated translations may still be imperfect for low-resource languages, for high-

resource languages they can already produce translations comparable in quality—sometimes even more natural and fluent—than professional tools such as Google Translate and DeepL (Hendy et al., 2023). In particular, at the document level, thanks to their strong long-context modeling capability, LLMs can outperform commercial translation tools such as Google Translate and DeepL in human evaluations (Wang et al., 2023). Even for literary translation, some studies suggest that DeepSeek has surpassed traditional neural machine translation in Chinese-to-English translation (Zhang & Zhao, 2025). In AI-assisted writing, compared with essays written by L2 learners, LLM-generated texts often show more coherent organization, more idiomatic expression, and richer vocabulary and syntactic structures (Herbold et al., 2023). In speaking, LLMs perform well in the accuracy of oral feedback and can provide structured feedback and scoring. Moreover, compared with traditional guidance provided by human experts, LLM-based speaking feedback systems are particularly beneficial for students from disadvantaged backgrounds (e.g., under-resourced or impoverished regions) (Mhasakar et al., 2024).

Despite their potential in translation, speaking practice, and other areas, actual learning needs vary across individuals. To preliminarily understand learners' practical use of LLMs across different dimensions in foreign language learning, we conducted an in-class poll collecting data on students' use of LLMs for grammar, pronunciation, writing, speaking, culture, etc. (see Appendix). Most students reported using AI tools primarily for grammar learning and writing correction, with relatively lower frequency for speaking practice and pronunciation. In addition, many widely used LLMs (both in China and internationally) are pretrained mainly on English and Chinese corpora, with relatively less French data. Therefore, a thorough evaluation of the actual effectiveness and limitations of LLMs in French grammar and writing correction is particularly necessary.

II. LITERATURE REVIEW AND RESEARCH QUESTIONS

Grammatical error correction (GEC) refers to the correction of various errors in texts, including lexical, syntactic, and orthographic errors (Bryant et al., 2023). According to Ducrot's French error taxonomy (Ducrot, 2005), lexical errors mainly involve inappropriate lexical choice (verbs, fixed expressions, etc.). Morphosyntactic errors can be divided into: agreement in gender and number (e.g., adjectives, past participles), selection of function words (e.g., prepositions, determiners), verb conjugation (present, future, present perfect, imperfect, etc.), and elision-related errors. Orthographic errors include missing or redundant letters, transposed letter order, or misspellings due to unfamiliarity (e.g., overall spelling deviation).

In second language acquisition, grammatical correction can effectively promote learners' understanding and internalization of grammatical mechanisms by identifying learner errors and providing timely feedback, helping them develop the ability to move from passive correction to autonomous self-revision (Miao & Yao, 2023). Current research on LLMs for GEC mainly focuses on English and Chinese (Wu et al., 2023; Fang et al., 2023; Qu & Wu, 2023), revealing a significant gap for French. In addition, in natural language processing, GEC is often trained and evaluated on official benchmark datasets. Such datasets may not accurately reflect typical error patterns of Chinese L1 French learners in syntactic structures, lexical choice, and gender/number agreement. To address this issue, the present study uses texts written by learners of French as a second foreign language (i.e., a second foreign language beyond English) as its corpus, aiming to fill the current gap in French GEC research. Using learner writings collected from authentic teaching settings, this study systematically evaluates several mainstream LLMs on French GEC performance. Based on the above analysis, the study proposes the following three research questions:

(1) What is the overall performance of several mainstream LLMs in French GEC? Are there significant differences?

(2) Do error types (orthographic, syntactic, lexical) affect the models' correction effectiveness?

(3) What are the main limitations of LLMs when processing French text?

III. METHODOLOGY

3.1 Corpus Data

This study selected written productions by learners of French as a second foreign language as the corpus. The learners' proficiency levels are at the A1-A2 levels of the Common European Framework of Reference for Languages (CEFR). The corpus contains 168 sentences in total. The total number of errors corrected by humans is 116. All human corrections were agreed upon after discussion between two native French speakers and one French teacher.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.5.1.6

3.2 Model Selection and Evaluation

Three LLM products were selected for evaluation: ChatGPT, DeepSeek, and Le Chat. Basic information is shown in Table 1. DeepSeek and ChatGPT are among the most widely used LLMs in China and around the world. Le Chat is an LLM developed by the French company Mistral AI, whose pretraining languages include English and French. Since all three models include English among their primary pretraining languages, this study used English as the base language for prompting to ensure each model can effectively process the input learner writings. The prompt is as follows:

> Please identify and correct any grammatical, lexical and orthographical error in the following text, while keeping the original sentence structures unchanged as much as possible.

Table 1. Large language models used in this study

AI Product	Model / Version	Company	Website	Primary Pretraining Languages	Pre-training Language Coverage
ChatGPT	4o	OpenAI	https://chatgpt.com/	English	Multilingual
DeepSeek	R1 (Deep Thinking version)	DeepSeek	https://chat.deepseek.com/	Chinese, English	Multilingual
Le Chat	Mistral Large 2	Mistral AI	https://chat.mistral.ai/	English, French	Multilingual

To systematically evaluate the effectiveness of the three models for French GEC, this study combined quantitative and qualitative analyses. In the quantitative analysis, errors detected by each model were categorized, annotated, and counted. Then, the model's precision, recall, and F0.5-score were calculated to measure its performance in correcting French grammatical errors. Precision refers to the proportion of correctly corrected errors among all corrections made by the model. Recall indicates the proportion of actual errors successfully identified by the model. The F0.5-score is a composite evaluation metric balancing precision and recall. Qualitative analysis, through specific case studies, compared the correction effectiveness and feedback quality of each model in practical application to provide a more comprehensive evaluation.

IV. RESULTS

4.1 Overall Performance of Mainstream LLMs in French GEC

Statistical results (Table 2) show that DeepSeek achieves the best overall performance, with an F0.5-score of 0.66, substantially higher than ChatGPT (0.512) and Le Chat (0.419). DeepSeek's precision is as high as 0.727, indicating it is better at making fewer but more accurate corrections—well suited to tasks prioritizing precise identification and correction of key errors. In terms of recall, ChatGPT performs better, suggesting it identifies more errors requiring correction. Although Le Chat is the only model whose primary pretraining languages include French, its precision, recall, and F0.5-score are relatively low, indicating substantial room for improvement when handling GEC tasks in French learner writings.

Table 2. Overall performance of three LLMs in French GEC

Model	Precision	Recall	F0.5-score
DeepSeek	0.727	0.483	0.66
ChatGPT	0.504	0.548	0.512
Le Chat	0.427	0.388	0.419

4.2 Impact of Different Error Types on Model Performance

Further examining orthographic, lexical, and syntactic error types, different LLMs show varying performance but also reveal some regular patterns in their ability to handle specific error types. For orthographic errors, all three models perform relatively well, especially DeepSeek, whose F0.5-score reaches 0.843. This indicates that LLMs have relatively strong abilities in identifying and correcting spelling errors, possibly because such errors are more pattern-based with clearer correction paths. For syntactic errors, DeepSeek still maintains the lead (F0.5-score = 0.661). ChatGPT (F0.5-score = 0.506) and Le Chat (F0.5-score = 0.513) show little difference, with the former having higher recall and the latter slightly better precision, indicating their respective strengths. However, as will be discussed in the following section, overall syntactic correction of

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.5.1.6

the three models remains challenging, especially for agreement and tense selection, where models often exhibit "false correction" or "under-correction". For lexical errors, all three models perform weakly, especially Le Chat, whose F0.5-score is only 0.211. DeepSeek maintains a relative advantage (F0.5-score = 0.449), but both its precision and recall are lower than its performance in syntactic and orthographic tasks. This result suggests that word choice errors (e.g., inappropriate word usage, semantic deviation) often involve subtle semantic differences and usage conventions, making them difficult for models to judge accurately and posing greater challenges.

Table 3. Performance by error type (orthographic, syntactic, lexical)

Error Type	Model	Precision	Recall	F0.5-score
Orthographic	DeepSeek	0.882	0.714	0.843
	ChatGPT	0.652	0.789	0.676
	Le Chat	0.545	0.600	0.556
Syntactic	DeepSeek	0.723	0.493	0.661
	ChatGPT	0.494	0.563	0.506
	Le Chat	0.551	0.403	0.513
Lexical	DeepSeek	0.538	0.269	0.449
	ChatGPT	0.381	0.320	0.367
	Le Chat	0.205	0.235	0.211

It is noteworthy that DeepSeek's comprehensive correction effectiveness is the most outstanding, whether in overall performance or across different error types (orthographic, syntactic, lexical). Its F0.5-scores for the four task categories are sequentially: 0.66 (overall), 0.843 (orthographic), 0.661 (syntactic), and 0.449 (lexical). ChatGPT consistently shows higher recall across overall and type-specific tasks, indicating stronger coverage in identifying potential errors. Le Chat's performance is relatively weak overall and across detailed categories, with the lowest F0.5-scores in almost every task, particularly in the lexical dimension.

V. DISCUSSION

Although the three mainstream LLMs demonstrate certain capabilities in French GEC, qualitative analysis of the model correction results reveals several main limitations when processing texts from French beginners.

(1) Weak detection of lexical errors and insufficient semantic discrimination

The phenomena observed in the corpus are highly consistent with the quantitative analysis results. Lexical errors are the most difficult category for all three models, with many word choice errors remaining undetected. For example, in example (a), the student used the verb *mettre* (put on) in the original text. Although both *mettre* (put on) and *porter* (wear) contain the semantic element of *chuān* (wear; put on) in Chinese, the former indicates the action of putting on clothes, while the latter indicates the state of wearing them. In the context, the student wanted to describe someone's outfit, making *porter* (wear) the correct expression. None of the three models identified this word choice error, showing significant shortcomings in distinguishing semantic differences between near-synonyms and judging word appropriateness based on context. Example (b) shows a word choice error regarding time expression. Unlike other times, French cannot use "24h" to denote midnight; the specific word *minuit* (midnight) should be used. Although "24h" seems intuitive from a learner's perspective, it does not conform to actual French lexical usage habits and is an error influenced by negative L1 transfer.

(a) **Original:** *Il* ***mettait*** *un t-shirt blanc, un petit short…* (He was putting on a white t-shirt and a small pair of shorts…)**Not detected by models** (DeepSeek, ChatGPT, Le Chat) **Correct usage:** *Il* ***portait*** *un t-shirt blanc, un petit short…* (He was wearing a white t-shirt and a small pair of shorts…)

(b) **Original:** *Je me suis couchée comme d'habitude* ***à 24h****.* (I went to bed at 24 o'clock as usual.)

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.5.1.6

Not detected by models (DeepSeek, ChatGPT, Le Chat)

Correct usage: *Je me suis couchée comme d'habitude* ***à minuit****.* (I went to bed as usual at midnight.)

(2) Unstable handling of gender/number agreement and limited gender inference

Gender/number agreement is one of the core difficulties of French syntax. In handling this error, the three models exhibited three different correction tendencies.

ChatGPT demonstrated some inference ability, meaning it could infer agreement information based on preceding context. In example (c), the preceding text mentions *mes amis* (my friends, a masculine plural group) and *allés* (went, a masculine plural past participle), from which it can be inferred that the group *nous* (we) is masculine plural. Therefore, the feminine plural past participle *promenées* (walked) later constitutes an agreement error. ChatGPT correctly changed the past participle to the masculine plural form *promenés* (walked) based on the preceding information. However, when subsequent text provides agreement information, ChatGPT did not backtrack to modify earlier text. In example (d), *déguisées* (dressed up) and *parties* (left) are both feminine plural past participles. In French, feminine plural agreement is only used when all members of a group are female, allowing the inference that *je* (I) in the preceding text is female. In this case, the model could not retrospectively adjust the preceding content and did not correct the agreement error in the masculine singular past participle *allé* (went). This suggests the current ChatGPT model relies more on linear sequence than holistic discourse understanding when handling agreement errors. Furthermore, if the text contains no explicit gender information, ChatGPT cannot infer based on content and thus will not correct agreement errors (as in example (e)).

ChatGPT: Inference based on context (difference between preceding and subsequent inference)

(c) **Preceding context:**

Original: D'abord, j'ai visité le Lac Yanqi avec **mes amis**. Nous sommes **allés** en voiture. Nous nous sommes **promenées** au bord de le lac… (First, I visited Yanqi Lake with my friends. We went by car. We walked along the edge of the lake...)

ChatGPT's correction: … Nous nous sommes **promenés** au bord du lac… (We walked along the edge of the lake...)

(d) **Subsequent context:**

Original: Je suis **allé** à Disneyland avec ma sœur. Après nous être **déguisées**, nous sommes **parties**. (I went to Disneyland with my sister. After dressing up, we left.)

Not revised. Correct usage: Je suis **allée** à Disneyland avec ma sœur. (I went to Disneyland with my sister.)

(e) **No inference, no agreement:**

Original: … j'étais très **occupé** (I was very busy.)

Not revised. Correct usage: … j'étais très **occupée** (I was very busy.)

In contrast, DeepSeek typically did not make inferences based on contextual information. For instance, in example (f), which is identical to the original sentence in (c), the model did not modify the erroneous *promenées* (walked, a feminine plural past participle) and left the sentence unchanged. However, sometimes the model provided two possible agreement forms. As in example (g), DeepSeek used inclusive writing to cover different gender possibilities. While the inclusive writing avoids gender misjudgment to some extent, it also indicates that the model lacks necessary discourse reasoning mechanisms and cannot make gender judgments and explicit agreements based on context.

DeepSeek: no inference, but sometimes offers two possibilities

(f) **Original:** D'abord, j'ai visité le Lac Yanqi avec **mes amis**. Nous sommes **allés** en voiture. Nous nous sommes **promenées** au bord de le lac… (First, I visited Yanqi Lake with my friends. We went by car. We walked along the edge of the lake...)

Not revised. Correct usage: … Nous nous sommes **promenés** au bord du lac… (We walked along the edge of the lake...)

(g) **Original:** Je **restais** dans le dortoir. (I was staying in the dormitory.)

DeepSeek's correction: Je suis **resté(e)** dans le dortoir. (I stayed in the dormitory.)

Le Chat demonstrated another type of error trend, where the model frequently mis-corrected originally accurate masculine gender-number agreement to feminine, resulting in overcorrection. In example (h), the *sentence Je suis retourné à l'école…* (I went back to school …) correctly uses the masculine singular form

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.5.1.6

retourné (returned), because the text is written by a male student; however, Le Chat incorrectly changed it to the feminine form *retournée* (returned). This type of error falls under "false positives" (FP) — where the model incorrectly identifies a correct form as an error and makes unnecessary changes. This reflects that Le Chat's judgment mechanism for handling gender information is unstable, prone to making unnecessary modifications even without definitive contextual support, potentially weakening the quality of originally correct text.

Le Chat: Incorrectly changed to feminine (increased FP)

(h) **Original:** Je suis **retourné** à l'école avec un panier de fraises. (I went back to school with a basket of strawberries.)
Le Chat's correction: Je suis **retournée** à l'école avec un panier de fraises. (I went back to school with a basket of strawberries.)

(3) ChatGPT sometimes confuses "error correction" with "optimization," leading to overcorrection

The GEC task in this study specifically emphasized that LLMs need to "keep the original sentence structures unchanged as much as possible", only making necessary corrections to the author's original text. This means if a model makes unnecessary modifications, it increases the number of "false positives" (FPs), negatively impacting its precision. For example, in sentence (i), the student previously wrote that it rained but did not mention the intensity or duration of getting wet. Therefore, the model's modification from *mouillées* (wet, varying degrees) to *trempées* (soaked, a stronger degree) constitutes over-correction. In sentence (j), the model changed *Nous nous sommes saoulés* (We got drunk) to *Nous nous sommes enivrés* (We got intoxicated), replacing a colloquial, slightly rough expression with a more formal, written-style one. This indicates that ChatGPT often combines "error correction" and "polishing" functions during correction, making it difficult to distinguish between genuine error correction and optimization adjustments.

ChatGPT: Correction" or "optimization?

(i) **Original:** Nous étions toutes **mouillées** (we were all wet)

ChatGPT's correction: Nous étions toutes **trempées** (we were all soaked)

(j) **Original:** Nous nous sommes **saoulés**. (We got drunk.)

ChatGPT's correction: Nous nous sommes **enivrés**. (We got intoxicated.)

VI. CONCLUSION

Based on authentic learner-writing data from students studying French as a second foreign language, this study systematically evaluated the performance of three LLMs — ChatGPT, DeepSeek, and Le Chat — in French GEC. Through a combination of quantitative and qualitative analysis, it explored their correction capabilities and limitations regarding orthographic, syntactic, and lexical errors. The research finds that LLMs show high application potential in French GEC tasks but still have some limitations. In overall and error-type-specific correction performance, DeepSeek has the strongest precision and comprehensive ability, showing a "less but more accurate" correction characteristic. ChatGPT has a clearer advantage in recall, being better at comprehensively identifying potential errors. Le Chat performs relatively weakly, especially in correcting lexical errors.

Error types also affect model correction effectiveness. Overall, models perform best in correcting orthographic errors with clear formal features. When modifying syntactic errors, models handle structural errors well but perform poorly on agreement and tense-related errors. Lexical errors, involving semantic discrimination and pragmatic conventions, pose the greatest challenge to models; all three failed to effectively handle some word choice deviations.

Combining the research results with current LLM development trends, future research and applications in the following directions can be outlined:

(1) Expand corpus data

This study is based on French learner writings at A1–A2 levels. Future work could expand to include a wider range of proficiency levels (e.g., B1–C1) and learners at different stages, exploring model performance differences in error detection and correction across proficiency levels. For example, comparing model error detection and correction patterns in beginner versus advanced learner writing could further verify their generalization ability and stability.

(2) Refined evaluation of model optimization capabilities

Subsequent research could further subdivide False Positives (mis-corrections) and False Negatives (missed corrections) in model output into more detailed categories like "false correction", "missed correction", or

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.5.1.6

"over-correction", with a focus on language polishing phenomena within "over-correction". By distinguishing the boundary between "error correction" and "optimization", we can better understand the true scope of LLMs' role in GEC tasks and provide a basis for their rational use in educational settings.

(3) Optimize prompts

Some research suggests that LLM performance may depend on prompt design. Future efforts could attempt to optimize prompts by providing demonstration examples (Brown et al., 2020; Davis et al., 2024; Su, 2024) to improve model correction accuracy and interpretability.

In terms of practical teaching applications, AI-assisted correction can complement manual evaluation. LLMs already have practical value for orthographic and some syntactic errors, potentially reducing teachers' workload on lower-level corrections and, to some extent, freeing them up (Jia, 2025). However, the final results still require manual verification by teachers or teaching assistants to ensure correction accuracy and reliability. Teachers may also guide students to use AI tools appropriately. In writing and self-revision stages, LLMs can serve as a "first-round check" tool, cultivating students' awareness of autonomous learning. At the same time, students should be cautioned not to adopt model modifications uncritically without reflection; teacher's feedback can be used to correct deviations in model output. In this way, learners can improve efficiency while also developing the ability to negotiate and collaborate in human–AI interaction (Wen & Liang, 2024).

REFERENCES

[1] Bryant C, Yuan Z, Qorib MR, et al (2023) Grammatical Error Correction: A Survey of the State of the Art. Computational Linguistics 49:643–701.

[2] Hendy A, Abdelrehim M, Sharaf A, et al (2023) How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation

[3] Wang L, Lyu C, Ji T, et al (2023) Document-Level Machine Translation with Large Language Models. In: Bouamor H, Pino J, Bali K (eds) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Singapore, pp 16646–16661

[4] Zhang S & Zhao C [张曙康, 赵朝永] (2025) 大语言模型之于文学翻译的适切性研究——基于多指标评估的《边城》多模型译文质量对比[Examining the Suitability of Large Language Models for Literary Translation:A Multi-Indicator Assessment of the English Translations of Biancheng]. 中国外语 [Foreign Languages in China] 22:85–95.

[5] Herbold S, Hautli-Janisz A, Heuer U, et al (2023) AI, write an essay for me: A large-scale comparison of human-written versus ChatGPT-generated essays

[6] Mhasakar M, Sharma S, Mehra A, et al (2024) Comuniqa : Exploring Large Language Models for improving speaking skills

[7] Ducrot J-M (2005) La pédagogie de l'erreur : corriger et remédier [Error-based pedagogy: correction and remediation]. Synergies FLE

[8] Miao J & Yao W [苗佳, 姚委委] (2023) 书面纠错反馈对英语写作语法准确度影响的元分析研究[A meta-analysis of the effects of written corrective feedback on grammar accuracy in English writing]. 外语教学理论与实践 [Foreign Language Learning Theory and Practice] 50–61

[9] Wu H, Wang W, Wan Y, et al (2023) ChatGPT or Grammarly? Evaluating ChatGPT on Grammatical Error Correction Benchmark

[10] Fang T, Yang S, Lan K, et al (2023) Is ChatGPT a Highly Fluent Grammatical Error Correction System? A Comprehensive Evaluation

[11] Qu F & Wu Y (2023) Evaluating the Capability of Large-scale Language Models on Chinese Grammatical Error Correction Task

[12] Brown T, Mann B, Ryder N, et al (2020) Language Models are Few-Shot Learners. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., pp 1877–1901

[13] Davis C, Caines A, Andersen ØE, et al (2024) Prompting open-source and commercial language models for grammatical error correction of English learner text. In: Ku L-W, Martins A, Srikumar V (eds) Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics, Bangkok, Thailand, pp 11952–11967

[14] Su Q [苏祺] (2024) 大语言模型在二语教学中的应用效能解析 [Efficacy assessment of large language models application in L2 teaching]. 外语界 [Foreign Language World]

[15] Jia Y [贾彦琪] (2025) 人工智能时代的教育变革及其回应 [Educational reform in the era of artificial intelligence and its response]. 未来与发展 [Future and Development]49:98–102

[16] Wen Q & Liang M [文秋芳, 梁茂成] (2024) 人机互动协商能力：ChatGPT 与外语教育 [Human-AI interactive negotiation competence: ChatGPT and foreign language education]. 外语教学与研究[Foreign Language Teaching and Research] 56:286-296+321

APPENDIX

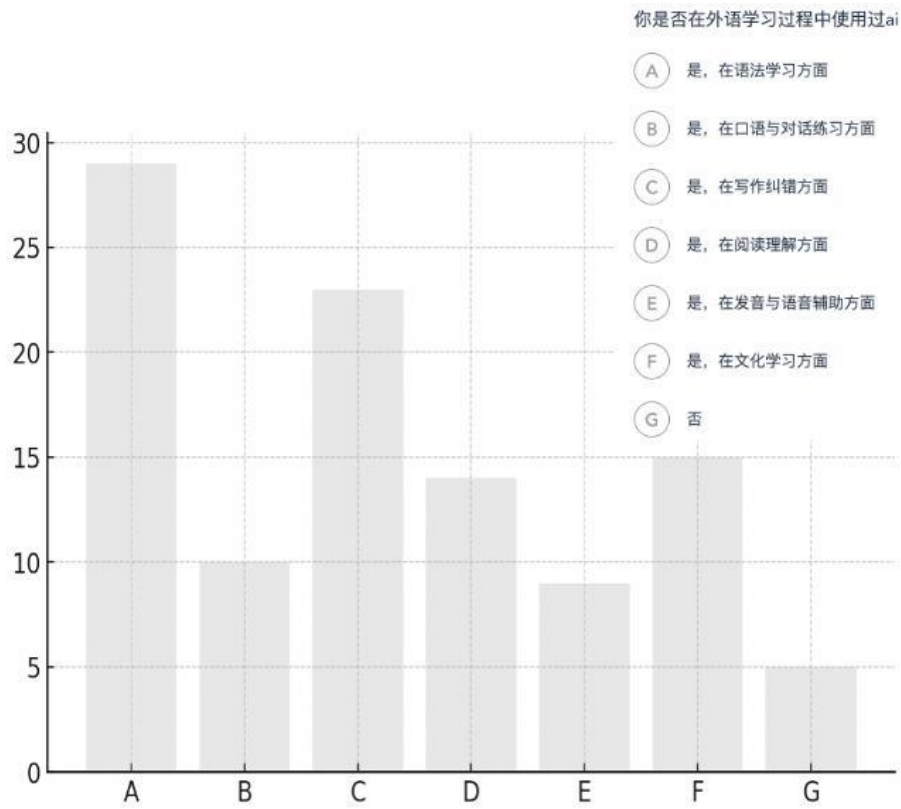

Fig.1. Use of large language models in foreign language learning

Have you used AI in your foreign language learning?

A. Yes, for grammar learning
B. Yes, for speaking and conversation practice
C. Yes, for writing correction
D. Yes, for reading comprehension
E. Yes, for pronunciation and speech assistance
F. Yes, for cultural learning
G. No

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.5.1.6