Malaysian Journal of Social Sciences and Humanities (MJSSH) (e-ISSN : 2504-8562)
2026, Volume 11, Issue 5, e003995
DOI: https://doi.org/10.47405/mjssh.v11i5.3995

RESEARCH ARTICLE

Micro-Corpus LLM Synergy for English Writing Instruction: A Controlled Experiment

Cheng Zhongfang[1], Chun Keat Yeap[2*], Amirah Mohd Juned[3]

[1]Academy of Language Studies, Universiti Teknologi MARA (UiTM) Melaka Branch, 78000 Alor Gajah, Melaka, Malaysia;
Department of Foreign Languages, Bozhou University, Bozhou, China
Email: czfang2026@gmail.com
[2]Academy of Language Studies, Universiti Teknologi MARA (UiTM) Melaka Branch, 78000 Alor Gajah, Melaka, Malaysia
Email: chunkeat@uitm.edu.my
[3]Academy of Language Studies, Universiti Teknologi MARA (UiTM) Melaka Branch, 78000 Alor Gajah, Melaka, Malaysia
Email: amirahjuned@uitm.edu.my

CORRESPONDING AUTHOR (*):
Chun Keat Yeap
(chunkeat@uitm.edu.my)

KEYWORDS:
Large Language Models
Corpus Linguistics
Retrieval-Augmented Generation (RAG)
Educational Technology
Controlled Experiment

CITATION:
Cheng, Z., Yeap, C. K., & Juned, A. M. (2026). Micro-corpus LLM Synergy for English Writing Instruction: A Controlled Experiment. *Malaysian Journal of Social Sciences and Humanities (MJSSH), 11*(5), e003995. https://doi.org/10.47405/mjssh.v11i5.3995

ABSTRACT

Even though Large Language Models (LLMs) demonstrate high efficiency and effectiveness in text generation, their tendencies toward hallucination and limited controllability raise concerns in educational contexts, where transparency and reliability are essential. Responding to claims that traditional corpora are becoming obsolete, this study re-examines the pedagogical value of corpora in the era of LLMs through a focused review of recent literature and a classroom-based randomized controlled trial. Specifically, representative studies published between 2020 and 2024 are systematically reviewed using a three-dimensional analytical framework comprising Efficiency, Controllability, and Educational Adaptability. In addition, a controlled experiment (N = 60) in a university English writing course compares a pure LLM-based instructional model with a corpus-augmented LLM model using Retrieval-Augmented Generation (RAG). The results show that the corpus-augmented model significantly reduced immediate post-test errors (9.7) compared to the LLM-only group (28.3), improved delayed retention, encouraged greater learner questioning behaviour, and reduced teachers' preparation time by approximately 30%. These findings demonstrate that small-scale, task-specific pedagogical corpora play a critical role in enhancing the controllability and instructional effectiveness of LLMs. The study proposes a micro-corpus–LLM synergy framework and provides an open micro-corpus to support classroom implementation and replication.

Contribution/Originality: This study contributes to the existing literature by demonstrating how task-specific pedagogical micro-corpora enhance the controllability and instructional effectiveness of LLM-based writing support. Through a classroom controlled experiment, it documents that corpus-augmented RAG reduces

writing errors, improves retention, increases learner questioning, and enhances teacher preparation efficiency in English writing instruction.

1. Introduction

The most notable consequence of the accelerated development of Large Language Models (LLMs), including ChatGPT, has been fundamental to the transformation of the practices in the field of language education and applied linguistics. Their scale generation of fluent, context-sensitive text has prompted the assertion that there is no longer any imperative to use a traditional corpora and corpus-based pedagogies. Nevertheless, there is increasingly strong evidence that demonstrates that LLMs have significant limitations in the educational setting, especially, their hallucinatory content, inability to provide clear reasoning, and feedback that is challenging to check by both learners and instructors.

These limitations are of critical concern in learning settings that are highly stakes, where accuracy, traceability, and alignment of pedagogy are critical factors. Recent studies have experimented with hybrid techniques incorporating external knowledge sources with LLMs, the most such techniques being Retrieval-Augmented Generation (RAG), which bases model outputs on curated datasets. Although RAG proves to be effective in its application in such areas like medicine or fact-checking, there is little empirical validation in real-life classroom context.

This study addresses the following research questions:

i. RQ1: Does a corpus-augmented LLM (RAG) significantly reduce grammatical error rates compared to a pure LLM-based instructional model?
ii. RQ2: Does the RAG-based approach enhance short-term retention and learner engagement in writing instruction?

This paper advocates combining a selective review of literature with a randomized controlled classroom experiment to argue that traditional corpora (and especially small-scale pedagogical micro-corpora) still have a critical educational role to play in the era of LLM.

2. Research Design

2.1. Literature Selection Procedure and Analytical Framework

It is narrowly based analytical literature research, as opposed to a complete PRISMA-guided systematic review. This methodology is not aimed at ensuring comprehensive access to all literature on large language models but reveals and surveys representative and high-impact studies that are specifically oriented on the combination of LLMs, corpus linguistics, and educational practice. This method is suitable because the study focused on conceptual parallels and pedagogical consequences, and not on bibliometric exhaustiveness.

Temporal, source-based, and thematic criteria were defined for literature selection. Articles were selected from 2019 to 2024 because this period captures the rapid growth of large language models and retrieval-augmented systems. Peer-reviewed conference proceedings and journal articles indexed in sources such as the ACL Anthology, EMNLP

proceedings, NeurIPS proceedings, and relevant journal databases were prioritised. Google Scholar was used as a supplementary search tool rather than a primary indexed source. Search terms included LLM hallucination, retrieval-augmented generation, learner corpus, corpus-based instruction, and LLMs in education to ensure relevance to both the technical and pedagogical aspects of the research topic.

Peer-reviewed conference and journal articles were prioritised as academically rigorous sources. In addition, one highly relevant arXiv survey was included to capture recent technical developments in retrieval-augmented generation. The chosen articles had to directly address limitations of LLMs, retrieval-augmented generation, or educational applications of corpora. The review excluded studies that focused only on engineering optimisation without an educational connection and opinion-based commentaries that were not supported by empirical or analytical evidence. After screening and eligibility assessment, eight representative studies remained for in-depth analysis and theoretical comparison. The selection of these studies was determined by their relevance, citation impact, and contribution to understanding the efficiency, controllability, and educational adaptability of LLM-based systems.

To critically examine the chosen literature, this paper generated a three-dimensional analytical framework that entails efficiency, controllability, and educational adaptability. The efficiency dimension looks at matters concerning the cost of computation, scaling, and speed of computation in language generation and instructional delivery. The controllability dimension is concerned with whether the model outputs can be interpreted and traced back to data sources and corrected in case of errors or hallucinations. Educational adaptability dimension measures the correspondence of each of the approaches to the pedagogical aims, to the involvement of learners, and the transparency of instruction in the classroom.

To increase the credibility of the analytic procedure, two investigators coded the chosen studies on their own with the help of the suggested structure. Inter-coder reliability was evaluated with the help of Cohen kappa coefficient, and the coefficient was 0.83, which suggests the high degree of agreement and testifies to the soundness of the analysis process.

3. Literature Review

3.1. Efficiency: Scale versus Pedagogical Precision

LLMs are efficient in generation and task generalisation. Brown et al. (2020) showed that GPT-3 could perform multiple language tasks without task-specific fine-tuning. However, this efficiency may come at the cost of pedagogical specificity. Corpus-based methods, especially domain-specific datasets, remain useful because they offer controllable examples for targeted instruction. In low-resource and specialised fields, curated information can be more appropriate than scale alone when instructional accuracy is required.

3.2. Controllability: Corpora as Factual Anchors

The limited interpretability of LLMs is one of the major issues affecting their implementation in education. Jain and Wallace (2019) warn that attention mechanisms are not reliable explanations of model reasoning. RAG frameworks address part of this

weakness by grounding generation in external corpora, which contributes to improved traceability. Gao et al. (2023) identify retrieval augmentation as a strategy for improving factual grounding and reducing hallucination risks, which is pedagogically significant because verifiable outputs are important.

3.3. Educational Adaptability: Micro-corpora for Learning Alignment

General-purpose LLM feedback is not always sensitive to specific teaching objectives or learners' proficiency levels. Bryant et al. (2023) review major grammatical error correction datasets and methods, showing that error-annotated data can inform the design of automated corrective feedback. Building on this insight, the present study examines task-based pedagogical micro-corpora dynamically incorporated through RAG.

Table 1 summarises the eight representative studies and shows how each source informs the dimensions of efficiency, controllability, and educational adaptability.

Table 1: Summary of Reviewed Studies (n = 8)

Author(s)	Year	Focus	Dataset Type	Key Contribution	Relevant Dimension
Brown et al.	2020	Few-shot LLMs	Web-scale text	Demonstrated scale efficiency	Efficiency
Jain & Wallace	2019	Model interpretability	NLP benchmarks	Attention ≠ explanation	Controllability
Lewis et al.	2020	RAG framework	Knowledge bases	Grounded generation	Controllability
Gao et al.	2023	RAG survey	Multi-domain RAG studies	Reviewed RAG paradigms and hallucination risks	Controllability
Ji et al.	2023	Hallucination survey	Multi-domain	Taxonomy of hallucination	Controllability
Bryant et al.	2023	Grammatical error correction	GEC datasets / learner error data	Surveyed GEC datasets, methods, and evaluation	Educational Adaptability
Sanh et al.	2022	Prompted multitask learning	Prompted datasets	Demonstrated task generalization	Efficiency
Bender et al.	2021	Ethics of LLMs	Web corpora	Risks of scale	Educational Adaptability

Large language models (LLMs) provide important gains in generation efficiency. Brown et al. (2020) demonstrated that GPT-3 could perform many tasks without gradient updates or task-specific fine-tuning. Nevertheless, curated corpora remain useful for meeting domain-specific instructional needs. Sanh et al. (2022) showed that prompted multitask training with structured task data can support zero-shot generalisation, indicating that data design can matter as much as model scale. This suggests that, in educational contexts, the quality and pedagogical relevance of data may be more important than scale alone.

3.4. Controllability Dimension: Corpora as "Factual Anchors"

LLM interpretability is another major challenge for educational technology. Although some studies have attempted to explain model decisions through attention mechanisms, Jain and Wallace (2019) argued that attention weights are not reliable explanations. Conversely, RAG improves generation by retrieving evidence snippets from external knowledge sources, such as corpora, making black-box generation more traceable. Gao et al. (2023) reviewed RAG research and identified retrieval augmentation as a way to improve factual grounding and reduce hallucination risks. This provides a technical justification for using verifiable answers in education.

3.5. Educational Adaptability Dimension: Precision Empowerment via Micro-corpora

Pedagogical specificity can be limited when LLMs are used directly for teaching feedback. In contrast, small, carefully designed pedagogical corpora can provide targeted support. In grammatical error correction, Bryant et al. (2023) summarised major GEC datasets, methods, and evaluation practices, showing that error-annotated data can inform automated corrective feedback. Extending this logic, the present experiment examines how micro-corpora can be added to dynamic teaching interactions through RAG.

4. Classroom Controlled Experiment

4.1. Participants and Design

The sample size in this research was 60 second year undergraduate students who were taking an English writing course in a university level. The sample size was similar in terms of academic background and all had been taught the English grammar earlier. In order to achieve group equivalency, students were randomly grouped into either one of two training conditions, namely, an LLM-only group consisting of 30 participants and a corpus-enhanced LLM group through retrieval-augmented generation (RAG) system, which also included 30 participants. Before the intervention, pre-test involving assessment of English articles usage was conducted on both groups. The pre-test results were statistically analyzed, showing that there was no significant difference between the groups ($p = .42$), which means that the groups are homogeneous in terms of the results. The experiment followed single-masked design whereby the participants had no knowledge of the instructional condition under assessment.

4.2. Micro-corpus Construction and Replicability

The research has used a task-specific pedagogical micro-corpus that is used to aid grammar-oriented writing in a way similar to conceptually different large-scale learner corpora like the International Corpus of Learner English (ICLE). The micro-corpus consisted of 2,200 sentences that showed the examples of both correct and incorrect English article usage (a, an and the) based on the typical error patterns in undergraduate academic writing. The sentences were based on classroom writing examples selected and edited by the instructor and were chosen to reflect on learner-like errors naturally found in instructional practice as opposed to the large-scale data of learners that typically occur naturally.

All the sentences of the corpus were hand-marked to reveal the kind of article error, the amended form and a short explanatory comment that was meant to justify the instructional feedback. The corpus thus obtained consisted of about 34,800 tokens and its mean length per sentence was 15.8 which was small enough to be easily retrieved and at the same time it was pedagogically relevant. The micro-corpus was incorporated into a retrieval-augmented system of generation, in which sentence embeddings were stored in a vector database and can be dynamically retrieved to base LLM-based feedback generation on concrete examples.

In order to provide replicability, the process of corpus construction was to be replicated in other teaching environments with ease. To reproduce or re-use the corpus, educators can use a systematic process which includes gathering brief texts by the learners which address a particular aspect of grammar, labeling and determining the frequent error patterns along with their corrections and explanations, indexing the annotated sentences by using a set of vector embeddings, and incorporating the indexed corpus into an LLM-based model by passing the indexed corpus through a retrieval-augmented generative pipeline. The given approach allows developing similarly scoped pedagogical micro-corpora in accordance with various teaching goals.

4.3. Procedure and Measures

At the stage of intervention, both groups of students were expected to perform an email-writing activity in a time frame of ten minutes, the task of which was to provoke as many examples of using English articles as possible. The learners who were provided with the RAG-based group were provided with automated feedback which was provided by the LLM, however, based on the retrieved examples within the pedagogical micro-corpus so that they could compare the model output with the material within the concrete corpus. Comparatively, the general-purpose LLM interface was used with students in the LLM-only group, where they could also obtain explanations and examples but not the curated corpus.

Multiple measures were employed in measuring learning outcomes to measure performance and engagement. The article usage error rate was the main outcome variable, and it was assessed by the blind rating of trained assessors, and the interrater reliability had a Cohen k -value of 0.85. The retention of knowledge was assessed with the help of delayed post-test one week after the intervention. Moreover, the learner's engagement was also operationalised based on the number of clarification questions asked by the students throughout the hours of the instructional session which were recorded by the teaching assistants. The teacher preparation time was also recorded in descriptive terms to investigate the efficiency gains that were likely to be achieved in regard to the corpus-enhanced instructional method.

4. Results

Independent samples t-tests with large effect sizes ($d > 0.8$) were used to find that the RAG group was significantly better than the LLM-only group on each outcome measure of learning. The amount of time spent on preparing teachers was also considerably reduced.

Table 2 presents the immediate and delayed post-test results, learner questioning rates, and teacher preparation time for the LLM-only and RAG groups.

Table 2: Comparison of Post-test Measures between Groups (M±SD)

Measure	LLM-only Group (n=30)	RAG Group (n=30)	*p*-value	Cohen's *d*
Immediate Post-test Error Rate (%)	28.3 ± 5.1	9.7 ± 3.8	< .001	1.12
Delayed Post-test Error Rate (%)	21.4 ± 4.6	8.1 ± 3.2	< .001	0.95
Avg. Questions per Student	2.1 ± 1.0	4.6 ± 1.4	< .001	1.05
Teacher Preparation Time (min)	45 ± 5	31 ± 6	-	-

5. Discussion: A Micro-Corpus–LLM Synergistic Educational Framework

The results of the study offer empirical evidence on the suggested analytical framework and showed how a corpus-enhanced LLM model can overcome the main shortcomings of the all-LLM-based instruction models in the educational context. When analyzed in the context of efficiency, the findings show that the incorporation of a task specific pedagogical micro-corpus in a retrieval augmented system of generation can lead to a decreased time of instructional preparation without affecting learning outcomes. The observed shorter teacher preparation time with the corpus-augmented condition implies that well-managed micro-corpora can serve as instructional resources which may be repeatedly used by educators to provide the target feedback more effectively than ad hoc prompt engineering in isolation. The result is consistent with previous studies highlighting the affordability of small but high-quality datasets in specialist areas and raises questions to the belief that pedagogical performance must be based on large-scale data.

In terms of controllability, the corpus-augmented methodology showed a significant decrease in the errors related to hallucinations as they were expressed in significantly lower post-test and delayed post-test error rates. The RAG system made the feedback process, previously a highly opaque generative process, into a semi-transparent teacher-learner interaction by basing the output of the LLM on retrieved examples of corpus. Not only were learners offered corrections, but also concrete, verifiable examples based on the corpus, which contributed to better learning and better retention, presumably. This result can respond directly to the questions of the literature about the interpretability and credibility of the LLM-generated feedback in the educational context and offer a concrete piece of evidence that corpus can be used as a viable source of factual support to be utilized in instruction.

Regarding the educational adaptability, it is possible to mention that more frequent cases of questions asked by the learners in the corpus-enhanced group may indicate that basing feedback on corpus evidence facilitates more engagement, as opposed to passive acceptance of automatic explanations. The presence of concrete examples does seem to have motivated a more active learning process where learners questioned the use of language, thus facilitating a more dialogic and reflective process of learning. Besides, it was found that the higher retention performance in delayed post-test refers to the fact that the superiority of the micro-corpus-LLM synergy is not only related to immediate

performance in a task but also to more sustainable learning results. Taken together, these results point to the fact that pedagogical micro-corpora, when combined via retrieval-augmented architectures, can not only help improve instructional accuracy but also learner agency and cognitive engagement.

6. Limitations and Future Directions

The results of this study are encouraging, but a number of limitations may be listed. To begin with, the experimental intervention involved one grammatical feature, that is, the use of English articles, and a rather brief instructional activity. This design decision was taken to maximise the internal validity and to measure the effects of learning with the greatest precision but this design inevitably constrains the generalisability of the results to other domains of linguistics. Further study is needed to apply this method to other grammatical, lexical, and discourse-level characteristics to determine how well the micro-corpus-augmented model can be applied to a larger number of instructional settings.

Second, the pedagogical micro-corpus that was employed in this study was instructional and based on one teaching context. Even though the corpus was created in a way that was replicable and adaptable, additional studies are required to compare the efficacy of task-specific pedagogical micro-corpora to larger groups of learners, as well as general reference corpora. These comparisons would aid in explaining the relative benefits of the various types of corpus when used in tandem with LLM using retrieval-augmented architectures. Moreover, the longitudinal studies should be considered longer to investigate the impact of corpus-enhanced LLM-feedback on the progression of language over longer periods of time.

Lastly, since the implementation of LLMs in teaching practice is still growing, the future research must focus on the problems of data control, data transparency, and the ethical use of the technology. To make sure that the instructional systems based on LLM are pedagogically sound and socially responsible, it will be necessary to develop the set of guidelines to be used in corpus construction, annotation, and responsible use of AI.

7. Conclusion

This paper shows that the development of huge language models does not make traditional corpora outdated. Rather, in the situation of careful incorporation in the form of retrieval-augmented generation, small-scale pedagogical micro-corpora are instrumental in improving the controllability, transparency, and educational quality of the instructional system built around LLM. Based on the evidence of curated corpus, the suggested micro-corpus-LLM synergy can overcome several major weaknesses of the LLM-only strategies, such as hallucination, interpretability, and incompatibility with didactic goals.

The results of the classroom-controlled experiment are empirical evidence that this synergistic model enhances the learning outcomes, facilitates the engagement of learners, and decreases the workload of the instructions. On a larger level, the research aids into the current discussion of the role of corpora in the era of artificial intelligence by proving that the quality of pedagogical value is in neither the amount of data, but the intentional plan and arrangement of domain-specific materials. The given framework

becomes a method to the replicated and evidence-based future research and practice in intelligent language education.

Ethics Approval and Consent to Participate

The study followed the ethical requirements of the authors' institution. Participants were informed of the study purpose, joined voluntarily, and provided informed consent. No personally identifiable information was reported.

Acknowledgement

The authors would like to thank the participating students and teaching assistants for their support during the classroom experiment.

Funding

This study received no funding.

Conflict of Interest

The authors declare no conflict of interest.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems,* 33, 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Bryant, C., Yuan, Z., Qorib, M. R., Cao, H., Ng, H. T., & Briscoe, T. (2023). Grammatical error correction: A survey of the state of the art. *Computational Linguistics, 49*(3), 643–701. https://doi.org/10.1162/coli_a_00478

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv*. https://doi.org/10.48550/arXiv.2312.10997

Jain, S., & Wallace, B. C. (2019). *Attention is not explanation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3543–3556). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1357

Ji, Z., Lee, N., Frieske, R. M., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information *Processing Systems*, 33, 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Le Scao, T., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma Sharma, S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N. V., ... Rush, A. M. (2022). Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=9Vrb9D0WI4