International Journal of Research in English Education

Online ISSN: 2538-3027

Homepage: https://ijreeonline.com

Scaffolded Large Language Model-Mediated Retrieval Practice Improves Collocational Competence in EFL Learners

Ali Hemmati[1]

1. Department of English Language Teaching, Farhangian University, Tehran, Iran

Email: a.hemmati@cfu.ac.ir

Article Info

Article type:
Research Article

Article history:
Received:
February 2, 2026
Revision:
March 26, 2026
Accepted:
March 30, 2026
Available online:
March 30, 2026

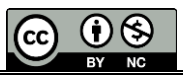
***Keywords*:**
collocational competence,
contextual variability,
EFL learning,
LLM-mediated instruction,
retrieval practice

ABSTRACT

This study examines how scaffolded retrieval practice mediated by a large language model (LLM) influences the receptive knowledge of verb–noun collocations among intermediate learners of English as a foreign language, in comparison with a closely matched corpus-based approach. Drawing on usage-based perspectives on lexical entrenchment and cognitive accounts of retrieval and contextual variation, it was hypothesized that stronger short-term learning and more durable retention would be produced when adaptive LLM scaffolding was combined with safeguards ensuring linguistic authenticity. Sixty-five intermediate learners were randomly assigned to either an LLM-mediated condition or a corpus-based condition and took part in six retrieval-oriented sessions over a two-week period. Receptive collocational knowledge was measured through parallel pretests, immediate posttests, and delayed posttests administered four weeks later, while process data documented retrieval attempts, contextual diversity, and time-on-task. In addition, a purposive subsample of 12 participants engaged in semi-structured interviews, which were analyzed using reflexive thematic analysis supported by MAXQDA 2022 software. The findings indicated that learners in the LLM condition achieved significantly larger gains on both posttests. These advantages were primarily associated with higher retrieval frequency and greater contextual variability, whereas time-on-task played only a minor role. Interview data suggested increased noticing, growing confidence in validated LLM examples, and sustained engagement throughout repeated retrieval cycles. The results indicate that carefully scaffolded and monitored LLM use, embedded within structured practice schedules, can support collocational learning and complement corpus-based resources in second-language instruction.

Keywords: collocational competence; contextual variability; EFL learning; LLM-mediated instruction; retrieval practice

Cite this article: Hemmati, A. (2026). Scaffolded Large Language Model-Mediated Retrieval Practice Improves Collocational Competence in EFL Learners. *International Journal of Research in English Education*, *11*(1), 1-22.



Publisher: Science Academy Publications.

1. Introduction

The Collocational competence, understood as the ability to recognize and produce conventional word combinations such as "reach an agreement" or "make a decision," is widely regarded as a crucial indicator of advanced lexical development and a prerequisite for idiomatic, fluent L2 English expression (Du, 2022). However, a significant body of classroom research shows that many intermediate EFL learners still struggle with even high-frequency verb–noun pairings. These ongoing gaps matter: lacking collocational knowledge often limits learners' accuracy, reduces naturalness, and hinders their expressive flexibility in both spoken and written communication (Sun & Park, 2023).

Corpus-based and data-driven learning (DDL) approaches have long been promoted as effective means to raise collocational awareness. By encouraging learners to analyze concordance lines and notice recurring lexical patterns, DDL can facilitate important receptive insights (Liu & Gablasova, 2025). However, evidence regarding its capacity to support sustained retention or meaningful productive gains remains mixed, particularly for learners who require more structured repetition and targeted support to consolidate new form–meaning mappings (Sun & Park, 2023). The core limitation appears to lie not in exposure alone but in the absence of sufficient, scaffolded retrieval opportunities necessary for forming stable lexical associations.

The rapid emergence of generative large language models (LLMs) has introduced a promising avenue for addressing this limitation. LLMs can supply individualized scaffolding, generate abundant contextual examples, and facilitate iterative retrieval cycles with a level of responsiveness and scale difficult to achieve through traditional corpus tools (Liu et al., 2024; Pack & Maloney, 2023). Meta-analytic findings further suggest generally positive learning outcomes when LLMs are integrated into instruction (Wang & Fan, 2025). Yet despite their expanding presence in educational practice, there remains a marked lack of rigorous, domain-specific evidence examining how LLMs can support the acquisition of multiword lexical units—an area where repeated retrieval is theoretically central to entrenchment.

This gap is especially salient given the alignment between cognitive memory research, which highlights retrieval and spacing as key mechanisms for durable learning, and usage-based accounts that view collocational competence as the product of repeated, meaningful encounters that strengthen associative links (Bego et al., 2024; Bybee, 2010). To date, few studies have directly compared LLM-mediated retrieval practice with established corpus-based methods under controlled conditions or investigated the process variables through which such tools may exert their effects.

The present study, therefore, provides a timely contribution by evaluating, through a randomized and pre-registered design, whether scaffolded LLM-based retrieval practice enhances

intermediate EFL learners' receptive verb–noun collocation knowledge more effectively than an equivalently sequenced corpus-based alternative. To orient the investigation, the study addresses the following research questions:

1. To what extent does scaffolded LLM-mediated retrieval practice improve intermediate EFL learners' receptive verb–noun collocation knowledge compared with corpus-based retrieval practice?
2. Which process variables (retrieval attempts, example variability, and time-on-task) more effectively mediate the relationship between instructional condition and receptive collocation gains?

2. Literature Review

2.1 Theoretical Fmework

Understanding how learners acquire collocations benefits from integrating usage-centered models of lexical organization with process-level accounts of attention, practice, and memory consolidation. Usage-based and constructionist frameworks hold that recurring word sequences become entrenched through repeated encounters in meaningful contexts; frequent co-occurrences and contextual diversity help transform loosely associated word pairs into chunked units available for efficient retrieval (Bybee, 2010). From this perspective, instructional input that supplies numerous, varied exemplars across communicative settings promotes robust representation formation because learners extract distributional regularities and map form to function through experience.

Yet exposure on its own rarely guarantees learning. As the Noticing Hypothesis proposes, learners must consciously register relevant features in the input before those features can be incorporated into their developing linguistic system (Schmidt, 1990). Recent research reinforces this point: studies show that when learners' attention is explicitly directed toward recurrent word combinations, their sensitivity to collocational patterns and their long-term retention improve markedly (Li et al., 2022; Wang, 2024). Because collocations are probabilistic and often escape learners' attention when they focus primarily on single-word meanings, instruction that foregrounds their salience, by highlighting co-occurrences, prompting comparisons, or supporting inductive pattern discovery, can play a decisive role. Corpus-informed classroom investigations further demonstrate that targeted awareness-raising tasks support deeper internalization of collocations than unguided exposure alone (Boone & Eyckmans, 2023). Complementing this, skill-acquisition theories provide an account of how initial declarative recognition of a collocation can, through repeated practice with feedback, be transformed into fluent and automatized retrieval (Jeong & DeKeyser, 2023). Taken together, these findings suggest that effective collocation pedagogy typically integrates both guided noticing and structured retrieval practice, ensuring that

learners not only detect collocational patterns but also come to access them swiftly and reliably in use.

Memory research provides two operationally powerful mechanisms that support such pedagogies. First, retrieval practice—actively recalling information rather than passively re-reading—substantially improves long-term retention; when retrieval is spaced across intervals, memory traces consolidate more effectively (Bego et al., 2024; Lyle et al., 2022). Second, variability in encoding contexts (different syntactic frames, topical scenarios) fosters abstraction and transfer, reducing the risk that learned pairings remain tethered to a narrow exemplar. Accordingly, collocation instruction that schedules spaced retrieval trials and exposes learners to target items in diverse contexts should produce better delayed recognition and more flexible comprehension.

LLMs map onto these theoretical levers in compelling ways: they can generate many native-like, contextualized exemplar sentences on demand, offer immediate corrective feedback, and adapt prompts to increase retrieval difficulty, affordances that theoretically support noticing, repeated retrieval, and contextual variability (Cong, 2024; Pack & Maloney, 2023). Yet theory also warns of risks. LLM outputs do not automatically reflect corpus-derived frequency distributions; they can present plausible but low-frequency or non-native collocations unless verified (Cong, 2024). Moreover, retrieval practice benefits depend critically on prompt design, feedback timing, and spacing schedules; absent careful instructional design, additional practice may yield limited or short-lived gains (Bego et al., 2024).

Taken together, usage-based and cognitive-memory theories suggest considerable potential for LLM-mediated collocation instruction, provided that such systems are embedded within a carefully scaffolded architecture that (a) directs learner attention to co-occurrence patterns, (b) ensures example authenticity and contextual variability, and (c) implements spaced retrieval with appropriate feedback, conditions shown to support durable collocation learning and delayed retention (Fang et al., 2024; Huang, 2025). The present study tests this theoretically informed hypothesis by comparing a pre-verified, scaffolded LLM retrieval-practice condition to a rigorously matched corpus-based retrieval practice condition, focusing on immediate and delayed receptive outcomes and on process variables theorized to mediate learning.

2.2 Review of Previous Studies

Research on collocations began with corpus-driven descriptions that mapped out what learners actually struggle with, helping define the instructional targets that later studies would pursue. Du's (2022) large-scale analysis of learner writing is a good example: the study showed that more proficient writers draw on frequent, prototypical verb–noun combinations and tend to use more abstract nominal elements. This provides a highly authentic picture of the collocational patterns

learners often fail to produce and, in doing so, helps teachers identify where instruction is most needed. Yet the study remains descriptive; because it does not evaluate pedagogical interventions, it cannot tell us which types of instruction genuinely move learners forward.

Building on this descriptive groundwork, Sun and Park (2023) synthesized 32 studies on corpus-based collocation teaching. Their review indicates that concordance-driven activities reliably enhance receptive knowledge, and their careful cataloging of task types, assessment windows, and learner-related moderators gives the field a clearer view of how corpus-based instruction is typically implemented. Even so, the primary research they reviewed varies widely in methodology, especially in how collocational knowledge is measured. Many of the included studies also lacked delayed posttests. As a result, conclusions about productive transfer and long-term retention remain provisional.

As generative AI tools started entering the applied-linguistics conversation, researchers initially focused on methodological possibilities. Pack and Maloney (2023) offered one of the earliest and most practical demonstrations of how tools like ChatGPT could be used to create tasks, generate examples, and support research workflows. Their work provided teachers and researchers with an accessible entry point into LLM use and highlighted the clear procedural benefits these tools offer. However, because the article centered on methodological potential rather than learning outcomes, it left open whether these affordances translate into measurable language gains.

Once LLMs were adopted informally by learners, empirical studies began capturing how students engaged with these tools outside formal instruction. Liu et al. (2024) surveyed and interviewed Chinese EFL learners about their everyday use of LLM platforms. Their mixed-methods design produced rich descriptions of learner strategies, perceptions of benefit, and factors shaping tool adoption. While this makes their study valuable for understanding feasibility and learner behavior, its reliance on self-reports and absence of objective pre-/post-testing means it cannot speak directly to actual improvements in collocational knowledge.

Conceptual contributions also emerged as educators sought guidance on how to integrate LLMs responsibly. Cong (2024) outlined a set of pedagogically principled recommendations, identifying both the potential advantages of LLMs, such as access to varied examples and adaptable scaffolding, and the risks, including hallucinated or infrequent outputs. These guidelines offer practical steps for teachers, such as cross-checking AI-generated examples against corpora. However, because these recommendations have not been empirically tested in collocation-focused instruction, their effectiveness remains theoretical.

As interest in LLM-based instruction grew, broader syntheses appeared. Wang and Fan's (2025) meta-analysis of 51 studies reported positive overall effects of LLM-supported learning and moderate improvements in higher-order thinking skills. This quantitative review is valuable for

demonstrating the general promise of LLMs and for providing subgroup analyses across instructional contexts. Nonetheless, the included studies span diverse domains, often adopt short-term designs, and rarely examine fine-grained lexical outcomes like collocation learning. Consequently, the meta-analysis cannot meaningfully inform pedagogical decision-making about multiword lexical development.

In parallel, cognitive-memory research continued to refine principles relevant to collocation learning. Studies by Lyle et al. (2022) and Bego et al. (2024) reaffirm that spaced retrieval practice reliably strengthens long-term retention, and that retrieval difficulty and scheduling are critical parameters for learning efficiency. These findings supply powerful theoretical mechanisms and offer design principles that can be incorporated into collocation instruction. Yet because much of this evidence comes from single-word learning or problem-solving contexts, it remains unclear whether the same retrieval schedules function optimally for multiword combinations.

A small number of intervention studies have begun testing AI-based instruction more directly. Early experiments, such as pilot RCTs and classroom trials, suggest that AI-generated feedback can benefit general writing and vocabulary development (Yang & Li, 2024). These studies provide encouraging causal evidence and confirm that AI can be integrated feasibly into instruction. However, they vary substantially in task design and research duration, seldom include delayed posttests, and rarely isolate mechanisms such as retrieval practice or example variability. Moreover, direct comparisons between LLM-based instruction and established corpus-driven approaches remain scarce, limiting the conclusions that can be drawn about durable, mechanism-specific gains in collocational competence.

Taken together, the literature paints a coherent picture. Corpus-based instruction expands learners' exposure to authentic collocations and supports receptive recognition, but the evidence for productive use and long-term retention is inconsistent, partly due to methodological variability and limited follow-up testing. LLMs promise rapid exemplar generation, scalable scaffolding, and efficient opportunities for spaced retrieval, yet domain-specific, experimentally controlled evaluations remain surprisingly limited. Although cognitive-memory research identifies retrieval practice and spacing as central drivers of durable learning, these principles have yet to be systematically applied and compared across LLM- and corpus-based approaches for multiword lexical items. This gap underscores the need for a randomized, tightly controlled comparison that standardizes target items, exposure time, and retrieval schedules, verifies LLM-generated exemplars for authenticity, and assesses both immediate and delayed outcomes in order to determine whether scaffolded LLM-based retrieval practice can deliver superior and lasting gains in verb–noun collocation knowledge.

3. Materials and Methods

3.1 Research Design

This study adopted a randomized-controlled design with an embedded qualitative component to investigate whether scaffolded retrieval practice mediated by a large language model (LLM) improves EFL learners' receptive knowledge of verb–noun collocations. Two parallel conditions were constructed to ensure comparability: an LLM-mediated retrieval-practice group and a corpus-based retrieval-practice group. The qualitative strand, consisting of in-depth interviews with a purposive subsample, allowed us to probe how learners experienced the intervention and how they evaluated the usefulness, trustworthiness, and authenticity of model-generated feedback. Collectively, this mixed-methods design supports a deeper understanding of both the measurable learning outcomes and the cognitive–affective processes shaping learners' engagement.

3.2 Participants

A total of 65 intermediate EFL learners (CEFR B1–B2), aged 18 to 30, participated in the study. Participants were recruited from two university English language programs in Kermanshah through course announcements and volunteer sign-up lists. At the time of data collection, the researcher was teaching at both institutions; however, participation was strictly voluntary, and students were explicitly informed that participation or non-participation would have no effect on course grades or evaluation.

Interested learners completed a placement test and a background questionnaire to confirm appropriate proficiency levels, comparable prior exposure to English, and no recent engagement in intensive collocation-focused instruction. Following eligibility screening, participants were assigned to instructional conditions using stratified randomization to ensure approximate balance in proficiency level and gender distribution.

Thirty-one learners were assigned to the LLM-mediated retrieval-practice group, and thirty-four learners were assigned to the corpus-based retrieval-practice group. Attendance was monitored throughout the two-week intervention, and attrition was minimal, resulting in a final sample that closely reflected the original group allocation. For the qualitative component, 12 participants (6 from each condition) were selected using maximum-variation sampling to capture a wide range of engagement profiles and performance levels. This purposive strategy ensured that learner perspectives were not drawn exclusively from high- or low-performing individuals, thereby enhancing the interpretive depth of the qualitative findings.

3.3 Materials and Instruments

3.3.1 Instructional Materials and Assessment Instruments

The instructional materials targeted 40 high-utility verb–noun collocations that are known to pose persistent difficulty for advanced L2 learners. Candidate items were first identified through frequency and dispersion analyses across a learner corpus and two large general reference corpora. This initial pool was subsequently reviewed by three senior applied linguistics specialists, who independently evaluated each item for pedagogical relevance, naturalness, and instructional value. Only collocations receiving full agreement were retained, ensuring both linguistic authenticity and curricular appropriateness.

Receptive collocational knowledge was assessed through three parallel test forms administered as a pretest, an immediate posttest, and a delayed posttest. Each form comprised 40 multiple-choice items embedded in short contextualized sentences. Participants were required to select the most natural verb–noun combination from four options. Distractors were systematically controlled for part of speech, semantic relatedness, and frequency band to minimize construct-irrelevant variance and guessing effects. Parallel-form equivalence was established through piloting with a comparable learner cohort before the main study.

Throughout the intervention, all learner interactions were automatically recorded by the digital practice platform. For participants in the LLM-mediated condition, log data captured the number of prompt attempts per item, system-generated feedback messages, response latency, and total retrieval cycles. For those in the corpus-based condition, logs documented search attempts, the number of concordance lines consulted, time spent per item, and navigation patterns. These behavioral data provided a complementary process-oriented perspective on learner engagement. Following the intervention, participants completed a brief post-task questionnaire assessing perceived noticing, feedback usefulness, and trust in the instructional input.

3.3.2 Interview Design and Validity Considerations

To explore learners' experiences of the instructional conditions in greater depth and to illuminate processes underlying the quantitative findings, semi-structured interviews were conducted with a purposive subsample of participants from both groups. The interview protocol comprised six broad, open-ended questions designed to invite reflection on learners' cognitive engagement during practice, perceptions of the authenticity and credibility of instructional input, and motivational experiences associated with retrieval and feedback (see Appendix for interview questions). The interview questions were explicitly aligned with the study's research questions and conceptual framework, with each prompt designed to elicit learners' reflections on cognitive engagement, perceptions of input authenticity, and motivational dynamics underlying retrieval practice. This alignment ensured that the qualitative data could meaningfully illuminate the mechanisms associated with observed learning gains and identified process mediators, while remaining sufficiently open to capture unanticipated aspects of learner experience.

Rather than targeting discrete constructs through narrowly specified prompts, the questions were intentionally framed to allow participants to direct the focus of their responses based on aspects they found most salient. Follow-up prompts were used flexibly to encourage elaboration, clarification, and exemplification when relevant.

The content validity of the interview instrument was established through close alignment with the study's research questions and conceptual framework. The interview guide was reviewed by two external researchers with expertise in qualitative methodology, who evaluated the clarity, scope, and coherence of the questions. Minor revisions were made to refine wording and reduce potential ambiguity. In addition, a pilot interview was conducted to confirm that participants interpreted the questions as intended and were able to draw meaningfully on their instructional experiences in responding.

3.4 Data Collection Procedure

The intervention spanned two weeks, during which participants completed six practice sessions of approximately 30–40 minutes each. In the LLM-mediated condition, learners were presented with a cue sentence missing the target noun. They attempted to retrieve the collocate, received immediate correctness feedback, and, if incorrect, were provided with an LLM-generated model sentence and a brief reflection prompt encouraging them to compare their attempt with the correct form. All LLM-generated examples were manually pre-screened by the research team to ensure accuracy and frequency realism.

The corpus-based condition followed a parallel structure. Learners attempted the retrieval task, received deterministic feedback showing the correct collocate, and examined three authentic concordance lines illustrating natural usage. Reflection prompts encouraged them to compare contexts and notice typical lexical patterns. Participants completed the receptive test one week before the intervention (pretest), immediately after the final session (posttest), and four weeks later (delayed posttest) to measure retention.

Qualitative interviews took place within two weeks after the delayed posttest to allow learners to reflect on their experiences while maintaining recall accuracy. Interviews were conducted using an interview guide that allowed flexibility for probing and follow-up questions, enabling deeper exploration of participants' experiences. All interviews lasted approximately 30–40 minutes, were audio-recorded with participant consent, and transcribed verbatim for subsequent analysis (Braun & Clarke, 2022).

3.5 Data Analysis

Quantitative analyses focused on receptive test scores and process measures. Each participant responded to multiple items across three testing occasions, producing a hierarchical data structure. Accordingly, receptive scores were analyzed using linear mixed-effects models (LMMs) with fixed

effects for condition and time, and crossed random intercepts for participants and items. This approach accounts for individual and item-level variability and provides more reliable estimates than traditional ANOVA-type procedures in repeated-measures language-learning data (Brysbaert, 2025). Model assumptions were examined through residual diagnostics, which indicated acceptable distributional patterns and no violations affecting interpretability. Given the large number of observations per participant (40 items × 3 tests), the sample size was sufficient for stable estimation of fixed and random effects.

Process measures, retrieval attempts, contextual variability, and time-on-task, were analyzed using parallel LMMs to evaluate group differences and then incorporated into a multilevel mediation model. Indirect effects were estimated using bootstrapped confidence intervals, a robust procedure for testing mediation in hierarchical data (Jia & Hui, 2025).

For the qualitative component, interview transcripts were analyzed using reflexive thematic analysis, with an emphasis on sustained, iterative, and interpretive engagement with the data (Braun & Clarke, 2024). Analysis was supported by MAXQDA 2022, which was used to organize transcripts, manage coding, and facilitate systematic comparison across data segments. The researcher conducted multiple rounds of close reading to build analytic familiarity, generated inductive initial codes, and progressively refined these codes into candidate themes through ongoing comparison and abstraction. Throughout the analytic process, emerging themes were critically examined for internal coherence, distinctiveness, and adequacy in representing participants' perspectives (Nowell et al., 2017). The final thematic structure was established through a careful review of coded extracts in relation to the evolving analytic framework, ensuring conceptual clarity and analytic rigor. This approach enabled the qualitative strand to complement and contextualize the quantitative findings meaningfully, offering a nuanced account of both learning outcomes and learners' experiences.

3.6 Validity, Reliability, and Rigor

To ensure measurement reliability, the collocation item pool was piloted with an independent sample before the main study. Items were calibrated to achieve comparable difficulty across test forms, and internal consistency was verified. In the intervention, fidelity was maintained by standardizing session procedures and pre-validating all LLM feedback and example sentences.

For the qualitative strand, credibility and trustworthiness were strengthened through strategies recommended in recent qualitative methodology research. These included member checking, in which participants reviewed preliminary thematic summaries for resonance and accuracy (Soysal & Türkmen, 2024); an audit trail documenting analytic decisions; and reflexive memoing to enhance transparency about researcher positionality and interpretive influence (Olmos-Vega et al., 2022). Peer debriefing sessions further enhanced dependability, while careful documentation of

coding procedures strengthened confirmability. Collectively, these steps ensured that both quantitative and qualitative results were robust, interpretable, and methodologically sound in line with contemporary standards of rigor (Nowell et al., 2017).

4. Results

The results of the study are presented in two complementary strands. Quantitative analyses indicate that learners in the LLM-mediated retrieval practice condition achieved larger immediate and delayed gains in receptive collocational knowledge than those in the corpus-based condition, with process measures indicating higher retrieval frequency and greater contextual variability. Qualitative findings provide insight into learners' experiences, revealing enhanced noticing, evolving trust in instructional input, and sustained motivational engagement, thereby contextualizing the mechanisms underlying the observed quantitative outcomes. Together, these findings provide a comprehensive understanding of both the effectiveness and learner-perceived affordances of LLM-supported collocation instruction.

4.1 Quantitative Results

The quantitative results are presented sequentially, beginning with descriptive statistics, followed by mixed-effects modeling, process measures, and mediation analyses. This organization allows a clear understanding of both overall performance patterns and the mechanisms underlying learning gains.

4.1.1 Descriptive Patterns in Receptive Collocational Knowledge

Table 1 presents mean scores and standard deviations for receptive collocation knowledge across the pretest, immediate posttest, and delayed posttest for both instructional conditions. Both groups demonstrated clear improvement from pretest to immediate posttest, with partial retention at the delayed posttest.

Mean scores for the LLM-mediated group were consistently higher than those of the corpus-based group, with larger gains observed immediately after the intervention and smaller declines at delayed testing. Standard deviations remained relatively stable over time, suggesting that improvements were broadly distributed rather than driven by a small subset of participants. These descriptive patterns indicate potentially divergent learning trajectories across conditions, which were formally tested using linear mixed-effects modeling.

Table 1. Descriptive statistics for receptive collocation test scores (0–40)

Group	Pretest M (SD)	Immediate Posttest M (SD)	Delayed Posttest M (SD)
LLM-mediated (n = 31)	21.7 (3.6)	33.6 (3.2)	31.7 (3.7)
Corpus-based (n = 34)	21.3 (3.8)	29.4 (4.0)	27.2 (4.5)

M = mean; SD = standard deviation

4.1.2 Condition-by-Time Effects: Linear Mixed-Effects Model

To formally evaluate whether the observed differences were statistically reliable while accounting for repeated measures and item-level variability, receptive test scores were analyzed using a linear mixed-effects model with fixed effects for condition, time, and their interaction, and random intercepts for participants and items. Model diagnostics confirmed acceptable residual distributions (Table 2).

Table 2. Fixed effects from mixed-effects model

Predictor	Estimate (β)	SE	t	p
Intercept (Corpus–Pretest)	21.31	0.53	40.21	< .001
Time: Immediate Posttest	8.07	0.61	13.22	< .001
Time: Delayed Posttest	5.91	0.57	10.37	< .001
Condition: LLM	0.35	0.78	0.45	.654
LLM × Time (Immediate)	+4.02	0.84	4.79	< .001
LLM × Time (Delayed)	+4.41	0.80	5.51	< .001

SE = standard error; β = unstandardized regression coefficient; p = probability value

The model revealed significant main effects of time, indicating that both groups improved from pretest to immediate posttest and retained gains at delayed posttest. Crucially, the interaction between condition and time was significant at both posttest points, demonstrating that learners in the LLM-mediated condition achieved larger gains than those in the corpus-based group and maintained superior retention over time. The absence of a significant main effect of condition at pretest confirms that groups were comparable at baseline, strengthening the interpretation that post-intervention differences reflect instructional effects.

4.1.3 Process Measures

To investigate potential mechanisms underlying the observed learning differences, group differences in process variables, retrieval attempts, contextual variability, and time-on-task were analyzed using parallel linear mixed-effects models (Table 3). These analyses tested whether engagement patterns differed systematically across instructional conditions while accounting for participant- and item-level variability.

Table 3. Group differences in process variables

Variable	LLM M (SD)	Corpus M (SD)	Group Difference (Cohen's d)
Retrieval attempts per item	2.48 (0.51)	1.92 (0.48)	1.08
Contextual variability	5.91 (1.02)	3.31 (0.71)	2.64
Time-on-task (min/session)	35.3 (4.1)	32.6 (4.2)	0.63

M = mean; SD = standard deviation

Results indicated that learners in the LLM-mediated condition engaged in significantly more retrieval attempts per item and experienced substantially greater contextual variability than learners in the corpus-based condition. Differences in time-on-task were modest. These findings suggest that the qualitative characteristics of engagement, intensive retrieval, and exposure to diverse exemplars played a more important role than overall session duration in driving learning differences.

4.1.4 Mediation Analyses

To test whether engagement patterns statistically mediated the effect of instructional condition on immediate posttest performance, a multilevel mediation model was fitted, with bootstrapped confidence intervals providing robust estimates of indirect effects (Table 4).

Table 4. Indirect effects of process variables on immediate posttest performance

Mediator	Indirect Effect (β)	95% CI	p
Retrieval attempts	0.97	0.46–1.65	.001
Contextual variability	1.89	1.02–2.78	< .001
Time-on-task	0.11	−0.05–0.31	.241

β = unstandardized regression coefficient; CI = confidence interval; p = probability value

Results indicated significant indirect effects for retrieval attempts and contextual variability, with confidence intervals excluding zero. This confirms that a substantial portion of the LLM-mediated advantage operated through increased retrieval frequency and exposure to a wider range of exemplars. The indirect effect of time-on-task was not statistically significant, indicating that total session duration did not meaningfully contribute to performance differences once other process variables were accounted for.

Taken together, these analyses provide robust inferential evidence that the LLM advantage was driven primarily by the intensity and diversity of engagement rather than total practice duration, complementing outcome-level findings and reinforcing causal interpretation.

4.2 Qualitative Results

The qualitative strand aimed to illuminate learners' experiences of the two instructional conditions and to clarify mechanisms contributing to the observed quantitative outcomes. Analysis of the semi-structured interviews generated three overarching themes, each with sub-themes supported by illustrative learner narratives (Table 5). These themes capture learners' cognitive and affective engagement, evolving judgments about input credibility, and motivational dynamics sustaining participation.

Table 5. Main themes, sub-themes, and sample narratives

Main Theme	Sub-theme	Sample Narrative (Illustrative Excerpt)
Heightened Noticing Through Adaptive Scaffolding	Salience of collocational patterns	"The examples were so focused that the key words felt like they almost jumped out at me." (Interviewee No. 2)
	Prompted comparison and reflective noticing	"When I chose the wrong word, the AI didn't just correct me—it made me think about why the right one worked." (Interviewee No. 5)
	Reduced cognitive clutter in input processing	"Corpus sentences had so much going on that I sometimes lost track of the target word." (Interviewee No. 1)
Authenticity, Trust, and the Credibility of Input	Early skepticism and gradual acceptance of AI-generated examples	"At first I doubted whether the sentences were natural, but once I learned they were checked, I started trusting them." (Interviewee No. 9)
	Corpus as inherently authoritative yet cognitively taxing	"I trusted the corpus because it's real language, but some sentences were too long or specialized to actually help me learn." (Interviewee No. 4)
	Negotiating the balance between naturalness and learnability	"The AI examples felt simpler, but in a good way—they were clear without sounding fake." (Interviewee No. 7)
Motivated Engagement in Retrieval Cycles	Retrieval as a challenge that sustains involvement	"Trying to guess the right collocation became almost addictive—I wanted to keep going." (Interviewee No. 10)
	Momentum supported by immediate, adaptive feedback	"The instant response kept my attention; I didn't drift like I do with printed exercises." (Interviewee No. 6)
	Emotional safety enabling risk-taking in practice	"I didn't feel judged for mistakes, so I didn't hesitate to try again and again." (Interviewee No. 3)

Theme 1: Heightened Noticing through Adaptive Scaffolding

Learners consistently described LLM-generated examples as clear, focused, and cognitively manageable, which facilitated the detection of target collocational patterns. By reducing lexical density and emphasizing the relevant verb–noun pairing, the AI scaffolded learners' attention and minimized extraneous processing demands. Prompts encouraging comparison between correct and incorrect responses further stimulated reflective noticing, allowing learners to actively engage with form-meaning relationships. In contrast, corpus-based examples, while authentic, were often perceived as cognitively demanding, potentially limiting attentional resources available for noticing. These accounts suggest that adaptive scaffolding via LLMs can enhance salience and accelerate the noticing process, a key mechanism in lexical acquisition.

Theme 2: Authenticity, Trust, and the Credibility of Input

Learners initially expressed skepticism toward AI-generated examples, questioning whether simplified sentences represented "real" English. Over time, pre-validation and consistent quality of LLM output increased trust and encouraged engagement. While corpus-based examples were valued for their authenticity, they sometimes imposed higher cognitive load, particularly when sentences were lengthy or syntactically complex. Participants negotiated a balance between authenticity and learnability, ultimately appreciating LLM-generated sentences for their pedagogically optimized naturalness, clear enough to support comprehension but sufficiently authentic to be credible. This theme highlights that perceived trustworthiness and instructional suitability jointly influenced learner engagement and uptake.

Theme 3: Motivated Engagement in Retrieval Cycles

The LLM environment fostered sustained engagement through immediacy, adaptiveness, and emotional safety. Learners reported that rapid, targeted feedback created a sense of momentum and maintained focus across repeated retrieval attempts. The non-judgmental tone encouraged risk-taking, allowing learners to experiment without fear of failure. This supportive environment appeared to enhance both cognitive persistence and affective motivation, ensuring learners actively participated in multiple retrieval cycles, a factor central to durable learning. In short, LLM-mediated design provided cognitive scaffolding and motivational support, amplifying the effectiveness of retrieval practice.

In sum, the qualitative findings elucidate the mechanisms underlying the LLM-mediated intervention's effectiveness. Theme 1 demonstrates that heightened noticing and reduced cognitive load made target collocations more salient, supporting higher posttest scores quantitatively. Theme 2 shows that trust in instructional input encouraged sustained engagement, while Theme 3 highlights that immediate, adaptive feedback and emotional safety fostered repeated retrieval

cycles. These experiential insights directly align with quantitative results: increased retrieval attempts and greater contextual variability statistically mediated the superior learning gains of the LLM group. Together, the qualitative and quantitative strands provide a cohesive, multidimensional account, showing that LLM scaffolding not only enhanced measurable collocational knowledge but also shaped learner behaviors and perceptions in ways that sustained active, effective engagement.

5. Discussion

The present study examined whether scaffolded, LLM-mediated retrieval practice could strengthen EFL learners' receptive verb–noun collocational knowledge more effectively than a tightly controlled corpus-based alternative. Across all quantitative analyses, learners in the LLM group (n = 31) demonstrated substantially larger gains from pretest to immediate posttest compared with the corpus group (n = 34), and these advantages were largely maintained at the four-week delayed posttest. These findings offer some of the clearest causal evidence to date that, when embedded within a principled pedagogical design, LLMs can deliver more durable multiword lexical gains than conventional corpus-based approaches, particularly when retrieval practice is structured to promote long-term retention (Fang et al., 2024; Huang, 2025).

A central contribution of the study lies in clarifying why the LLM condition outperformed the corpus condition. Mediation analyses revealed that learners in the LLM condition completed more successful retrieval attempts and encountered a wider range of contextualized exemplars—two variables that have long been recognized as key drivers of durable lexical learning. These patterns parallel usage-based accounts emphasizing that robust entrenchment emerges from repeated exposure to variable input (Bybee, 2010) and align with cognitive-memory research demonstrating that spaced, effortful retrieval consolidates long-term retention (Bego et al., 2024; Fang et al., 2024; Lyle et al., 2022). From a cognitive psychology perspective, retrieval practice supported by appropriately spaced schedules is especially effective when learners are repeatedly required to reconstruct form–meaning mappings, a mechanism that LLM scaffolding appears well suited to operationalize (Huang, 2025). The current findings extend this body of work by showing that LLMs, when appropriately scaffolded and quality-controlled, can instantiate these theoretically established mechanisms at scale and with a high degree of consistency.

Equally noteworthy is what did not explain the learning gains. Time-on-task showed no significant mediating effect, indicating that the LLM advantage cannot simply be attributed to increased effort or prolonged engagement. Instead, the qualitative characteristics of practice, particularly adaptive feedback and contextual variability, proved to be the meaningful levers. This distinction is important because existing research often attributes the benefits of digital tools to convenience or increased practice volume; the present study demonstrates that LLM-supported retrieval practice can influence deeper cognitive processes central to multiword learning, consistent

with findings that retrieval quality matters more than sheer quantity (Fang et al., 2024; Huang, 2025).

These results speak directly to long-standing challenges in collocational acquisition. Decades of research show that learners struggle to notice, store, and retrieve multiword combinations whose boundaries are often not perceptually salient (Du, 2022; Sun & Park, 2023). The LLM-mediated tasks appeared to counter several of these challenges. Immediate, tailored feedback likely heightened the salience of co-occurrence patterns, helping learners overcome the limitations of exposure-based approaches that often leave collocational probabilities under-noticed (Schmidt, 1990). Likewise, verified LLM exemplars offered contextual diversity that textbooks or fixed corpus concordances rarely achieve. Emerging discussions in applied linguistics suggest that LLMs may be uniquely positioned to deliver such variability while maintaining pedagogical control (Cong, 2024), and the present findings provide empirical grounding for that claim.

The comparison with corpus-based retrieval practice also offers practical insights. Although corpus-based methods remain a cornerstone of collocation pedagogy and reliably enhance receptive knowledge (Sun & Park, 2023), they may lack the adaptive qualities that characterize LLM-generated feedback. The corpus condition in this study was intentionally designed to reflect best practice—controlled exposure, matched retrieval scheduling, and authentic concordance lines, yet the LLM condition still produced more robust and persistent gains. This suggests that responsiveness and instructional sensitivity, rather than authenticity alone, may be critical for maximizing collocational learning. Importantly, this does not imply that LLMs should replace corpora; rather, LLM output must be grounded in corpus evidence to ensure lexical realism, particularly given well-documented risks of frequency distortions or hallucinations.

The qualitative findings further nuance these interpretations. Learners valued the clarity, immediacy, and motivational affordances of LLM feedback but continued to express awareness of potential inaccuracies. This aligns with broader discussions in the AI-in-education literature cautioning that LLMs, while pedagogically powerful, require structured verification and transparent use to ensure alignment with linguistic norms (Kukulska-Hulme, 2024). In the present study, all LLM examples were pre-verified, and learners' comments suggested that this vetting process was essential for establishing trust and sustaining engagement. Without such safeguards, it remains uncertain whether similar learning gains would emerge.

Methodologically, the study contributes to a growing call for more rigorous evaluations of AI-based instruction. Much existing work relies on informal classroom observations, exploratory case studies, or self-reported improvement (Li et al., 2022). By employing a randomized controlled design, carefully controlling input and retrieval schedules, and including delayed testing, the present study provides stronger evidence regarding the causal impact of LLM-mediated practice. The sizeable delayed-retention advantage is especially noteworthy, given consistent findings that

collocational gains often decay rapidly in the absence of high-quality retrieval practice (Fang et al., 2024; Sun & Park, 2023).

From a theoretical perspective, the findings support the integration of usage-based and memory-based frameworks. LLM-mediated practice enhanced both noticing and entrenchment, the two cognitive phases most critical for internalizing collocations (Bybee, 2010; Schmidt, 1990). By combining repeated exposure to variable input with structured retrieval and feedback, LLMs appear capable of operationalizing principles long advocated in the literature but difficult to implement at scale, a point also emphasized in recent cognitive accounts of AI-supported learning (Huang, 2025).

Nevertheless, LLMs are not a panacea. Their effectiveness depends on careful instructional design, controlled prompting, and corpus-based verification. The findings underscore the importance of principled integration rather than enthusiasm-driven adoption. When used deliberately, LLMs may function as a pedagogically enriched layer atop traditional methods rather than a wholesale replacement.

Taken together, the quantitative outcomes and qualitative insights converge to demonstrate that scaffolded LLM use enhances collocational learning by increasing retrieval success, providing diverse and meaningful exemplars, and fostering sustained cognitive and motivational engagement.

6. Conclusion

This study demonstrates that scaffolded LLM-mediated retrieval practice can significantly enhance EFL learners' receptive verb–noun collocational knowledge compared with a rigorously matched corpus-based alternative. Learners in the LLM condition (N = 31) not only achieved larger immediate gains than those in the corpus condition (N = 34) but also retained these improvements more effectively over time**.** The mediation analyses further revealed that increased successful retrieval attempts and richer contextual variability were central to these gains, highlighting the capacity of LLMs to operationalize core mechanisms emphasized in usage-based and cognitive-memory theories.

These findings carry several pedagogical implications. For practitioners, LLMs can serve as a powerful complement to existing corpus-based materials by offering adaptive feedback, varied exemplars, and responsive scaffolding. For curriculum designers, the results underscore the value of designing retrieval-based activities that incorporate LLM-generated examples while ensuring authenticity through systematic verification. More broadly, the study contributes to ongoing discussions about principled AI integration, showing that LLMs are most effective when embedded within structured, theory-driven instructional frameworks.

At the same time, several limitations warrant caution. The study focused on receptive knowledge of a single collocation type, leaving open the question of whether similar benefits

extend to productive use or to other multiword constructions. The intervention duration was relatively short, and while delayed gains were robust, longer-term developmental trajectories remain unexplored. Furthermore, although the qualitative insights were rich, they were drawn from a modest sample, and broader learner populations may display different patterns of engagement or trust in LLM output.

Future research should examine productive collocational use, compare different forms of LLM scaffolding (e.g., prompt engineering, staged feedback, error-contingent support), and investigate how proficiency or learning style interacts with LLM-mediated practice. Longitudinal studies tracking retention over extended periods would offer further insight into the durability of gains.

In sum, the present study provides strong evidence that LLM-mediated retrieval practice, when carefully designed, verified, and theoretically grounded, can meaningfully advance collocational competence. As AI becomes increasingly embedded in language education, such principled, empirically tested models will be essential for guiding responsible and effective pedagogical use.

Acknowledgements

The authors would like to thank the students who participated in this study.

Conflict of interest

The author declares no conflict of interest.

References

Bego, C. R., Lyle, K. B., Ralston, P. A. S., & Immekus, J. C., Chastain, R. J., Haynes, L. D., Hoyt, L. K., Pigg, R. M., Rabin, S. D., Scobee, M. W., Starr, T. L. (2024). Single-paper meta-analyses of the effects of spaced retrieval practice in nine introductory STEM courses: Is the glass half full or half empty? *International Journal of STEM Education, 11*(1), Article 9. https://doi.org/10.1186/s40594-024-00468-5

Boone, G., & Eyckmans, J. (2023). Exploring collocation development in L2 German from students' perspective: A contrasting case study. *Studies in Second Language Learning and Teaching, 13*(3), 571–599. https://doi.org/10.14746/ssllt.32539

Braun, V., & Clarke, V. (2022). *Thematic analysis: A practical guide* (2nd ed.). SAGE Publications. https://www.amazon.com/Thematic-Analysis-Practical-Virginia-Braun/dp/1473953243

Braun, V., & Clarke, V. (2024). Supporting best practice in reflexive thematic analysis reporting in palliative medicine: A review of published research and introduction to the reflexive thematic analysis reporting guidelines (RTARG). *Palliative Medicine, 38*(6), 608–616. https://doi.org/10.1177/02692163241234800

Brysbaert, M. (2025). Applying mixed-effects models in research on second language acquisition: A tutorial for beginners. *Languages, 10*(2), Article 20. https://doi.org/10.3390/languages10020020

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press. https://doi.org/10.1017/CBO9780511750526

Cong, Y. (2024). AI language models: An opportunity to enhance language learning. *Informatics, 11*(3), Article 49. https://doi.org/10.3390/informatics11030049

Du, X. (2022). Collocation use in EFL learners' writing across multiple language proficiencies: A corpus-driven study. *Frontiers in Psychology, 13*, Article 752134. https://doi.org/10.3389/fpsyg.2022.752134

Fang, N., Elgort, I., & Chen, Z. (2024). Effects of retrieval schedules on the acquisition of explicit, automatized-explicit, and implicit knowledge of L2 collocations. *Studies in Second Language Acquisition, 46*(3), 663–685. https://doi.org/10.1017/S0272263124000184

Huang, M. (2025). Spaced repetition and retrieval practice: Efficient learning mechanisms from a cognitive psychology perspective and their empowerment by AI. *International Journal of Asian Social Science Research, 2*(6), 31-37. https://doi.org/10.70267/ijassr.250206.3137

Jeong, H., & DeKeyser, R. (2023). Development of automaticity in processing L2 collocations: The roles of L1 collocational knowledge and practice condition. *Studies in Second Language Acquisition, 45*(4), 930–954. https://doi.org/10.1017/S0272263122000547

Jia, R., & Hui, B. (2025). Modeling relationships between learning conditions, processes, and outcomes: An introduction to mediation analysis in SLA research. *Studies in Second Language Acquisition, 47*(3), 912–947. https://doi.org/10.1017/S0272263125100867

Kukulska-Hulme, A., Friend Wise, A. F., Coughlan, T., Biswas, G., Bossu, C., Burriss, S. K., & Whitelock, D. (2024). *Innovating Pedagogy 2024: Exploring new forms of teaching, learning and assessment to guide educators and policy makers* (Open University Innovation Report 12). The Open University. https://doi.org/10.13140/RG.2.2.15627.30246

Li, H., Paterson, K. B., Warrington, K. L., & Wang, X. (2022). Insights into the processing of collocations during L2 English reading: Evidence from eye movements. *Frontiers in Psychology, 13*, 845590. https://doi.org/10.3389/fpsyg.2022.845590

Liu, G., Darvin. R., & Ma, C. (2024). Exploring AI-mediated informal digital learning of English: Acceptance and use of LLM platforms (AI-IDLE): a mixed-method investigation of Chinese EFL learners' AI adaptation and experiences. *Computer Assisted Language Learning*, *38*(7), 1632-1660. https://doi.org/10.1080/09588221.2024.2310288

Liu, T., & Gablasova, D. (2025). Data-driven learning of collocations by Chinese learners of English: A longitudinal perspective. *Computer Assisted Language Learning, 38*(3), 612–637. https://doi.org/10.1080/09588221.2023.2214605

Lyle, K. B., Bego, C. R., Ralston, P. A. S., & Immekus, J. C. (2022). Spaced retrieval practice imposes desirable difficulty in calculus learning. *Educational Psychology Review, 34*(4), 1799–1812. https://doi.org/10.1007/s10648-022-09677-2

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods, 16*(1), 1–13. https://doi.org/10.1177/1609406917733847

Olmos-Vega, F. M., Stalmeijer, R. E., Varpio, L., & Kahlke, R. (2022). A practical guide to reflexivity in qualitative research: AMEE Guide No. 149. *Medical Teacher, 45*(3), 241–251. https://doi.org/10.1080/0142159X.2022.2057287

Pack, A., & Maloney, J. (2023). Using generative artificial intelligence for language education research: Insights from using OpenAI's ChatGPT. *TESOL Quarterly, 57*(4), 1571-1582. https://doi.org/10.1002/tesq.3253

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Soysal, Y., & Türkmen, G. (2024). Reinterpreting the member checking validation strategy in qualitative research through the hermeneutics lens. *Qualitative Inquiry in Education: Theory & Practice / QIETP*, *2*(1), 42-63. https://doi.org/10.1016/j.system.2023.103208

Sun, W., & Park, E. (2023). EFL learners' collocation acquisition and learning in corpus-based instruction: A systematic review. *Sustainability, 15*(17), 13242. https://doi.org/10.3390/su151713242

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications, 12*, Article 621. https://doi.org/10.1057/s41599-025-04787-y

Wang, W. (2024). Recent research on L2 vocabulary collocations acquisition: Implications for language teaching. *International Journal of Education and Humanities, 16*(2), 305–308. https://doi.org/10.54097/16mmbj24

Yang, L., & Li, R. (2024). ChatGPT for L2 learning: Current status and implications. *System, 124*, 103351. https://doi.org/10.1016/j.system.2024.103351

Appendix

Interview Questions

1. Can you tell me about your overall experience with the retrieval practice activities? Which parts helped you notice or remember verb–noun combinations the most?

2. How did the examples and prompts you worked with influence the way you recognized or thought about collocations?

3. When you chose a wrong answer, how did the feedback you received shape your understanding or help you figure out the correct collocation?

4. What were your first impressions of the examples—did they feel natural or trustworthy? Did your perception change as you went through the sessions?

5. How did the learning activities affect your motivation to keep trying, especially when you were unsure or made mistakes?

6. Did the pace of the tasks or the amount of time you spent on them affect your focus or persistence? How so?